



# **Bioinformatics Databases and Data Resources**

## **Reference Atlas**

**First Edition – Integrated Master Edition**

**A Practical Guide to Biological Data Resources, Repositories, Knowledgebases,  
Ontologies, Tools, Their Uses, Strengths, Limitations, and Scientific Applications**

<b>Author</b>	<b>Mohamed Mostafa</b>
<b>Edition Date</b>	<b>May 2026</b>
<b>Verification Date</b>	<b>May 2026</b>

## First Edition Editorial Note

This document represents the first integrated master edition of the Bioinformatics Databases and Data Resources Reference Atlas. It is built on the core atlas draft and incorporates expanded coverage across controlled-access human genomics repositories, proteomics repositories, metabolomics resources, reference gene annotation, expression and proteome atlases, GWAS and rare disease resources, ontology ecosystem resources, structural biology archives, pharmacogenomics and drug-target resources, cancer model systems, non-coding RNA databases, and comparative genomics and orthology resources. The Atlas draft remains the structural and editorial foundation of this document. The expanded material has been integrated to improve scientific coverage, database-selection guidance, comparison tables, reproducibility guidance, quality-control warnings, and appendices.

This atlas is curated and representative, not an exhaustive encyclopedia of every biological database ever created. Database URLs, APIs, access policies, release versions, and citations change frequently; users should verify current information before publication, production use, or clinical interpretation.

### Bioinformatics Databases and Data Resources Reference Atlas

**Author:** Mohamed Mostafa

**Version:** First Edition – Integrated Master Edition

**Publication Year:** 2026

**DOI:** <https://doi.org/10.5281/zenodo.20533722>

**Zenodo Record:** <https://zenodo.org/records/20533722>

**License:** Creative Commons Attribution 4.0 International (CC BY 4.0)

**Citation:** Mohamed Mostafa Mohamed. (2026). Bioinformatics Databases and Data Resources Reference Atlas: A Practical Guide to Biological Data Resources, Repositories, Knowledgebases, Ontologies, Tools, Their Uses, Strengths, Limitations, and Scientific Applications (First Edition – Integrated Master Edition) [Computer software]. Self-published. <https://doi.org/10.5281/zenodo.20533722>

# Table of Contents

First Edition Editorial Note.....	2
Table of Contents .....	3
Start Here: How to Navigate This Atlas.....	13
Quick Navigation .....	15
How to Use This Atlas.....	16
Section 1: Executive Summary .....	17
Section 2: Scope and Limitations .....	19
Section 3: Standard 31-Field Database Card Template.....	20
Section 4: Database and Resource Type Definitions.....	21
Section 5: Database Types — Definitions and Distinctions .....	23
Section 6: FAIR and Reproducibility Principles.....	30
Section 7: Database Release and Verification Policy.....	32
Section 8: Decision Maps — Start with the Scientific Question.....	34
Critical Scientific Misuse Warnings .....	47
Category A: General Bioinformatics Portals and Integrated Resources .....	48
Category Overview .....	48
A1 – NCBI (National Center for Biotechnology Information) .....	49
A2 – EMBL-EBI (European Bioinformatics Institute).....	52
A3 – DDBJ (DNA Data Bank of Japan).....	55
A4 – ExPASy (Expert Protein Analysis System).....	58
A5 – Ensembl.....	61
Category B: Literature and Scientific Publications .....	66
Category Overview .....	66
B1 – PubMed .....	67
B2 – PubMed Central (PMC) .....	70
B3 – Europe PMC .....	72
B4 – Google Scholar.....	74
B5 – Semantic Scholar.....	76
B6 – Scopus [COMMERCIAL — Institutional Access Required] .....	78
B7 – Web of Science [COMMERCIAL — Institutional Access Required].....	80
Category C: Nucleotide Sequence Databases .....	84
Category Overview .....	84
C1 – GenBank.....	85
C2 – EMBL/ENA (European Nucleotide Archive).....	87
C3 – DDBJ Sequence Database.....	89
C4 – RefSeq .....	91
C5 – NCBI Nucleotide.....	94



Category D: Sequence Similarity and Search Tools .....	98
Category Overview .....	98
D1 – NCBI BLAST (Basic Local Alignment Search Tool).....	99
D2 – EBI BLAST (NCBI-BLAST at EMBL-EBI) .....	102
D3 – HMMER.....	104
D4 – PSI-BLAST (Position-Specific Iterated BLAST) .....	106
D5 – DIAMOND (Double Index Alignment of Next-generation sequencing Data).....	108
D6 – FASTA (EBI Sequence Similarity Search) .....	110
Category E: Genome Browsers and Genome Annotation .....	114
Category Overview .....	114
E1 – Ensembl Genome Browser .....	115
E2 – UCSC Genome Browser .....	117
E3 – NCBI Genome Data Viewer .....	119
E4 – JBrowse .....	121
E5 – IGV (Integrative Genomics Viewer).....	123
Category F: Gene and Genome Databases .....	127
Category Overview .....	127
F1 – NCBI Gene.....	128
F2 – GeneCards .....	130
F3 – HGNC (HUGO Gene Nomenclature Committee) .....	132
F4 – OMIM (Online Mendelian Inheritance in Man) — Cross-reference Entry .....	134
F5 – Ensembl Genes .....	134
Category G: Transcriptomics and Gene Expression Databases .....	138
Category Overview .....	138
G1 – GEO (Gene Expression Omnibus).....	139
G2 – ArrayExpress / BioStudies [Note: ArrayExpress has been integrated into BioStudies at EMBL-EBI] .....	142
G3 – Expression Atlas .....	144
G4 – GTEx (Genotype-Tissue Expression) .....	146
G5 – SRA (Sequence Read Archive) — Cross-reference Entry.....	148
Category H: Raw Sequencing Data Repositories.....	150
CATEGORY OVERVIEW .....	150
H1 – SRA (Sequence Read Archive) .....	151
H2 – ENA (European Nucleotide Archive) .....	155
H3 – DRA (DDBJ Sequence Read Archive) .....	158
Category I: Protein Sequence and Function Databases .....	162
CATEGORY OVERVIEW .....	162
I1 – UniProt (Universal Protein Resource).....	163
I2 – Swiss-Prot (manually reviewed section of UniProtKB) .....	167
I3 – TrEMBL (computationally annotated section of UniProtKB) .....	170



I4 – InterPro — Cross-reference Entry .....	172
I5 – PROSITE — Cross-reference Entry.....	172
I6 – SMART (Simple Modular Architecture Research Tool) .....	173
Category J: Protein Family, Domain, and Motif Databases.....	177
CATEGORY OVERVIEW .....	177
J1 – Pfam (Protein Families Database) NOTE: Pfam has been integrated into InterPro.....	178
J2 – InterPro .....	180
J3 – CDD (Conserved Domain Database) .....	181
J4 – PROSITE (in Category J context).....	183
J5 – PRINTS (Protein Fingerprint Database.....	184
J6: SUPERFAMILY .....	186
Category K: Protein Structure Databases .....	190
CATEGORY OVERVIEW .....	190
K1: RCSB PDB (Protein Data Bank).....	191
K2 – PDBe (Protein Data Bank in Europe) .....	194
K3 – PDBj (Protein Data Bank Japan) .....	197
K4 – AlphaFold Protein Structure Database.....	199
K5 – SWISS-MODEL Repository .....	202
Category L: Variant and Mutation Databases.....	206
CATEGORY OVERVIEW .....	206
dbSNP (Database of Single Nucleotide Polymorphisms) .....	207
ClinVar .....	210
L3: gnomAD (Genome Aggregation Database).....	213
L4: COSMIC (Catalogue of Somatic Mutations in Cancer) — Cross-reference Entry.....	215
L5: Ensembl VEP (Variant Effect Predictor).....	216
L6: LOVD (Leiden Open Variation Database) .....	219
L7: ClinGen (Clinical Genome Resource) — Cross-reference Entry .....	220
Category M: Disease and Clinical Genomics Databases .....	222
CATEGORY OVERVIEW .....	222
M1: OMIM (Online Mendelian Inheritance in Man) .....	223
M2: Orphanet .....	226
M3: ClinGen — Cross-reference Entry.....	228
M4: DisGeNET .....	229
M5: MalaCards (Human Disease Database).....	232
Category N: Pathway and Systems Biology Databases.....	236
CATEGORY OVERVIEW .....	236
N1: KEGG (Kyoto Encyclopedia of Genes and Genomes) .....	237
N2: Reactome.....	240



N3: BioCyc (Collection of Pathway/Genome Databases) NOTE: BioCyc has tiered access; some features and databases require a subscription. EcoCyc (E. coli) and MetaCyc (metabolic pathways) have more open access. ....	243
N4: WikiPathways.....	246
N5 – STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) — Cross-reference Entry.....	248
Category O: Protein–Protein Interaction Databases.....	250
OVERVIEW .....	250
O1 – STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) .....	251
O2 – BioGRID (Biological General Repository for Interaction Datasets) .....	254
O3 – IntAct .....	257
O4 – DIP (Database of Interacting Proteins) .....	260
O5 – MINT (Molecular INTERaction database) .....	262
Category P: Drug, Compound, and Target Databases .....	265
OVERVIEW .....	265
P1 – DrugBank.....	266
P2 – ChEMBL.....	269
P3 – PubChem.....	272
P4 – BindingDB .....	275
P5 –TTD (Therapeutic Target Database) .....	277
Category Q: Ontologies and Controlled Vocabularies .....	280
OVERVIEW .....	280
Q1 – Gene Ontology (GO) .....	281
Q2 – Sequence Ontology (SO) .....	283
Q3 – Human Phenotype Ontology (HPO) .....	285
Q4 – Disease Ontology (DO) .....	287
Q5 – Uberon.....	289
MeSH .....	292
Category R: Epigenomics Databases .....	295
OVERVIEW .....	295
R1 – ENCODE (Encyclopedia of DNA Elements) .....	296
R2 – Roadmap Epigenomics .....	298
R3 – Cistrome .....	300
R4 – JASPAR (Transcription Factor Binding Profiles Database) .....	302
R5 – ChIP-Atlas .....	304
Category S: Single-cell and Spatial Omics Resources .....	307
OVERVIEW .....	307
S1 – Human Cell Atlas (HCA) .....	308
S2 – CellxGene (CELLxGENE).....	310
S3 – PanglaoDB .....	312
S4 – Single Cell Expression Atlas (SCEA) .....	314



Category T: Microbiome and Metagenomics Databases .....	317
OVERVIEW .....	317
T1 – MGnify (formerly EBI Metagenomics).....	318
T2 – SILVA .....	320
T3 – GTDB (Genome Taxonomy Database) .....	322
T4 – RDP (Ribosomal Database Project) .....	324
T5 – MG-RAST (Metagenomics Rapid Annotation using Subsystem Technology) .....	326
Short Index Entries — Category T .....	328
T6 – Greengenes2.....	328
T7 – IMG/M (Integrated Microbial Genomes and Microbiomes).....	328
Category U: Taxonomy and Organism Databases .....	330
OVERVIEW .....	330
U1 – NCBI Taxonomy.....	331
U2 – UniProt Taxonomy.....	333
U3 – Catalogue of Life .....	335
Category V: Antimicrobial Peptide and Peptide Databases.....	338
OVERVIEW .....	338
V1 – DRAMP (Data Repository of Antimicrobial Peptides).....	339
V2 – APD/APD6 (Antimicrobial Peptide Database; APD3 is historical, APD6 is the current platform) .....	341
V3 – dbAMP .....	343
V4 – CAMPR4 (Collection of Anti-Microbial Peptides Release 4; CAMP/CAMPR4 are historical names).....	345
V5 – DBAASP (Database of Antimicrobial Activity and Structure of Peptides).....	347
Category W: Cancer Genomics Databases .....	350
OVERVIEW .....	350
W1 – TCGA (The Cancer Genome Atlas).....	351
W2 – ICGC (International Cancer Genome Consortium).....	353
W3 – cBioPortal for Cancer Genomics .....	355
W4 – COSMIC (Catalogue of Somatic Mutations in Cancer) .....	357
W5 – OncoKB.....	359
Category X: Model Organism Databases .....	362
OVERVIEW .....	362
X1 – FlyBase.....	363
X2 – WormBase .....	365
X3 – MGI (Mouse Genome Informatics) .....	367
X4 – ZFIN (Zebrafish Information Network).....	369
X5 – SGD (Saccharomyces Genome Database) .....	371
X6 – TAIR (The Arabidopsis Information Resource) .....	373
X7 – PomBase.....	375
X8 – RGD (Rat Genome Database) .....	377



X9 – Xenbase .....	377
X10 – Gramene .....	377
X11 – VEuPathDB (Eukaryotic Pathogen, Vector and Host Informatics Resource) .....	378
X12 – BV-BRC (Bacterial and Viral Bioinformatics Resource Center) .....	378
Category Y: Data Standards, Repositories, and FAIR Resources .....	380
OVERVIEW .....	380
Y1 – FAIRsharing .....	381
Y2 – BioSamples .....	383
Y3 – BioProject .....	385
Y4 – BioStudies .....	387
Category Z: Database Directories and Resource Catalogs .....	390
OVERVIEW .....	390
Z1 – NAR Molecular Biology Database Collection .....	391
Z2 – bio.tools .....	393
Z3 – Database Commons .....	395
PART II — Expanded Omics, Clinical, Proteomics, Metabolomics, and Translational Data Resources .....	398
Category AA: Controlled Human Genomics and Phenotype Repositories .....	398
Category Overview .....	398
AA1 — EGA (European Genome-phenome Archive) .....	400
AA2 — dbGaP (database of Genotypes and Phenotypes) .....	403
Short Index Entries — Category AA .....	406
JGA (Japanese Genotype-phenotype Archive) .....	406
AnVIL (NHGRI Analysis, Visualization, and Informatics Lab-space) .....	406
DUOS (Data Use Oversight System) .....	406
Category AB: Proteomics Repositories and Mass Spectrometry Resources .....	408
Category Overview .....	408
Key distinctions in this category: .....	408
AB1 — ProteomeXchange Consortium .....	409
AB2 — PRIDE (PRoteomics IDentifications Database) .....	411
AB3 — PeptideAtlas .....	414
AB4 — MassIVE (Mass Spectrometry Interactive Virtual Environment) .....	417
Short Index Entries — Category AB .....	419
jPOST (Japan ProteOme STandard Repository/Database) .....	419
iProX (Integrated Proteome Resources) .....	419
Panorama Public .....	419
PASSEL (PeptideAtlas SRM Experiment Library) .....	420
ProteomicsDB .....	420
CPTAC (Clinical Proteomic Tumor Analysis Consortium) .....	421
Category AC: Metabolomics and Small-Molecule Omics Resources .....	422





Category Overview .....	422
Key distinctions in this category: .....	422
Metabolite annotation confidence levels (Metabolomics Standards Initiative): .....	422
AC1 — MetaboLights .....	423
AC2 — Metabolomics Workbench .....	426
AC3 — HMDB (Human Metabolome Database) .....	428
AC4 — GNPS (Global Natural Products Social Molecular Networking) .....	431
Short Index Entries — Category AC .....	434
MassBank .....	434
MoNA (MassBank of North America) .....	434
LipidMaps .....	434
KEGG Compound .....	435
ChEBI (Chemical Entities of Biological Interest) .....	435
Category AD: Reference Gene Annotation and Genome Annotation Resources .....	436
Category Overview .....	436
Key distinctions: .....	436
AD1 — GENCODE .....	437
AD2 — MANE (Matched Annotation from NCBI and EMBL-EBI) .....	440
AD3 — NCBI Assembly .....	442
AD4 — NCBI Datasets .....	444
Short Index Entries — Category AD .....	446
RefSeq Select .....	446
UCSC Table Browser .....	446
Ensembl BioMart .....	446
Category AE: Expression and Proteome Atlases .....	447
Category Overview .....	447
Key distinctions: .....	447
AE1 — Human Protein Atlas .....	448
AE2 — Bgee (Gene Expression Evolution Database) .....	451
AE3 — recount3 .....	454
Short Index Entries — Category AE .....	456
Tabula Sapiens .....	456
ARCHS4 (All RNA-seq and ChIP-seq Sample and Signature Search) .....	456
BioGPS .....	456
Category AF: GWAS, Rare Disease, and Clinical Variant Interpretation Resources .....	457
Category Overview .....	457
Critical distinctions: .....	457
AF1 — NHGRI-EBI GWAS Catalog .....	458
AF2 — DECIPHER (DatabasE of genom <i>i</i> C variation and Phenotype in Humans using Ensembl Resources) .....	461



AF3 — ClinGen (Clinical Genome Resource) .....	464
Short Index Entries — Category AF .....	466
LOVD (Leiden Open Variation Database) .....	466
HGMD (Human Gene Mutation Database) .....	466
CIViC (Clinical Interpretation of Variants in Cancer).....	467
Orphanet .....	467
Category AG: Ontology Ecosystem and Controlled Vocabularies.....	468
Category Overview .....	468
Critical distinctions:.....	468
AG1 — OBO Foundry (Open Biological and Biomedical Ontologies Foundry) .....	469
AG2 — MONDO Disease Ontology .....	471
AG3 — Cell Ontology (CL).....	473
AG4 — ECO (Evidence and Conclusion Ontology).....	475
AG5 — EDAM (Bioinformatics Operations, Data, Topics, and Formats Ontology) .....	477
AG6 — OLS (Ontology Lookup Service).....	479
AG7 — BioPortal.....	481
Short Index Entries — Category AG .....	483
Gene Ontology (GO) .....	483
HPO (Human Phenotype Ontology) .....	483
DOID (Disease Ontology).....	484
Uberon (Uber-anatomy Ontology).....	484
Sequence Ontology (SO).....	484
MeSH (Medical Subject Headings) .....	485
Category AH: Structural Biology Data Resources .....	486
Category Overview .....	486
Critical distinctions:.....	486
AH1 — EMDB (Electron Microscopy Data Bank).....	487
AH2 — BMRB (Biological Magnetic Resonance Data Bank) .....	489
AH3 — CATH (Class, Architecture, Topology, Homologous Superfamily).....	491
AH4 — SCOPe (Structural Classification of Proteins — extended) .....	493
Short Index Entries — Category AH .....	495
SASBDB (Small Angle Scattering Biological Data Bank).....	495
PDB-Dev (Prototype System for Archiving Integrative/Hybrid Structural Models) .....	495
ModelArchive .....	495
PDBsum.....	496
Category AI: Pharmacogenomics and Drug-Target Interaction Resources .....	497
Category Overview .....	497
Critical distinctions:.....	497
AI1 — Open Targets Platform .....	498



AI2 — PharmGKB (Pharmacogenomics Knowledgebase) .....	500
AI3 — DGIdb (Drug-Gene Interaction Database) .....	501
AI4 — DrugCentral .....	504
Short Index Entries — Category AI .....	506
IUPHAR/BPS Guide to Pharmacology .....	506
DrugBank .....	506
ChEMBL .....	507
BindingDB .....	507
Category AJ: Cancer Model Systems and Cell Line Resources .....	508
Category Overview .....	508
Critical distinctions: .....	508
AJ1 — DepMap (Cancer Dependency Map) .....	509
AJ2 — GDSC (Genomics of Drug Sensitivity in Cancer) .....	511
AJ3 — Cellosaurus .....	513
Short Index Entries — Category AJ .....	515
CCLE (Cancer Cell Line Encyclopedia) .....	515
AACR Project GENIE .....	515
ICGC/ARGO (International Cancer Genome Consortium / Accelerating Research in Genomic Oncology) .....	516
Category AK: Non-coding RNA Databases .....	517
Category Overview .....	517
Critical distinctions: .....	517
AK1 — RNAcentral .....	518
AK2 — Rfam .....	520
AK3 — miRBase .....	522
Short Index Entries — Category AK .....	524
miRGeneDB .....	524
LNCipedia .....	524
circBase .....	524
snoDB .....	525
Category AL: Comparative Genomics and Orthology Resources .....	526
Category Overview .....	526
Critical distinctions: .....	526
AL1 — OrthoDB .....	527
AL2 — eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) .....	529
AL3 — PANTHER (Protein Analysis THrough Evolutionary Relationships) .....	531
Short Index Entries — Category AL .....	533
OMA (Orthologous Matrix) .....	533
TimeTree .....	533
InParanoid .....	533



PhylomeDB.....	534
Part V: Practical Reproducibility and Quality-Control Guidelines for Database-Based Research .....	535
Introductory Note .....	535
Section 1: The Reproducibility Crisis in Bioinformatics Database Usage .....	535
Section 2: Minimum Reproducibility Requirements .....	536
Section 3: Database-Specific QC Warnings .....	537
Section 4: Applying FAIR Principles in Database-Based Research .....	539
Common Mistakes and Misconceptions .....	540
Recommended Learning Path.....	546
LEVEL A: COMPLETE BEGINNERS (No bioinformatics background).....	546
LEVEL B: BIOLOGY STUDENTS WITH SOME COMPUTATIONAL EXPOSURE .....	548
LEVEL C: EXPERIENCED RESEARCHERS NEW TO A SPECIFIC DOMAIN.....	549
References and Further Reading .....	550
PRIMARY DATABASE PAPERS.....	550
REVIEW ARTICLES AND METHODOLOGY PAPERS .....	552
Appendix: Master Comparison Table of All Databases .....	553
Appendix A: Master Database Table .....	563
Appendix B: Raw Links for Copy/Paste .....	570
Appendix C: Deprecated, Limited-Update, and Restricted Resources .....	574
Appendix D: Glossary .....	576
About This Atlas .....	578

## Start Here: How to Navigate This Atlas

This atlas is designed for practical use. Do not start by searching for a familiar database name. Start with the scientific question, identify the data type, then use the Decision Maps, category overview, database card, and comparison tables to select the appropriate resource.

### Recommended navigation path:

1. If you are unsure where to start, read the Quick Navigation and Table of Contents.
2. If you have a specific research question, go to Section 8: Decision Maps.
3. If you know the data domain, go directly to the relevant category.
4. If multiple resources seem appropriate, use the comparison tables.
5. If you will publish or report your analysis, use Part V to record reproducibility and quality-control details.
6. If a resource appears outdated, restricted, or restructured, check the deprecated/restricted resources appendix before using it.

### List of Decision Maps

DM-01: If I have a protein sequence, where should I go?

DM-02: If I have a DNA/nucleotide sequence, where should I go?

DM-03: If I want scientific papers, where should I go?

DM-04: If I want raw RNA-seq data, where should I go?

DM-05: If I want processed gene expression data, where should I go?

DM-06: If I want protein 3D structure, where should I go?

DM-07: If I want to interpret a genetic variant, where should I go?

DM-08: If I have human genomic data with patient/phenotype information, where should I deposit it?

DM-09: If I want controlled-access human genotype-phenotype data, where should I search?

DM-10: If I want pathway/functional context, where should I go?

DM-11: If I want disease-gene associations, where should I go?

DM-12: If I want antimicrobial peptide data, where should I go?

DM-13: If I want cancer genomics data, where should I go?

DM-14: If I want microbiome data, where should I go?

DM-15: If I want tissue-level protein expression, should I use GTEx or Human Protein Atlas?

DM-16: If I want cross-species expression comparison, which resource is best?

DM-17: If I want to interpret a GWAS hit, what resources should I use?

DM-18: If I want to interpret a rare disease variant, what resources should I use?

DM-19: If I want an ontology for disease, phenotype, anatomy, cell type, chemical entities, or evidence codes, where should I go?

DM-20: If I want cryo-EM maps, NMR data, SAXS/SANS structures, or predicted protein models, which structural resource should I use?

DM-21: If I want drug-target evidence, drug-gene interactions, or pharmacogenomic evidence, which resources should I use?

DM-22: If I want cancer cell-line dependency or drug-sensitivity data, where should I go?

DM-23: If I want non-coding RNA annotation, which resource should I use?

DM-24: If I want orthologs, paralogs, or protein family evolution, which resource should I use?

DM-25: If I work on a model organism, when should I use the organism-specific database instead of NCBI/Ensembl/UniProt?

DM-26: If I work on microbiome taxonomy, when should I use SILVA, GTDB, RDP, Greengenes2, MGnify, or MG-RAST?

DM-27: If I have raw mass spectrometry proteomics data, where should I deposit it?

DM-28: If I want to reanalyze public proteomics datasets, where should I go?

DM-29: If I have metabolomics LC-MS or NMR data, which repository is appropriate?

DM-30: If I want to identify a metabolite, which databases should I use?

DM-31: If I want reliable human gene annotation, should I use Ensembl, RefSeq, GENCODE, or MANE?

## Quick Navigation

This reference is organized for scientific purposes. Use this map before going into the detailed cards.

- **Front Matter** — scope, use policy, editorial standard, database-card templates, resource-type definitions, FAIR principles, and verification policy.
- **Decision Maps** — choose resources by scientific question: sequence, protein, variant, expression, structure, pathway, drug, disease, AMP, cancer, microbiome, literature, controlled-access human data, proteomics, metabolomics, gene annotation, expression/proteome atlases, GWAS, rare disease, ontology selection, structural biology archives, pharmacogenomics, cancer models, non-coding RNA, orthology, model organisms, and microbiome taxonomy.
- **Core Category Atlas** — full-depth database cards for the original A–Z resource categories.
- **Expanded Coverage Sections** — additional first-edition categories covering underrepresented domains such as controlled-access human genomics, proteomics, metabolomics, gene annotation, expression/proteome atlases, GWAS/rare disease, ontology ecosystems, structural archives, pharmacogenomics, cancer model systems, non-coding RNA, and comparative genomics.
- **Comparison Tables** — side-by-side comparisons of related resources across major domains.
- **Reproducibility Framework** — minimum reporting checklists and example methods statements.
- **Quality-Control Warnings** — critical warnings for common database misuse.
- **Appendices** — master index, raw links, deprecated/restricted resources, glossary, correction log, future roadmap, editorial QC report, and references.

## How to Use This Atlas

This atlas is designed to be used as a reference, not read cover to cover. The most efficient approach is to start with the scientific question and work toward the specific resource.

### Step 1: Identify your scientific question

What are you trying to find out? Examples: What is the function of this protein? What variants are associated with this disease? What is the expression of this gene in liver tissue? What is the 3D structure of this enzyme? What metabolites are altered in this condition?

### Step 2: Use the Decision Maps

The Decision Maps (Section 8) provide structured guidance for 31 common scientific questions. Each map identifies the appropriate starting resource and explains when to use alternatives.

### Step 3: Read the Category Overview

Each category begins with an overview explaining the conceptual landscape of that data domain, common mistakes, and when to use each type of resource within the category.

### Step 4: Read the Database Card

Full database cards provide 31 fields of information including definition, use cases, strengths, limitations, programmatic access, citations, and reproducibility notes. Index cards provide 10 fields for less central resources.

### Step 5: Check the Comparison Tables

When multiple resources could serve your purpose, the comparison tables provide side-by-side comparisons to help you choose.

### Step 6: Record your choices for reproducibility

Use Part V: Practical Reproducibility and Quality-Control Guidelines to record the database name, release/version, access date, query terms, filters, genome build, annotation release, API endpoint, and quality-control decisions used in your analysis.

For publication: cite the specific database paper (not just the portal), include the version/release number, and record the access date in your methods section.



## Section 1: Executive Summary

The Bioinformatics Databases and Data Resources Reference Atlas is a comprehensive, structured reference guide designed to help researchers, students, clinicians, and data scientists navigate the rapidly expanding landscape of biological databases, repositories, and analytical tools. Modern biology generates data at an unprecedented scale — from whole-genome sequences and transcriptomic profiles to protein structures, clinical variants, and antimicrobial peptide libraries — and the number of resources available to store, query, and interpret this data has grown correspondingly. This atlas provides a structured, curated reference that describes what each major resource is, what it contains, how to access it, and when to use it. It is intended for anyone who works with biological data, from undergraduate students encountering bioinformatics for the first time to experienced researchers who need a reliable reference when venturing into unfamiliar data domains.

Choosing the right database is not a trivial decision. Using an inappropriate resource can lead to wasted time, incorrect conclusions, or missed data. For example, submitting a protein sequence to a nucleotide database, searching for raw sequencing reads in a processed expression database, or relying on a computationally predicted annotation when manually reviewed data is available — all of these are common errors with real consequences for research quality. Different databases serve fundamentally different purposes: some store primary experimental data, others curate and integrate information from multiple sources, and still others provide analytical tools that operate on data retrieved elsewhere. Understanding these distinctions is essential for designing efficient, reproducible bioinformatics workflows. The quality, completeness, and curation level of a database directly affects the reliability of any downstream analysis.

This atlas is organized into thematic categories, each covering a distinct domain of biological data. The categories progress from general integrated portals through literature resources, nucleotide and protein sequence databases, structural databases, variant and disease resources, pathway and functional databases, expression and transcriptomics resources, and specialized domains including antimicrobial peptides, microbiome data, and single-cell genomics. Each category begins with a conceptual overview and decision guide, followed by detailed database cards that describe every major resource in that domain. The atlas also includes cross-cutting sections on how to think about databases conceptually, a decision map organized by scientific question, and practical workflow examples throughout.

To use this atlas effectively, begin with the Decision Map in Section 8 if you have a specific scientific question and need to identify the right starting point. If you are exploring a new data domain, read the category overview first to understand the landscape before examining individual database cards. Each database card follows a standardized detailed format covering purpose, content, strengths, limitations, access methods, curation level, reproducibility considerations, quality-control notes, and practical



workflow examples — use these fields to compare resources systematically. The comparison tables at the end of each category provide a quick side-by-side view of key attributes. Throughout the atlas, resources that are commercial, deprecated, or require institutional access are clearly marked. Version information and update frequencies are provided where known, as these details are critical for reproducible research.

## Section 2: Scope and Limitations

This atlas covers biological databases, repositories, knowledgebases, portals, ontologies, registries, and major analysis resources relevant to genomics, transcriptomics, proteomics, metabolomics, structural biology, variant interpretation, pathway analysis, drug discovery, cancer genomics, microbiome research, and related fields. It is not a software manual, a bioinformatics methods textbook, or a comprehensive encyclopedia of every biological database ever created.

### Scope boundaries:

- Included: Databases, repositories, knowledgebases, portals, ontologies, genome browsers, literature search engines, dataset collections, benchmark datasets, registries/catalogs, and controlled-access archives that are widely used, scientifically important, or methodologically distinct.
- Included: Resources with stable, publicly documented access (open or controlled-access with defined application procedures).
- Included: Resources with peer-reviewed documentation in major journals (Nucleic Acids Research, Nature Biotechnology, Nature Methods, Database, or equivalent).
- Excluded: Databases with no stable URL, no peer-reviewed documentation, or no clear update policy.
- Partially covered: Commercial databases are included where they are widely used in research (e.g., DrugBank, COSMIC), but their licensing restrictions are clearly noted.

### Known limitations of this atlas:

- The atlas reflects the state of the database landscape as of May 2026. Database URLs, APIs, entry counts, and access policies change frequently. Always verify current status before use.
- Entry counts and coverage statistics are approximate and based on published figures or official documentation. Exact current numbers may differ.
- Some databases in rapidly evolving fields (e.g., single-cell omics, AI-predicted structures) may have changed significantly since this edition was prepared.
- The atlas does not cover every database in any given domain. For comprehensive discovery, use the NAR Molecular Biology Database Collection, bio.tools, or Database Commons.
- Citations are to the most recent available official database papers as of May 2026. Newer papers may have been published after this edition.
- Some foundational citations are retained where they remain scientifically relevant or historically important. Users should verify the latest official database publication before formal citation in manuscripts, theses, grant proposals, or clinical reports.

## Section 3: Standard 31-Field Database Card Template

The final target structure for every major database card is the 31-field template below. Major cards should be treated as complete cards; short entries or comparison-only resources are flagged in the QC matrix as partial and should be expanded in future editions if they become central to the atlas.

Field	Field
1. Database/Resource Name	17. When to Use It
2. Official Website URL	18. When NOT to Use It
3. Correct Resource Type	19. Related Databases or Alternatives
4. Main Biological Domain	20. How It Connects to Other Resources
5. Short Definition	21. API / FTP / Bulk Download / Programmatic Access
6. What It Is Used For	22. Evidence or Curation Level
7. What Data It Contains	23. Update Status
8. Main Scientific Question It Helps Answer	24. Licensing or Access Restrictions
9. Typical Users	25. Citation / Recommended Reference
10. Example Scientific Questions	26. Beginner-Friendly Explanation
11. Example Use Cases	27. Advanced Technical Explanation
12. Input Data Accepted	28. Practical Workflow Example
13. Output Data Provided	29. Reproducibility Notes
14. Strengths	30. Quality-Control Notes
15. Limitations	31. Access Date / Verification Date
16. Common Beginner Mistakes	

## Section 4: Database and Resource Type Definitions

The terms 'database,' 'repository,' 'knowledgebase,' 'tool,' 'portal,' 'genome browser,' 'literature search engine,' 'ontology,' and 'dataset collection' are often used interchangeably in casual conversation, but they refer to meaningfully different types of resources. Understanding these distinctions is essential for designing efficient, reproducible bioinformatics workflows, for accurate citation, and for correctly interpreting the reliability and scope of any resource.

Resource Type	Definition
<b>Database</b>	A structured collection of data organized for efficient retrieval, typically with defined schemas, identifiers, and query interfaces.
<b>Repository</b>	A storage system for depositing and retrieving data objects — often raw or minimally processed — such as sequencing reads or microarray files.
<b>Knowledgebase</b>	Integrates data with curated biological knowledge, relationships, and interpretations, often drawing from multiple primary sources.
<b>Tool</b>	A software application that performs computational operations on data, such as sequence alignment or structure prediction. tools may be web-based or command-line, and they may query databases internally.
<b>Portal</b>	An integrated web interface that provides access to multiple databases and tools under a single umbrella, such as NCBI or EMBL-EBI.
<b>Genome Browser</b>	A specialized visualization tool that displays genomic features — genes, variants, regulatory elements — along a chromosomal coordinate system.
<b>Literature Search Engine</b>	A literature search engine indexes scientific publications and enables retrieval by keyword, author, journal, topic, MeSH term, citation network, DOI, or semantic similarity. It does not store biological data; it helps users find and evaluate evidence from the scientific literature.
<b>Ontology</b>	A formal, hierarchical vocabulary of biological terms and their relationships, used to standardize annotations across databases.
<b>Dataset Collection</b>	An aggregation of curated datasets, often with metadata standards, intended for reuse in benchmarking or meta-analysis.
<b>Benchmark Dataset</b>	A curated collection of biological data with known, validated properties, used to evaluate computational methods.



This section provides a concise classification table for the major resource types used throughout the atlas. Section 5 expands these definitions with detailed explanations, examples, misconceptions, and practical interpretation notes.

## Section 5: Database Types — Definitions and Distinctions

### PRIMARY DATABASE

A primary database stores original, experimentally determined biological data submitted directly by researchers or sequencing centers. The data is typically deposited as part of the publication process or as a condition of funding, and it represents the raw output of experiments — sequences, structures, expression measurements — with minimal post-submission curation. Primary databases serve as the authoritative archive of record for a given data type and are the ultimate source from which secondary databases draw their content. They are essential for data provenance: if you need to know exactly what was measured in an experiment, the primary database is where you go.

**Example:** GenBank (<https://www.ncbi.nlm.nih.gov/genbank>) is the primary nucleotide sequence database.

**Common misconception:** Many users assume that data in a primary database has been validated or reviewed for biological accuracy. In reality, primary databases perform only format and basic consistency checks; the biological interpretation of the data is the submitter's responsibility. Errors, chimeric sequences, and mislabeled entries do exist in primary databases.

### SECONDARY / CURATED DATABASE

A secondary database is built by integrating, re-annotating, and curating data drawn from one or more primary databases. Expert biocurators or automated pipelines add functional annotations, cross-references, quality filters, and standardized vocabulary to the raw data. Secondary databases sacrifice some breadth and currency (they may not include the most recently deposited sequences) in exchange for higher reliability and richer annotation. They are the preferred starting point for functional interpretation of biological data.

**Example:** UniProtKB/Swiss-Prot (<https://www.uniprot.org>) is a manually reviewed secondary protein database. Each entry is curated by expert annotators who review the primary literature and assign function, subcellular localization, post-translational modifications, and other attributes with explicit evidence codes.

**Common misconception:** Researchers sometimes assume that because a secondary database is curated, all of its entries are equally reliable. In practice, curation depth varies high-profile proteins (e.g., human TP53) may have hundreds of curated annotations, while obscure proteins from poorly studied organisms may have only computationally transferred annotations with minimal manual review.

## KNOWLEDGEBASE

---

A knowledgebase integrates biological data with curated knowledge — relationships, pathways, mechanisms, and interpretations — drawn from experimental data, literature, and expert curation. Unlike a simple database that stores data objects, a knowledgebase represents biological knowledge as a network of entities and relationships, enabling complex queries about biological functions, disease mechanisms, and molecular interactions. Knowledgebases are particularly valuable for systems biology and translational research, where understanding the context of a gene or protein is as important as knowing its sequence.

**Example:** The Gene Ontology (GO) knowledgebase (<https://geneontology.org>) provides a structured, hierarchical vocabulary of biological processes, molecular functions, and cellular components, along with curated annotations linking genes to GO terms based on experimental evidence.

**Common misconception:** Knowledgebases are sometimes confused with databases because they store data. The key distinction is that a knowledgebase explicitly represents relationships and reasoning — it is designed to answer "why" and "how" questions, not just "what" questions. A database tells you what sequences exist; a knowledgebase tells you what those sequences do and how they relate to each other.

## REPOSITORY

---

A repository is a storage and retrieval system for depositing and accessing data objects, typically with minimal processing or curation. Repositories prioritize completeness and accessibility over annotation richness. They are the primary mechanism for data sharing in compliance with journal and funder mandates. Repositories typically provide accession numbers, basic metadata, and download access, but do not add biological interpretation to the deposited data.

**Example:** The Sequence Read Archive (SRA, <https://www.ncbi.nlm.nih.gov/sra>) is a repository for raw next-generation sequencing data. Researchers deposit FASTQ or BAM files, and the repository stores them with minimal processing, making them available for download and reanalysis.

**Common misconception:** Repositories are sometimes assumed to be equivalent to databases. The distinction matters because repositories do not guarantee data quality or biological accuracy — they are archives, not curated resources. Data retrieved from a repository requires independent quality assessment before use in analysis.

## SEARCH ENGINE / SIMILARITY TOOL

---

A sequence similarity search engine is a computational tool that identifies sequences in a database that are similar to a query sequence, using algorithms such as BLAST (Basic Local Alignment Search Tool)



or HMMER. These tools are not databases themselves — they are analytical engines that operate on database content. They are among the most widely used tools in bioinformatics, enabling researchers to infer function, identify homologs, and place sequences in evolutionary context based on similarity to known sequences.

**Example:** NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>) allows users to submit a nucleotide or protein sequence and identify similar sequences in GenBank, RefSeq, or other NCBI databases, returning alignments with statistical significance scores (E-values).

**Common misconception:** Many beginners assume that a BLAST hit with a low E-value proves functional equivalence between two sequences. In reality, sequence similarity is evidence for homology, not proof of identical function. Paralogs (duplicated genes with diverged functions) can produce strong BLAST hits while having distinct biological roles.

## GENOME BROWSER

A genome browser is a specialized visualization tool that displays genomic features — genes, transcripts, regulatory elements, variants, epigenomic marks, and sequence alignments — along a chromosomal coordinate system. Genome browsers allow researchers to navigate the genome interactively, zoom in on specific regions, overlay multiple data tracks, and examine the genomic context of features of interest. They are essential for interpreting genomic data in its chromosomal context and for integrating multiple data types visually.

**Example:** The UCSC Genome Browser (<https://genome.ucsc.edu>) provides an interactive interface for exploring the human and many other genomes, with hundreds of pre-loaded annotation tracks covering genes, conservation, regulatory elements, variants, and user-uploaded custom data.

**Common misconception:** Genome browsers are sometimes confused with genome databases. A genome browser is a visualization interface; the underlying data is stored in genome databases. The same genomic data may be viewable in multiple browsers (e.g., Ensembl, UCSC, NCBI GDV), but the browsers may display different annotation tracks or use different coordinate systems, leading to apparent discrepancies.

## LITERATURE DATABASE

A literature database indexes scientific publications — journal articles, preprints, reviews, and conference proceedings — and provides search, retrieval, and citation analysis capabilities. Literature databases vary in their scope (biomedical vs. all sciences), indexing criteria (peer-reviewed only vs. including preprints), full-text availability, and analytical features (citation networks, semantic search, AI-assisted summarization). They are the primary tools for literature review, evidence synthesis, and staying current with research developments.

**Example:** PubMed (<https://pubmed.ncbi.nlm.nih.gov>) is the primary biomedical literature database maintained by NCBI, indexing more than 40 million citations and abstracts from MEDLINE and other life science journals, with links to full text where available.

**Common misconception:** Many researchers assume that PubMed indexes all biomedical literature. In practice, PubMed covers journals that meet MEDLINE indexing criteria; some legitimate journals, preprint servers, and conference proceedings are not indexed. For comprehensive literature searches, multiple databases should be consulted.

## ONTOLOGY RESOURCE

An ontology resource provides a formal, hierarchical vocabulary of biological terms and their relationships, used to standardize annotations across databases and enable computational reasoning. Ontologies define terms precisely, specify relationships between terms (e.g., "is a," "part of," "regulates"), and provide unique identifiers for each term. They are the backbone of interoperability in bioinformatics: by annotating data with ontology terms, different databases can be queried and integrated consistently.

**Example:** The Gene Ontology (GO, <https://geneontology.org>) defines terms for biological processes (e.g., "DNA repair"), molecular functions (e.g., "ATP binding"), and cellular components (e.g., "nucleus"), and provides curated annotations linking genes to these terms across thousands of species.

**Common misconception:** Ontologies are sometimes perceived as mere controlled vocabularies or thesauri. In fact, ontologies have formal logical structures that enables automated reasoning — for example, inferring that a gene annotated with "double-strand break repair" is also involved in "DNA repair" without requiring an explicit annotation to the parent term.

## INTEGRATED PORTAL

An integrated portal is a web-based platform that provides unified access to multiple databases, tools, and services under a single interface. Portals typically offer cross-database search, data integration, and links between related resources, reducing the need for researchers to navigate multiple independent websites. Major bioinformatics portals such as NCBI and EMBL-EBI host dozens of individual databases and tools, providing a coherent ecosystem for biological data access.

**Example:** NCBI (<https://www.ncbi.nlm.nih.gov>) is an integrated portal that provides access to over 40 databases including GenBank, PubMed, RefSeq, dbSNP, GEO, and SRA, along with tools such as BLAST, Primer-BLAST, and the Sequence Viewer, all accessible through a unified search interface.

**Common misconception:** Researchers sometimes treat an integrated portal as a single database and cite it as such. In reality, a portal is a collection of distinct resources, each with its own data model, update cycle, and citation. When reporting results, the specific database within the portal (e.g., "NCBI RefSeq" rather than just "NCBI") should be cited.

## BENCHMARK DATASET

A benchmark dataset is a curated collection of biological data with known, validated properties, used to evaluate the performance of computational methods. Benchmark datasets are essential for method development and comparison: they provide a standardized test set against which different algorithms can be measured objectively. High-quality benchmark datasets have clear provenance, defined positive and negative sets, and documented curation criteria.

**Example:** The Critical Assessment of Protein Structure Prediction (CASP) datasets provide experimentally determined protein structures that are withheld from public databases until after computational predictions are submitted, enabling unbiased evaluation of structure prediction methods.

**Common misconception:** Benchmark datasets are sometimes used as training data for machine learning models, which can lead to overfitting and inflated performance estimates. A benchmark dataset should be used only for evaluation, not for training, to ensure that performance estimates are unbiased.

## ANALYSIS TOOL

An analysis tool is a software application that performs computational operations on biological data — alignment, assembly, variant calling, differential expression analysis, structure prediction, and so on. Analysis tools may be web-based (accessible through a browser interface) or command-line (requiring local installation or access to a computing cluster). They are distinct from databases in that they process data rather than store it, although many tools are tightly integrated with specific databases.

**Example:** HMMER (<https://hmmer.org>) is a profile hidden Markov model-based tool for sensitive sequence similarity searches. It builds statistical models of protein families from multiple sequence alignments and uses these models to search sequence databases for distant homologs.

**Common misconception:** Analysis tools are sometimes confused with databases because they are accessed through the same web portals. The distinction matters for citation and reproducibility: when reporting results, both the tool (with version number) and the database it searched (with version) must be cited separately.

## COMMERCIAL OR SEMI-COMMERCIAL TOOL

A commercial or semi-commercial tool is a resource that requires payment, institutional subscription, or licensing for full access. Some commercial tools offer limited free tiers for academic use, while others require institutional agreements. Commercial resources are common in literature databases (Scopus, Web of Science), sequence analysis software (CLC Genomics Workbench, Geneious), and clinical genomics platforms. They are marked throughout this atlas to alert users who may not have access.

**Example:** Scopus (<https://www.scopus.com>) is a commercial abstract and citation database operated by Elsevier, requiring institutional subscription for full access. It provides broader coverage of non-English and non-biomedical literature than PubMed, but its commercial nature limits accessibility for researchers without institutional subscriptions.

**Common misconception: Commercial tools are sometimes assumed to be superior to free alternatives simply because they cost money. In practice, many open-source and freely available tools (e.g., BLAST, HMMER, Bioconductor packages) are scientifically equivalent or superior to commercial alternatives for most bioinformatics tasks. The choice between commercial and free tools should be based on specific feature requirements, not cost alone.**

Resource Type	Definition	Example	Common Misconception
Primary Database	Stores original, experimentally determined biological data submitted directly by researchers; minimal post-submission curation beyond format validation.	GenBank, PDB, GEO, SRA	Primary records are not automatically biologically validated.
Secondary / Curated Database	Integrates, re-annotates, and curates data from primary sources using expert curation or automated pipelines.	UniProtKB/Swiss-Prot, Pfam, KEGG	Curated does not mean every entry has equal evidence depth.
Repository	Stores and retrieves data objects, often raw or minimally processed; prioritizes archiving and access.	SRA, ENA, PRIDE, MetaboLights, EGA	A repository is not necessarily a curated biological knowledgebase.
Knowledgebase	Integrates data with curated biological knowledge, relationships, pathways, mechanisms, and interpretations.	Gene Ontology, Reactome, PharmGKB, ClinGen	A knowledgebase is not just a large table; it represents relationships and evidence.
Tool	Performs computational operations on data; may query databases internally.	BLAST, HMMER, Ensembl VEP, AlphaFold2	A tool is a database. Tools operate on data; databases store data. Many tools query databases internally, but the tool itself is not the database.
Portal	Provides unified access to multiple databases, tools, and services.	NCBI, EMBL-EBI, ExPASy, Open Targets	Do not cite the portal when a specific database was used.
Genome Browser	Visualizes genomic features along chromosomal coordinates.	UCSC Genome Browser, Ensembl, NCBI GDV, IGV	A browser is a visualization interface, not the underlying data source itself.

Resource Type	Definition	Example	Common Misconception
Literature Search Engine	Indexes scientific publications and enables retrieval by keyword, author, journal, topic, MeSH term, citation network, DOI, or semantic similarity. Does not store biological data.	PubMed, Europe PMC, Semantic Scholar, Google Scholar	It indexes evidence sources; it does not replace biological databases.
Ontology	Formal vocabulary of biological terms and relationships used for standard annotations and reasoning.	Gene Ontology, HPO, MONDO, Cell Ontology, ChEBI	Ontologies are not simple keyword lists.
Dataset Collection	An aggregation of curated or reprocessed datasets, often with metadata standards, intended for reuse in benchmarking, meta-analysis, or cross-study comparison.	recount3, ARCHS4, Tabula Sapiens	A dataset collection is usually secondary, not a primary archive.
Benchmark Dataset	Curated dataset with known properties used to evaluate computational methods.	CASP targets, CAFA benchmark sets	Using a benchmark for training invalidates unbiased evaluation.
Controlled-Access Archive	Repository for sensitive data requiring formal access approval and consent compliance.	EGA, dbGaP, JGA, AnVIL	Not just a slower public repository; it exists for privacy and ethics.
Registry / Catalog	Curated index of resources, tools, databases, or standards; usually does not store biological data itself.	NAR DB Collection, bio.tools, FAIRsharing, Database Commons	A registry catalogs resources; it does not contain the described biological data.
Integrated Portal	A web-based platform providing unified access to multiple databases, tools, and services, often with cross-database search and data integration capabilities.	NCBI, EMBL-EBI, Open Targets Platform	Citing a portal is sufficient for publication. When reporting results, the specific database within the portal must be cited, not just the portal.

## Section 6: FAIR and Reproducibility Principles

The FAIR principles — Findable, Accessible, Interoperable, and Reusable — were formally articulated in a 2016 paper by Wilkinson et al. in *Scientific Data* and have since been adopted by major funding agencies, journals, and data repositories worldwide. When choosing a database for data deposition or retrieval, assessing its FAIR compliance is a practical way to evaluate its long-term utility and the reproducibility of any analysis built upon it.

### The Four FAIR Principles

- **Findable:** Data and metadata are assigned globally unique and persistent identifiers (e.g., accession numbers, DOIs). Data is described with rich metadata and indexed in searchable resources.
- **Accessible:** Data is retrievable by its identifier using open, standardized protocols (e.g., HTTP, FTP). Metadata remains accessible even when data is no longer available. Access conditions are clearly stated.
- **Interoperable:** Data uses formal, shared, and broadly applicable languages for knowledge representation. Data uses vocabulary that follows FAIR principles (e.g., ontologies). Data includes qualified references to other data.
- **Reusable:** Data is richly described with accurate and relevant attributes. Data is released with a clear and accessible data usage license. Data meets domain-relevant community standards.

### Primary vs. Secondary Databases: Data Provenance

Understanding whether you are working with primary or secondary data is essential for assessing the confidence level of any annotation. A primary database stores original, experimentally determined data submitted directly by researchers. A secondary database is built by integrating, curating, and annotating data drawn from primary sources. Secondary databases are generally more reliable for functional interpretation but may lag primary databases in coverage of newly deposited data.

### Curation Levels

Curation level is one of the most important attributes of any biological database, and it varies enormously across resources:

- **Manually reviewed (expert-curated):** Expert biocurators read primary literature, evaluate experimental evidence, and assign annotations with explicit evidence codes. Examples: UniProtKB/Swiss-Prot, OMIM, ClinGen. Highly reliable but slow and limited in coverage.
- **Computationally predicted:** Generated by automated pipelines — sequence similarity searches, machine learning models, or rule-based systems. Can cover millions of sequences rapidly but with higher error rates. Examples: UniProtKB/TrEMBL, InterPro automated annotations.



- **Community-submitted:** User-contributed data with variable quality. Examples: Wikipedia-style databases, user-contributed pathway databases. Broad coverage but requires critical evaluation.
- **Mixed:** Combines automated prediction with selective manual curation of high-priority entries. Most modern databases use this approach.

## Data Versioning

---

Data versioning is a critical but frequently overlooked aspect of reproducible bioinformatics. Databases are not static: sequences are updated, annotations are revised, entries are merged or split, and entire datasets are replaced as new evidence accumulates. A gene annotation that was correct in Ensembl release 95 may differ from the annotation in release 110. For reproducible research, always record the database version, release number, and access date used in any analysis.



## Section 7: Database Release and Verification Policy

Every database card in this atlas includes an Access Date / Verification Date field recording when the card was last checked against official documentation. This policy ensures transparency about the currency of database information. Because biological databases change frequently, users must record database releases, access dates, API endpoints, query parameters, and licensing conditions whenever database-derived information is used in research, teaching, reporting, or publication.

### Minimum Information for Every Database Card

---

- Database/resource name (official name, not informal abbreviation)
- Official website URL (verified as of the access date)
- Resource type (from the controlled vocabulary in Section 4)
- Version or release number where applicable
- Access date / verification date
- Whether the resource has stable archived releases
- API, FTP, or download portal URL where available
- Licensing or access restrictions
- Citation to the most recent official database paper

### Verification Status Codes Used in This Atlas

---

**Verified:** Card content confirmed against official documentation as of May 2026.

**Requires verification:** Specific claim could not be confirmed from available documentation; marked for future verification.

**Deprecated:** Resource is no longer actively maintained; see Appendix C for details.

**Restricted:** Resource requires subscription, institutional access, or data access committee approval.

**Uncertain status:** Resource operational status could not be confirmed; verify before use.

### API Stability Warning

---

**WARNING:** API endpoints, FTP paths, and download URLs change frequently. All API and FTP information in this atlas was verified as of May 2026. Always verify the current endpoint before building automated pipelines. Record the exact endpoint, query parameters, and date in your methods.



## Citation Policy

---

For every database used in a publication:

1. Cite the most recent official database paper (typically the NAR Database Issue paper). Do not cite only the portal or website URL.
2. Include the database version or release number in the methods section.
3. Include the access date.
4. For controlled-access data, include the data access approval number.
5. For APIs, include the endpoint URL and query parameters.
6. Do not cite only the umbrella portal when a specific database within the portal was used.

## Section 8: Decision Maps — Start with the Scientific Question

Use these decision maps to identify the appropriate starting resource for your scientific question. Each map provides a primary recommendation, alternatives, and guidance on when to use each option. Always read the full database card before committing to a resource.

**NOTE:** Decision maps provide starting points, not definitive answers. The best resource depends on your specific question, organism, data type, and analysis context. When in doubt, consult the relevant category overview and comparison tables.

### DM-01: IF I HAVE A PROTEIN SEQUENCE, WHERE SHOULD I GO?

If you have a protein sequence and want to identify what it is or what it does, begin with NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>) using the BLASTp option to search against the non-redundant (nr) protein database or UniProtKB. A strong hit (E-value < 1e-5, identity > 30% over a significant alignment length) suggests homology to a characterized protein. For more sensitive searches against distantly related proteins, use HMMER (<https://www.ebi.ac.uk/Tools/hmmer>) to search against Pfam or other profile databases. Once you have identified the protein or its family, go to UniProtKB/Swiss-Prot (<https://www.uniprot.org>) for manually reviewed functional annotations, and to the Protein Data Bank (<https://www.rcsb.org>) if you want structural information.

### DM-02: IF I HAVE A DNA/NUCLEOTIDE SEQUENCE, WHERE SHOULD I GO?

For a DNA or nucleotide sequence, NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>) with the BLASTn option against the nucleotide (nt) database is the standard first step for identifying the sequence or finding similar sequences. If the sequence is from a genome and you want to know what genes or features it contains, use a genome browser such as the UCSC Genome Browser (<https://genome.ucsc.edu>) or Ensembl (<https://www.ensembl.org>) to map it to a reference genome. For submitting a new sequence to the public record, use GenBank (<https://www.ncbi.nlm.nih.gov/genbank>) or ENA (<https://www.ebi.ac.uk/ena>). If the sequence is a coding sequence and you want to predict the protein it encodes, use the NCBI ORF Finder or translate it and proceed with protein analysis tools.

### DM-03: IF I WANT SCIENTIFIC PAPERS, WHERE SHOULD I GO?

For biomedical literature, PubMed (<https://pubmed.ncbi.nlm.nih.gov>) is the primary resource and should be the first stop for most life science literature searches. For full-text access to open-access articles, use PubMed Central (<https://www.ncbi.nlm.nih.gov/pmc>) or Europe PMC (<https://europepmc.org>). For broader coverage including non-biomedical sciences, preprints, and grey literature, Google Scholar (<https://scholar.google.com>) provides the widest scope but with less rigorous indexing. For citation

analysis and impact metrics, Scopus or Web of Science are the standard tools, though both require institutional subscriptions. Semantic Scholar (<https://www.semanticscholar.org>) offers AI-assisted literature discovery and is freely accessible.

## DM-04: IF I WANT RAW RNA-SEQ DATA, WHERE SHOULD I GO?

---

Raw RNA-seq data (FASTQ files) is primarily archived in the Sequence Read Archive (SRA, <https://www.ncbi.nlm.nih.gov/sra>) at NCBI, the European Nucleotide Archive (ENA, <https://www.ebi.ac.uk/ena>), or the DDBJ Sequence Read Archive (DRA). These three archives are part of the INSDC and mirror each other's data, so the same dataset is accessible from any of the three. SRA is the most commonly used entry point for North American researchers, while ENA is often preferred for its more user-friendly download interface. Use the SRA Toolkit or the ENA browser download tools to retrieve FASTQ files for reanalysis.

**WARNING: SRA/ENA/DRA contain public sequencing data only. For controlled-access human genomic data with phenotype information, use EGA or dbGaP (see DM-08).**

## DM-05: IF I WANT PROCESSED GENE EXPRESSION DATA, WHERE SHOULD I GO?

---

For processed gene expression data — normalized count matrices, differential expression results, or summarized expression profiles — the Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo>) is the primary resource, hosting data from microarray, RNA-seq, and other expression profiling experiments. ArrayExpress (now integrated into BioStudies at EMBL-EBI, <https://www.ebi.ac.uk/arrayexpress>) is the European equivalent and mirrors much of the same data. For pre-computed, uniformly processed expression data across many experiments and species, the Expression Atlas (<https://www.ebi.ac.uk/gxa>) provides a curated resource with consistent analysis pipelines. For tissue-specific expression in humans, GTEx (<https://gtexportal.org>) is the definitive resource. For protein-level expression: Human Protein Atlas (<https://www.proteinatlas.org>) — see DM-15. For cross-species expression: Bgee (<https://www.bgee.org>) — see DM-16.

## DM-06: IF I WANT PROTEIN 3D STRUCTURE, WHERE SHOULD I GO?

---

The Protein Data Bank (PDB, <https://www.rcsb.org>) is the single global archive for experimentally determined 3D structures of proteins, nucleic acids, and their complexes, determined by X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy. For structures predicted by computational methods, AlphaFold DB (<https://alphafold.ebi.ac.uk>) provides high-confidence predicted structures for hundreds of millions of proteins. The PDBe (<https://www.ebi.ac.uk/pdbe>) and PDBj (<https://pdbj.org>) are European and Japanese mirrors of the PDB with additional analysis tools. For

structure visualization, PyMOL, UCSF ChimeraX, or the web-based Mol\* viewer (embedded in the PDB website) are standard tools.

**WARNING: Do not treat predicted protein structures as equivalent to experimentally solved structures. AlphaFold pLDDT scores indicate per-residue confidence; regions with pLDDT < 50 are likely disordered and should not be used for structural interpretation.**

## DM-07: IF I WANT TO INTERPRET A GENETIC VARIANT, WHERE SHOULD I GO?

For interpreting human genetic variants, begin with ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar>) for clinically interpreted variants with pathogenicity classifications, and dbSNP (<https://www.ncbi.nlm.nih.gov/snp>) for a comprehensive catalog of known variants. For population frequency data — essential for distinguishing rare pathogenic variants from common benign polymorphisms — use gnomAD (<https://gnomad.broadinstitute.org>) or the 1000 Genomes Project data. For functional impact prediction, tools such as SIFT, PolyPhen-2, and CADD provide computational pathogenicity scores. For cancer-specific somatic variants, COSMIC (<https://cancer.sanger.ac.uk/cosmic>) and cBioPortal (<https://www.cbioportal.org>) are the primary resources. For functional impact prediction: Ensembl VEP (<https://www.ensembl.org/vep>) integrates multiple prediction tools. For cancer somatic variants: COSMIC (<https://cancer.sanger.ac.uk/cosmic>).

**WARNING: Do not use population frequency databases as clinical pathogenicity databases. A variant being rare does not make it pathogenic. Do not treat computational pathogenicity predictions (SIFT, PolyPhen-2, CADD) as clinical diagnosis.**

## DM-08: If I have human genomic data with patient/phenotype information, where should I deposit it?

Primary recommendation: EGA (European Genome-phenome Archive, <https://ega-archive.org>) for European researchers; dbGaP (<https://www.ncbi.nlm.nih.gov/gap>) for US-funded research.

For Japanese researchers: JGA (Japanese Genotype-phenotype Archive, <https://www.ddbj.nig.ac.jp/jga/>). For NIH-funded cloud-based analysis: AnVIL (<https://anvilproject.org>).

**WARNING: Do NOT deposit sensitive human genomic data with phenotype information in public repositories (SRA, ENA, GEO). This violates participant consent, data access agreements, and in many jurisdictions, data protection law. Controlled-access archives (EGA, dbGaP) provide the appropriate infrastructure for human subject data.**

Key requirements for controlled-access deposition:

- Obtain institutional review board (IRB) or ethics committee approval.

- Confirm that participant consent covers data sharing.
- Establish a Data Access Committee (DAC) to review access requests.
- Define consent restrictions (e.g., disease-specific use, no commercial use).
- Register the study and obtain accession numbers before submission.

## DM-09: If I want controlled-access human genotype-phenotype data, where should I search?

Primary recommendation: EGA (<https://ega-archive.org>) for European cohorts; dbGaP (<https://www.ncbi.nlm.nih.gov/gap>) for US cohorts.

Process: Submit a data access request to the relevant Data Access Committee (DAC). Approval typically requires institutional affiliation, research purpose statement, and data use agreement.

For GWAS summary statistics (not individual-level data): GWAS Catalog (<https://www.ebi.ac.uk/gwas>) provides open-access summary statistics for many studies.

**WARNING: Controlled-access data requires formal approval before access. Do not attempt to access controlled-access data without an approved data access request. Violations can result in loss of access privileges and legal consequences.**

## DM-10: IF I WANT PATHWAY/FUNCTIONAL CONTEXT, WHERE SHOULD I GO?

For pathway and functional context, KEGG (<https://www.kegg.jp>) provides manually curated metabolic and signaling pathway maps with gene-to-pathway mappings across many organisms. Reactome (<https://reactome.org>) offers a freely accessible, manually curated pathway database with detailed molecular mechanisms and strong integration with other resources. The Gene Ontology (<https://geneontology.org>) provides functional annotations at the level of molecular function, biological process, and cellular component. For protein-protein interaction networks, STRING (<https://string-db.org>) integrates experimental and predicted interactions. For enrichment analysis of gene lists, tools such as GSEA, g:Profiler, or Enrichr use these pathway and ontology resources as their underlying databases. For pathway enrichment analysis: g:Profiler (<https://biit.cs.ut.ee/gprofiler>), DAVID (<https://david.ncifcrf.gov>), or clusterProfiler (R package).

## DM-11: IF I WANT DISEASE-GENE ASSOCIATIONS, WHERE SHOULD I GO?

For Mendelian disease-gene associations, OMIM (<https://www.omim.org>) is the authoritative resource for Mendelian disease genetics, providing manually curated entries for genes and phenotypes with detailed clinical and molecular information. DisGeNET (<https://www.disgenet.org>) provides a broader, computationally integrated resource covering associations from multiple sources including GWAS,

literature mining, and clinical databases. The GWAS Catalog (<https://www.ebi.ac.uk/gwas>) catalogs genome-wide association study results for complex traits and diseases. GeneCards (<https://www.genecards.org>) provides an integrated summary of disease associations for individual genes. For rare diseases specifically, Orphanet (<https://www.orpha.net>) is the primary European resource. For drug target evidence: Open Targets Platform (<https://platform.opentargets.org>) — see DM-21.

## DM-12: IF I WANT ANTIMICROBIAL PEPTIDE DATA, WHERE SHOULD I GO?

For antimicrobial peptide (AMP) data, the APD/APD6 (Antimicrobial Peptide Database, <https://aps.unmc.edu>) is one of the most comprehensive manually curated resources, covering natural AMPs with activity data, structural information, and physicochemical properties. DRAMP (Data Repository of Antimicrobial Peptides, <https://dramp.cpu-bioinfor.org>) provides a broader collection including both natural and synthetic peptides. CAMPR4 (<https://camp.bicnirrh.res.in>) focuses on prediction and classification of AMPs. For peptide sequence similarity searches, BLAST against these specialized databases or the broader UniProtKB is appropriate. When working with AMPs, note that activity data is highly condition-dependent (organism, assay conditions, MIC values) and should be interpreted carefully. For structure-activity data: DBAASP (<https://dbaasp.org>).

**WARNING: AMP activity values are not comparable unless assay conditions, organism, medium, MIC definition, and peptide form are identical. Do not compare MIC values across databases without verifying assay conditions.**

## DM-13: IF I WANT CANCER GENOMICS DATA, WHERE SHOULD I GO?

For cancer genomics data, The Cancer Genome Atlas (TCGA, <https://portal.gdc.cancer.gov>) is the primary resource for multi-omic data across 33 cancer types, accessible through the GDC Data Portal. cBioPortal (<https://www.cbioportal.org>) provides a user-friendly interface for exploring TCGA and other cancer genomics datasets, with tools for mutation analysis, copy number analysis, and survival analysis. COSMIC (<https://cancer.sanger.ac.uk/cosmic>) is the most comprehensive catalog of somatic mutations in human cancer, with manually curated data from the literature. For pediatric cancers, the Pediatric Cancer Data Commons (<https://commons.cri.uchicago.edu/pcdc>) and St. Jude Cloud (<https://www.stjude.cloud>) are important resources. The International Cancer Genome Consortium (ICGC) data portal (<https://dcc.icgc.org>) provides access to cancer genome data from projects worldwide. For cancer cell-line data: DepMap (<https://depmap.org>) and CCLE — see DM-22. For drug sensitivity: GDSC (<https://www.cancerrxgene.org>) — see DM-22.

## DM-14: If I want microbiome data, where should I go?

---

For processed metagenomics data: MGnify (<https://www.ebi.ac.uk/metagenomics>).

For 16S rRNA taxonomy: SILVA (<https://www.arb-silva.de>) — most comprehensive and actively maintained.

For genome-resolved taxonomy: GTDB (<https://gtdb.ecogenomic.org>) — phylogenetically consistent bacterial/archaeal taxonomy.

For raw metagenomics data: SRA/ENA (public) or EGA (controlled-access human microbiome).

**WARNING: Taxonomy assignments differ between SILVA, GTDB, RDP, and Greengenes2. Always specify which taxonomy database and version was used. Do not mix taxonomy assignments from different databases in the same analysis.**

## DM-15: If I want tissue-level protein expression, should I use GTEx or Human Protein Atlas?

---

Use GTEx (<https://gtexportal.org>) when: You want mRNA expression data across human tissues from healthy adult donors; you want eQTL data linking genetic variants to expression; you want tissue-specific splicing data.

Use Human Protein Atlas (<https://www.proteinatlas.org>) when: You want protein-level expression data (antibody-based immunohistochemistry or mass spectrometry); you want subcellular localization data; you want single-cell expression data; you want cancer tissue expression.

Key distinction: GTEx measures mRNA (transcriptomics); Human Protein Atlas measures protein (proteomics/immunohistochemistry). mRNA and protein levels do not always correlate due to post-transcriptional regulation.

**WARNING: Do not compare gene expression datasets without checking normalization, tissue definitions, batch effects, platform, and metadata. GTEx and Human Protein Atlas use different tissues, donors, and measurement technologies.**

## DM-16: If I want cross-species expression comparison, which resource is best?

---

Primary recommendation: Bgee (<https://www.bgee.org>) — integrates expression data across species using anatomical ontologies for cross-species comparison.

For human-focused multi-tissue: GTEx (<https://gtexportal.org>) or Human Protein Atlas.

For uniformly processed multi-species: Expression Atlas (<https://www.ebi.ac.uk/gxa>).

For large-scale reprocessed human RNA-seq: recount3 (<https://rna.recount.bio>) or ARCHS4 (<https://maayanlab.cloud/archs4/>).



**WARNING: Cross-species expression comparison requires careful attention to ortholog mapping, tissue homology, developmental stage, and normalization. Do not assume that expression patterns are directly comparable across species without accounting for these factors.**

### DM-17: If I want to interpret a GWAS hit, what resources should I use?

---

Step 1 — Find the association: GWAS Catalog (<https://www.ebi.ac.uk/gwas>) for published GWAS results.

Step 2 — Check population frequency: gnomAD (<https://gnomad.broadinstitute.org>) for allele frequencies.

Step 3 — Identify candidate genes: Ensembl VEP for functional annotation; GTEx for eQTL evidence; Open Targets for target prioritization.

Step 4 — Check clinical significance: ClinVar for known pathogenicity; OMIM for Mendelian disease context.

Step 5 — Explore disease context: DECIPHER (<https://www.deciphergenomics.org>) for rare disease variants; DisGeNET for disease-gene associations.

**WARNING: GWAS associations identify statistical associations between variants and traits, not causal variants or causal genes. A GWAS hit may be in linkage disequilibrium with the causal variant. Do not assume the lead SNP is the causal variant.**

### DM-18: If I want to interpret a rare disease variant, what resources should I use?

---

Primary resources: ClinVar (pathogenicity classifications), OMIM (Mendelian disease genetics), ClinGen (clinical validity), gnomAD (population frequency).

For rare disease phenotype context: Orphanet (<https://www.orpha.net>), DECIPHER (<https://www.deciphergenomics.org>).

For variant databases: LOVD (<https://www.lovd.nl>) for locus-specific variant databases; HGMD (restricted access) for comprehensive variant catalog.

For variant interpretation guidelines: ACMG/AMP 2015 guidelines (Richards et al., Genetics in Medicine, 2015).

**WARNING: Do not treat computational pathogenicity predictions (SIFT, PolyPhen-2, CADD, REVEL) as clinical diagnosis. These tools provide supporting evidence, not definitive pathogenicity classification. Clinical variant interpretation requires integration of multiple lines of evidence following ACMG/AMP guidelines.**



## DM-19: If I want an ontology for disease, phenotype, anatomy, cell type, chemical entities, or evidence codes, where should I go?

Disease: MONDO Disease Ontology (<https://mondo.monarchinitiative.org>) for cross-ontology harmonized disease terms; Disease Ontology (<https://disease-ontology.org>) for human diseases.

Phenotype: HPO (<https://hpo.jax.org>) for human phenotypes; MP (Mammalian Phenotype Ontology) for mouse phenotypes.

Anatomy: Uberon (<https://obophenotype.github.io/uberont>) for cross-species anatomy; FMA for human anatomy.

Cell type: Cell Ontology (<https://cell-ontology.github.io>) for cell type classification.

Chemical entities: ChEBI (<https://www.ebi.ac.uk/chebi>) for chemical entities of biological interest.

Evidence codes: ECO (Evidence and Conclusion Ontology, <https://evidenceontology.org>) for standardized evidence annotation. Gene function: Gene Ontology (<https://geneontology.org>) for molecular function, biological process, cellular component.

For ontology discovery: OLS (Ontology Lookup Service, <https://www.ebi.ac.uk/ols4>) or BioPortal (<https://bioportal.bioontology.org>).

## DM-20: If I want cryo-EM maps, NMR data, SAXS/SANS structures, or predicted protein models, which structural resource should I use?

Cryo-EM maps (density maps, not atomic models): EMDB (<https://www.ebi.ac.uk/emdb/>) — the global archive for electron microscopy density maps.

NMR chemical shifts and constraints: BMRB (<https://bmrb.io>) — Biological Magnetic Resonance Data Bank.

SAXS/SANS low-resolution solution structures: SASBDB (<https://www.sasbdb.org>) — Small Angle Scattering Biological Data Bank.

Integrative/hybrid models: PDB-Dev (<https://pdb-dev.wwpdb.org>) — for structures determined by integrative methods.

Predicted models (not deposited in PDB): ModelArchive (<https://modelarchive.org>) — for computational models.

AI-predicted structures: AlphaFold DB (<https://alphafold.ebi.ac.uk>) for AlphaFold2 predictions.

Domain classification: CATH (<https://www.cathdb.info>) or SCOPe (<https://scop.berkeley.edu>) for structural domain classification.

## DM-21: If I want drug-target evidence, drug-gene interactions, or pharmacogenomic evidence, which resources should I use?

For integrated drug-target evidence: Open Targets Platform (<https://platform.opentargets.org>) — integrates genetic, genomic, and clinical evidence for target prioritization.

For pharmacogenomic variants: PharmGKB (<https://www.pharmgkb.org>) — curated pharmacogenomic knowledge.

For drug-gene interactions: DGIdb (<https://www.dgidb.org>) — aggregated drug-gene interaction database.

For drug mechanisms and targets: DrugCentral (<https://drugcentral.org>) — curated drug information.

For receptor pharmacology: IUPHAR/BPS Guide to Pharmacology (<https://www.guidetopharmacology.org>).

For bioactivity data: ChEMBL (<https://www.ebi.ac.uk/chembl>) — large-scale bioactivity database.

For cancer variant actionability: CIViC (<https://civicdb.org>) — clinical interpretation of variants in cancer.

## DM-22: If I want cancer cell-line dependency or drug-sensitivity data, where should I go?

For CRISPR dependency screens: DepMap (<https://depmap.org>) — Cancer Dependency Map with CRISPR and RNAi screens. For cell-line genomic profiles: CCLE (Cancer Cell Line Encyclopedia, <https://sites.broadinstitute.org/ccle>) — integrated with DepMap. For drug sensitivity: GDSC (Genomics of Drug Sensitivity in Cancer, <https://www.cancerrxgene.org>) — drug response data for 1,000+ cell lines. For cell-line identity: Cellosaurus (<https://www.cellosaurus.org>) — authoritative cell-line registry.

**WARNING: Cell-line data has significant translational limitations. Cell lines may not accurately represent the tumor microenvironment, clonal heterogeneity, or in vivo drug pharmacokinetics. Always validate cell-line findings in more physiologically relevant models. Verify cell-line identity using STR profiling; misidentified and contaminated cell lines are a documented problem in cancer research.**

## DM-23: If I want non-coding RNA annotation, which resource should I use?

For integrated ncRNA sequences: RNAcentral (<https://rnacentral.org>) — integrates ncRNA sequences from all major databases. For RNA families and covariance models: Rfam (<https://rfam.org>) — RNA family database. For miRNA: miRBase (<https://www.mirbase.org>) — primary miRNA registry;

miRGeneDB (<https://mirgenedb.org>) for high-confidence miRNA genes. For lncRNA: LNCipedia (<https://lncipedia.org>) or NONCODE (<http://www.noncode.org>).

**WARNING: miRNA annotation quality varies significantly. miRBase contains many low-confidence entries. For high-confidence miRNA genes, use miRGeneDB. For lncRNA, note that the catalog is rapidly evolving and many lncRNAs have uncertain functional significance.**

### DM-24: If I want orthologs, paralogs, or protein family evolution, which resource should I use?

For hierarchical ortholog groups: OrthoDB (<https://www.orthodb.org>) — hierarchical ortholog database. For functional ortholog annotation: eggNOG (<http://eggnog5.embl.de>) — evolutionary genealogy of genes. For pairwise orthologs: OMA (<https://omabrowser.org>) — Orthologous MATrix. For phylogenetic trees: TimeTree (<https://timetree.org>) — timetree of life. For functional annotation transfer: PANTHER (<https://www.pantherdb.org>) — protein family and function.

**WARNING: Do not assume the best BLAST hit equals an ortholog. Orthologs are genes related by speciation; paralogs are related by duplication. Functional annotation transfer from paralogs can be misleading. Always use a dedicated ortholog database rather than BLAST alone for ortholog identification.**

### DM-25: If I work on a model organism, when should I use the organism-specific database instead of NCBI/Ensembl/UniProt?

Use the organism-specific database when: You need curated phenotype data for mutant alleles; you need genetic interaction data; you need organism-specific gene nomenclature; you need community-curated functional annotations not yet in general databases.

Key organism-specific databases: FlyBase (Drosophila), WormBase (C. elegans), MGI (mouse), ZFIN (zebrafish), SGD (S. cerevisiae), TAIR (Arabidopsis), PomBase (S. pombe), RGD (rat), Xenbase (Xenopus), Gramene/MaizeGDB (plants), VEuPathDB (eukaryotic pathogens), BV-BRC (bacteria/viruses).

**WARNING: Do not assume model organism databases are redundant with NCBI or Ensembl. Model organism databases contain deeper curated phenotypes, allele, and genetic interaction data that is not available in general databases. For organism-specific research, the organism-specific database is often an authoritative source.**

## DM-26: If I work on microbiome taxonomy, when should I use SILVA, GTDB, RDP, Greengenes2, MGnify, or MG-RAST?

SILVA (<https://www.arb-silva.de>): Use for 16S/18S/23S/28S rRNA-based taxonomy; most comprehensive and actively maintained rRNA database; widely used for amplicon sequencing.

GTDB (<https://gtdb.ecogenomic.org>): Use for genome-resolved metagenomics and whole-genome-based taxonomy; phylogenetically consistent bacterial and archaeal taxonomy; preferred for MAG (metagenome-assembled genome) classification.

RDP (<https://rdp.cme.msu.edu>): Use with caution; limited updates since ~2014; SILVA is preferred for new analyses.

Greengenes2 (<https://greengenes2.ucsd.edu>): Updated replacement for the original Greengenes; use for 16S amplicon analysis when Greengenes compatibility is required.

MGnify (<https://www.ebi.ac.uk/metagenomics>): Use for accessing processed metagenomics datasets and functional annotations from the EBI metagenomics pipeline.

MG-RAST (<https://www.mg-rast.org>): Use for depositing and analyzing metagenomics data; verify current operational status before use.

**WARNING: Taxonomy assignments differ between SILVA, GTDB, RDP, and Greengenes2. Do not mix taxonomy assignments from different databases. Always specify the taxonomy database, version, and classifier used. Taxonomy is not stable across database versions.**

## DM-27: If I have raw mass spectrometry proteomics data, where should I deposit it?

Primary recommendation: PRIDE (<https://www.ebi.ac.uk/pride>) — the primary European proteomics repository and the most widely used ProteomeXchange member. For US-based deposition: MassIVE (<https://massive.ucsd.edu>) — UCSD-based ProteomeXchange member. For targeted proteomics (SRM/MRM): PASSEL (<http://www.peptideatlas.org/passel/>) — PeptideAtlas SRM Experiment Library. For DIA proteomics: Panorama Public (<https://panoramaweb.org/home/project-begin.view>) — Skyline-based DIA data repository.

All ProteomeXchange members assign PXD accession numbers for citation.

**WARNING: Proteomics data deposition is required by most major journals. Submit raw files (vendor format or mzML), processed results (mzIdentML, mzTab), and complete metadata. Do not submit only processed results without raw data.**

## DM-28: If I want to reanalyze public proteomics datasets, where should I go?

Primary recommendation: PRIDE (<https://www.ebi.ac.uk/pride>) for the largest collection of public proteomics datasets. For peptide-level evidence: PeptideAtlas (<http://www.peptideatlas.org>) — reanalyzed and integrated peptide evidence. For protein expression profiles: ProteomicsDB (<https://www.proteomicsdb.org>) — human proteome expression database. For cancer proteomics: CPTAC (<https://proteomics.cancer.gov/programs/cptac>) — Clinical Proteomic Tumor Analysis Consortium.

**WARNING: Proteomics datasets vary significantly in quality, instrument type, sample preparation, and search parameters. Always check the original publication for experimental details before reanalysis. Reanalysis with different search parameters can produce substantially different results.**

## DM-29: If I have metabolomics LC-MS or NMR data, which repository is appropriate?

Primary recommendation: MetaboLights (<https://www.ebi.ac.uk/metabolights>) — EMBL-EBI metabolomics repository; required by many European journals. For US-based deposition: Metabolomics Workbench (<https://www.metabolomicsworkbench.org>) — NIH-funded metabolomics repository. For mass spectrometry spectral data: GNPS (<https://gnps.ucsd.edu>) — Global Natural Products Social Molecular Networking. For NMR spectral data: MetaboLights accepts NMR data; BMRB accepts NMR metabolomics data.

## DM-30: If I want to identify a metabolite, which databases should I use?

For human metabolites: HMDB (<https://hmdb.ca>) — Human Metabolome Database; most comprehensive human metabolite resource. For spectral matching: GNPS (<https://gnps.ucsd.edu>) for MS/MS spectral matching; MassBank (<https://massbank.eu>) for reference spectra. For chemical structure: PubChem (<https://pubchem.ncbi.nlm.nih.gov>) or ChEBI (<https://www.ebi.ac.uk/chebi>). For lipids: LipidMaps (<https://www.lipidmaps.org>) — comprehensive lipid classification and database. For KEGG metabolic context: KEGG Compound (<https://www.kegg.jp/kegg/compound/>).

**WARNING: Metabolite annotation confidence levels vary. Level 1 (confirmed by reference standard) is the gold standard; Level 2 (putative annotation by spectral match) is common in untargeted metabolomics; Level 3 (putative compound class) and Level 4 (unknown) are less informative. Always report the annotation confidence level.**

## DM-31: If I want reliable human gene annotation, should I use Ensembl, RefSeq, GENCODE, or MANE?

GENCODE (<https://www.encodegenes.org>): Use for comprehensive human and mouse gene annotation; the reference annotation for ENCODE and many RNA-seq pipelines; includes all transcript isoforms. MANE (<https://www.ncbi.nlm.nih.gov/refseq/MANE/>): Use when you need a single representative transcript per gene; MANE Select is the jointly agreed Ensembl/RefSeq transcript; MANE Plus Clinical adds clinically important transcripts. Ensembl (<https://www.ensembl.org>): Use for genome browser visualization, VEP variant annotation, BioMart data retrieval, and multi-species analysis. RefSeq (<https://www.ncbi.nlm.nih.gov/refseq>): Use for NCBI-centric workflows, clinical variant annotation, and when RefSeq accessions are required.

**WARNING: Ensembl IDs, RefSeq IDs, and GENCODE annotations may differ for the same gene. Ensembl and RefSeq use different transcript models and may disagree on isoform boundaries, UTR definitions, and non-coding transcript classification. Always record the annotation version used. Do not mix annotation versions within the same analysis.**

## Critical Scientific Misuse Warnings

Before using any biological database, read the following warnings carefully. These are common sources of incorrect conclusions in database-based bioinformatics research.

1. Do not assume that a database entry is experimentally validated simply because it exists.
2. Do not confuse primary submitted records with curated reference records.
3. Do not infer biological function from BLAST similarity alone.
4. Do not use population frequency databases as clinical pathogenicity databases.
5. Do not treat computational pathogenicity predictions as clinical diagnosis.
6. Do not treat predicted protein structures as equivalent to experimentally solved structures.
7. Do not compare expression datasets without checking normalization, batch effects, tissue definitions, platform, and metadata.
8. Do not mix genome builds or annotation releases within the same analysis.
9. Do not use benchmark datasets for model training if they are intended for unbiased evaluation.
10. Do not deposit sensitive human genomic data in public repositories.
11. Do not compare AMP activity values unless assay conditions, organism, medium, MIC definition, and peptide form are comparable.
12. Do not assume microbiome taxonomy is stable across database versions.
13. Do not assume organism-specific databases are redundant with NCBI, Ensembl, or UniProt.
14. Do not cite only an umbrella portal when a specific database within the portal was used.
15. Do not build automated pipelines without recording API endpoints, query parameters, access dates, and database releases.



## Category A: General Bioinformatics Portals and Integrated Resources

### Category Overview

General bioinformatics portals and integrated resources represent the broadest tier of biological data infrastructure. These platforms do not focus on a single data type or organism but instead provide unified access to a wide range of databases, tools, and services spanning genomics, proteomics, literature, and beyond. They are the natural starting point for researchers who are new to a data domain, who need to cross-reference information across multiple biological levels, or who want to access multiple resources through a single, consistent interface. The major portals — NCBI, EMBL-EBI, DDBJ, ExPASy, and Ensembl — collectively host the majority of the world's publicly accessible biological data and are maintained by national or international funding bodies, ensuring long-term stability and open access.

A researcher needs an integrated portal when the scientific question spans multiple data types or when the appropriate specialized database is not immediately obvious. For example, a researcher who has identified a novel gene from a sequencing experiment may need to search for its sequence in a nucleotide database, find its protein product in a protein database, look up its known function in a knowledgebase, check for associated variants in a variant database, and review the relevant literature — all of which can be initiated from a single portal. Portals also provide standardized cross-links between their component databases, so that a gene record in NCBI Gene automatically links to its sequence in RefSeq, its protein in UniProt, its variants in dbSNP, and its expression data in GEO. This interconnectedness is one of the most powerful features of integrated portals.

The distinction between a portal and a database is important for accurate citation and data provenance. A portal is an access layer — it provides the interface and the links — but the actual data resides in specific databases within the portal. When reporting results, researchers should cite the specific database (e.g., "NCBI RefSeq, release 220") rather than the portal (e.g., "NCBI"). Similarly, when comparing portals, it is important to recognize that the same underlying data may be accessible through multiple portals: for example, the same nucleotide sequence may be retrieved from NCBI, EMBL-EBI, or DDBJ because all three are members of the INSDC data-sharing consortium. The choice of portal often comes down to interface preference, geographic proximity (for download speed), or the availability of specific tools and analysis pipelines.



## A1 – NCBI (National Center for Biotechnology Information)

**Official Website URL:** <https://www.ncbi.nlm.nih.gov>

**Resource Type:** Portal

**Main Biological Domain:** Omics (DNA sequences, RNA/transcriptomics, Proteins, Variants, Literature, Clinical genomics)

**What It Is Used For:** NCBI is the primary US government-funded portal for biological data, providing access to over 40 databases covering nucleotide sequences, protein sequences, genome assemblies, gene annotations, variants, gene expression data, biomedical literature, and clinical information. It is used for sequence submission, similarity searching, literature retrieval, variant interpretation, and data download across virtually all areas of molecular biology and genomics. NCBI also provides a suite of analytical tools including BLAST, Primer-BLAST, and Sequence Viewer.

**What Data It Contains:** NCBI hosts GenBank (nucleotide sequences), RefSeq (reference sequences), dbSNP (variants), dbVar (structural variants), ClinVar (clinical variants), GEO (gene expression), SRA (raw sequencing reads), PubMed (literature), PubMed Central (full-text articles), NCBI Gene (gene annotations), NCBI Protein (protein sequences), NCBI Taxonomy (organism classification), and many other specialized databases. The total data holdings span billions of nucleotide sequences, hundreds of millions of protein sequences, and tens of millions of literature citations.

**Main question it helps answer:** What is known about this gene, sequence, variant, or organism across all levels of biological data?

**Typical user:** Beginner student / Researcher / Clinician / Bioinformatician / Wet-lab scientist / Data analyst

**Example scientific questions:**

- What is the function of the human BRCA1 gene, and what variants are associated with disease?
- What nucleotide sequences are available for a specific bacterial species?
- What published studies have investigated a particular protein family?

**Example use cases:**

- Submitting a newly sequenced genome to GenBank for public archiving
- Running BLAST to identify the closest known relative of an unknown sequence
- Downloading raw RNA-seq data from SRA for reanalysis
- **Input Data Accepted:** Sequence queries (nucleotide or protein), gene names, accession numbers, organism names, keywords, PubMed IDs, variant identifiers
- **Output Data Provided:** Sequence records, alignment results, gene annotations, variant classifications, expression datasets, literature citations, genome assemblies, taxonomy records

**Strengths:**

- Largest and most comprehensive biological data portal in the world
- Free and openly accessible with no registration required for most resources
- Integrated cross-database linking enables seamless navigation between data types

- BLAST and other tools are directly integrated with the databases
- Programmatic access via E-utilities API and FTP enables large-scale data retrieval

### Limitations:

- Interface can be overwhelming for beginners due to the sheer number of databases and tools
- Data quality varies across databases; primary databases contain unreviewed submissions
- Some databases (e.g., dbSNP) are extremely large and can be slow to query interactively
- RefSeq annotation quality varies by organism; model organisms are well-annotated, others less so
- The web interface has changed significantly over the years, making older tutorials outdated

### Common beginner mistakes:

- Searching "NCBI" without specifying which database, leading to irrelevant results
- Confusing GenBank (primary submissions) with RefSeq (curated reference sequences)
- Using BLAST results without checking E-values and percent identity thresholds
- Downloading data without recording the database version or accession numbers

**When to Use It:** Use NCBI as the starting point for any bioinformatics analysis involving sequences, variants, gene annotations, or biomedical literature. It is particularly valuable when you need to cross-reference information across multiple data types or when you need to access US government-funded genomics datasets.

**When NOT to Use It:** NCBI is not the best choice for highly specialized data types such as protein-protein interactions (use STRING or IntAct), metabolic pathways (use KEGG or Reactome), or detailed protein functional annotations (use UniProtKB/Swiss-Prot). For European-centric data or resources, EMBL-EBI may provide better integration.

### Related databases / alternatives:

- EMBL-EBI (<https://www.ebi.ac.uk>): European equivalent, mirrors much of the same data with different tools
- DDBJ (<https://www.ddbj.nig.ac.jp>): Japanese equivalent, INSDC member
- Ensembl (<https://www.ensembl.org>): Specialized for genome annotation and browsing
- ExPASy (<https://www.expasy.org>): Specialized for protein analysis

**How It Connects to Other Resources:** NCBI is a member of the INSDC consortium, meaning nucleotide sequences in GenBank are automatically mirrored to ENA (EMBL-EBI) and DDBJ. NCBI Gene records link to UniProtKB protein entries, Ensembl gene models, OMIM disease entries, and GO annotations. PubMed records link to full text in PMC and to related sequences, genes, and structures.

**API / FTP / programmatic access:** E-utilities API (<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/>) provides RESTful access to all NCBI databases; supports XML, JSON, and text output. The Biopython library (Bio.Entrez module) provides a Python interface to E-utilities. FTP access at <ftp://ftp.ncbi.nlm.nih.gov> provides bulk downloads of all major databases. NCBI BLAST API available at <https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi>.

**Evidence/curation level:** Mixed — ranges from community-submitted (GenBank, SRA) to manually reviewed (RefSeq selected entries, ClinVar expert panel reviews) to computationally predicted (RefSeq automated annotations)

**Data Update Status:** Continuous updates; GenBank and SRA updated daily; RefSeq updated in regular releases (approximately monthly); PubMed updated daily

**Licensing / access restrictions:** Mostly open access; no registration required for most databases. Some clinical databases (e.g., dbGaP) require data access agreements for controlled-access data.

**Citation / Recommended Reference:** Sayers EW et al. (2022) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50(D1):D20–D26. doi:10.1093/nar/gkab1112

**Beginner-Friendly Explanation:** NCBI is like a giant library for biological data, maintained by the US government. If you have a DNA or protein sequence and want to know what it is, you can use NCBI's BLAST tool to search for similar sequences. If you want to read scientific papers about a gene or disease, you can use PubMed, which is also part of NCBI. It is the most commonly used starting point in bioinformatics.

**Advanced Technical Explanation:** NCBI's E-utilities API implements the Entrez Programming Utilities, providing programmatic access to all Entrez databases through HTTP GET/POST requests. The API supports esearch (text search), efetch (record retrieval), elink (cross-database linking), and einfo (database metadata) operations. Rate limits apply (3 requests/second without API key, 10/second with key). The NCBI BLAST API supports asynchronous job submission with RID-based result retrieval, enabling integration into automated pipelines.

**One practical workflow example:**

Step 1: Navigate to <https://www.ncbi.nlm.nih.gov> and select "Gene" from the database dropdown.

Step 2: Search for your gene of interest (e.g., "BRCA1 Homo sapiens") to retrieve the NCBI Gene record.

Step 3: From the Gene record, click "RefSeq" to access the curated reference sequence, or "ClinVar" to view associated clinical variants.

Step 4: Use the "Cited in PubMed" link to find relevant literature or click "GEO Profiles" to find expression data.

Step 5: For programmatic access, use the E-utilities API: fetch the gene record with efetch, then follow elink connections to retrieve associated sequences and variants.

## A2 – EMBL-EBI (European Bioinformatics Institute)

**Official Website URL:** <https://www.ebi.ac.uk>

**Resource Type:** Portal

**Main Biological Domain:** Omics (DNA sequences, RNA/transcriptomics, Proteins, Structures, Variants, Pathways, Literature)

**What It Is Used For:** EMBL-EBI is the European hub for biological data, providing access to over 100 databases and tools covering nucleotide sequences, protein sequences and structures, gene expression, pathways, variants, chemical biology, and more. It is used for data submission, retrieval, analysis, and integration across all areas of molecular biology. EMBL-EBI also develops and maintains widely used bioinformatics tools and standards, including Ensembl, UniProt (jointly with SIB and PIR), InterPro, Reactome, and the European Nucleotide Archive.

**What Data It Contains:** EMBL-EBI hosts the European Nucleotide Archive (ENA), UniProtKB (jointly), the Protein Data Bank in Europe (PDBe), ArrayExpress/BioStudies, Expression Atlas, IntAct (protein interactions), ChEMBL (chemical biology), Reactome (pathways), InterPro (protein families), Ensembl (genome annotation), and many other specialized databases. It is the European node of several global data-sharing consortia including INSDC, wwPDB, and UniProt.

**Main question it helps answer:** What biological data is available for this sequence, gene, protein, or organism from European and global resources?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist / Data analyst

**Example scientific questions:**

- What protein families does this sequence belong to, based on InterPro domain analysis?
- What gene expression datasets are available for specific tissue or disease condition?
- What chemical compounds interact with this protein target?

**Example use cases:**

- Submitting RNA-seq data to ArrayExpress/BioStudies for public archiving
- Searching UniProtKB for manually reviewed protein annotations
- Using the EBI BLAST service to search for protein sequences against UniRef databases

**Input Data Accepted:** Sequence queries, accession numbers, gene/protein names, chemical identifiers, organism names, keywords

**Output Data Provided:** Sequence records, protein annotations, structure data, expression datasets, pathway information, chemical biology data, interaction networks

**Strengths:**

- Hosts some of the world's most important biological databases (UniProt, ENA, PDBe, Ensembl)
- Strong emphasis on data standards, ontologies, and interoperability
- Excellent programmatic access through well-documented REST APIs
- European data sovereignty and GDPR compliance for sensitive data
- Active development of new tools and standards for emerging data types

**Limitations:**

- Interface complexity comparable to NCBI; can be overwhelming for beginners
- Some tools and databases are better maintained than others; less-used resources may have slower update cycles
- Geographic distance may affect download speeds for non-European users (though mirrors exist)
- Some resources (e.g., ArrayExpress) have undergone significant restructuring, causing confusion about current URLs and data locations

**Common beginner mistakes:**

- Confusing EMBL (the research organization) with EMBL-EBI (the bioinformatics institute)
- Not realizing that ENA, UniProt, and Ensembl are separate databases within the EBI ecosystem
- Using outdated URLs for databases that have been restructured (e.g., ArrayExpress → BioStudies)
- Overlooking the EBI's specialized tools (e.g., InterPro for domain analysis) in favor of more familiar NCBI tools

**When to Use It:** Use EMBL-EBI when working with European datasets, when you need access to resources not available at NCBI (e.g., ChEMBL for chemical biology, IntAct for protein interactions, Reactome for pathways), or when you prefer the EBI's interface and tools for protein analysis (UniProt, InterPro, PDBe).

**When NOT to Use It:** EMBL-EBI is not the best choice for US-specific clinical genomics resources (use NCBI ClinVar, dbGaP) or for accessing NCBI-specific databases such as PubMed or NCBI Gene directly. For some data types, NCBI's tools may be more familiar or better documented for North American users.

**Related databases / alternatives:**

- NCBI (<https://www.ncbi.nlm.nih.gov>): US equivalent, INSDC member
- DDBJ (<https://www.ddbj.nig.ac.jp>): Japanese equivalent, INSDC member
- ExPASy (<https://www.expasy.org>): Swiss Institute of Bioinformatics portal, strong protein focus

**How It Connects to Other Resources:** EMBL-EBI is a member of INSDC (with NCBI and DDBJ), wwPDB (with RCSB and PDBj), and UniProt (with SIB and PIR). ENA sequences are mirrored to GenBank and DDBJ. UniProt entries link to ENA, PDBe, Ensembl, InterPro, Reactome, and many other resources. The EBI's BioStudies database links to ENA, ArrayExpress, and other data archives.

**API / FTP / programmatic access:** EBI Search API (<https://www.ebi.ac.uk/ebisearch/swagger.ebi>) provides cross-database search. Individual databases have their own REST APIs (e.g., UniProt API at <https://rest.uniprot.org>, ENA API at <https://www.ebi.ac.uk/ena/portal/api>). FTP access at <ftp://ftp.ebi.ac.uk>. The Biopython library supports access to several EBI services.

**Evidence/curation level:** Mixed — ranges from community-submitted (ENA, ArrayExpress) to manually reviewed (UniProtKB/Swiss-Prot, Reactome) to computationally predicted (UniProtKB/TrEMBL, InterPro automated annotations)

**Data Update Status:** Continuous updates; ENA updated daily; UniProt updated in regular releases (approximately every 8 weeks); Ensembl updated in numbered releases (approximately quarterly)

**Licensing / access restrictions:** Mostly open access; data available under EMBL-EBI terms of use which permit free academic and commercial use for most databases. Some controlled-access datasets require data access agreements.

**Citation / Recommended Reference:** Burley SK et al. (2022) Open-access data: A fundamental principle of publicly funded research. EMBL-EBI resources are described in annual Nucleic Acids Research database issues; see <https://www.ebi.ac.uk/about/publications> for current citations.

**Beginner-Friendly Explanation:** EMBL-EBI is Europe's main center for biological data, similar to NCBI in the United States. It hosts many important databases including UniProt (for protein information), the European Nucleotide Archive (for DNA sequences), and Ensembl (for genome browsing). If you are looking for protein function information or want to submit data to a European archive, EMBL-EBI is the place to go.

**Advanced Technical Explanation:** EMBL-EBI implements a service-oriented architecture with individual REST APIs for each major database, enabling programmatic integration across resources. The EBI Search API provides federated search across all EBI databases using a unified query syntax. UniProt's REST API supports complex queries using the UniProt query language, with output in multiple formats (JSON, TSV, FASTA, GFF). The ENA portal API supports programmatic access to sequence metadata and file download URLs.

**One practical workflow example:**

Step 1: Navigate to <https://www.ebi.ac.uk> and use the search bar to find your protein of interest.

Step 2: Follow the link to the UniProtKB entry for detailed functional annotation, including domain structure, active sites, and disease associations.

Step 3: From the UniProt entry, click "Structure" to view available PDB structures in PDBe, or "Pathways" to see Reactome pathway involvement.

Step 4: Use the InterPro link to view domain and family classifications, which provide functional context for uncharacterized proteins.

Step 5: For bulk data retrieval, use the UniProt REST API to download all proteins in a given family or organism in FASTA format.



## A3 – DDBJ (DNA Data Bank of Japan)

**Official Website URL:** <https://www.ddbj.nig.ac.jp>

**Resource Type:** Portal / Database

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** DDBJ is the Japanese member of the International Nucleotide Sequence Database Collaboration (INSDC), responsible for collecting, maintaining, and distributing nucleotide sequence data from Japanese researchers and the broader international community. It provides sequence submission services, similarity search tools, and access to the shared INSDC nucleotide sequence database. DDBJ also hosts the DDBJ Sequence Read Archive (DRA) for next-generation sequencing data and the Japanese Genotype-phenotype Archive (JGA) for controlled-access human genomic data.

**What Data It Contains:** DDBJ contains the same nucleotide sequence data as GenBank and ENA through the INSDC data-sharing agreement, plus Japanese-specific resources including the DDBJ Sequence Read Archive (DRA), the Japanese Genotype-phenotype Archive (JGA), and the DDBJ BioProject and BioSample databases. The DDBJ Center also maintains specialized databases for mass spectrometry data (jPOST) and metabolomics data.

**Main question it helps answer:** What nucleotide sequences are available for this organism or gene, particularly from Japanese research groups?

**Typical user:** Researcher / Bioinformatician (particularly in Japan and Asia-Pacific region)

**Example scientific questions:**

- What sequences have been submitted by Japanese research groups for this organism?
- How do I submit my sequencing data to comply with journal requirements for data deposition?
- What raw sequencing datasets are available in the DDBJ Sequence Read Archive?

**Example use cases:**

- Submitting nucleotide sequences from a Japanese research project to the international sequence database
- Downloading raw sequencing data from DRA for reanalysis
- Accessing controlled-access human genomic data through JGA

**Input Data Accepted:** Nucleotide sequences, accession numbers, organism names, keywords; submission of FASTA/flat file format sequences

**Output Data Provided:** Sequence records in DDBJ flat file format, FASTA sequences, raw sequencing data (FASTQ), metadata records

**Strengths:**

- Full INSDC membership ensures access to the complete international nucleotide sequence database
- Specialized resources for Japanese and Asia-Pacific research communities
- JGA provides a controlled-access archive for human genomic data compliant with Japanese regulations
- DDBJ tools (ARSA search, BLAST) provide alternative access points to INSDC data
- Strong support for data submission from Japanese researchers

**Limitations:**

- Interface and documentation primarily designed for Japanese researchers; English documentation less comprehensive than NCBI or EBI
- Less widely used internationally than NCBI or EMBL-EBI, so community support is more limited
- Some specialized DDBJ tools are less well-known and less frequently updated than NCBI/EBI equivalents
- Geographic distance may affect download speeds for non-Asian users

**Common beginner mistakes:**

- Not realizing that DDBJ, GenBank, and ENA contain the same INSDC sequence data
- Overlooking DDBJ as a data source when NCBI or EBI is unavailable or slow
- Confusing DDBJ (nucleotide sequences) with DRA (raw sequencing reads) — these are separate databases within the DDBJ ecosystem

**When to Use It:** Use DDBJ when you are based in Japan or the Asia-Pacific region and need fast access to INSDC sequence data, when you need to submit data to comply with Japanese funding or journal requirements, or when you need to access the JGA for controlled-access Japanese human genomic data.

**When NOT to Use It:** DDBJ is not the best choice for accessing non-sequence biological data (use NCBI or EMBL-EBI for literature, variants, expression data), or for researchers who need extensive English-language documentation and community support.

**Related databases / alternatives:**

- GenBank/NCBI (<https://www.ncbi.nlm.nih.gov>): US INSDC member, same sequence data
- ENA/EMBL-EBI (<https://www.ebi.ac.uk/ena>): European INSDC member, same sequence data

**How It Connects to Other Resources:** DDBJ is a full member of INSDC, meaning all sequences submitted to DDBJ are automatically mirrored to GenBank and ENA. DDBJ accession numbers are recognized by all INSDC members. DRA data is mirrored to SRA (NCBI) and ENA. DDBJ BioProject and BioSample records are synchronized with NCBI BioProject and BioSample.

**API / FTP / programmatic access:** DDBJ Web API (<https://ddbj.nig.ac.jp/search>) provides search and retrieval. FTP access at <ftp://ftp.ddbj.nig.ac.jp>. The ARSA (Annotated/Assembled Sequences Retrieval Application) provides advanced search. DRA data accessible via the DRA Search interface and FTP.

**Evidence/curation level:** Community-submitted (primary sequences); INSDC data sharing means the same curation standards as GenBank and ENA apply

**Data Update Status:** Continuous synchronization with INSDC partners; daily updates

**Licensing / access restrictions:** Open access for most data; JGA requires data access agreements for controlled-access human genomic data

**Citation / Recommended Reference:** Fukuda A et al. (2021) DDBJ update in 2021. Nucleic Acids Research, 49(D1):D71–D75. doi:10.1093/nar/gkaa982

**Beginner-Friendly Explanation:** DDBJ is Japan's national DNA sequence database and is one of three international partners (along with NCBI in the US and EMBL-EBI in Europe) that share all publicly submitted DNA sequences. If you submit a sequence to DDBJ, it will automatically appear in the other two databases as well. DDBJ is particularly important for researchers in Japan and the Asia-Pacific region.



**Advanced Technical Explanation:** DDBJ implements the INSDC Feature Table format for sequence annotation, which is the same standard used by GenBank and ENA. The DDBJ Sequence Read Archive (DRA) uses the SRA XML metadata schema for experiment, run, and sample records, ensuring interoperability with NCBI SRA and ENA. The JGA implements a controlled-access model with data access committee review, similar to dbGaP at NCBI.

**One practical workflow example:**

Step 1: Navigate to <https://www.ddbj.nig.ac.jp> and select the ARSA search tool.

Step 2: Search for sequences by organism name, gene name, or accession number.

Step 3: Download sequences in FASTA or DDBJ flat file format for local analysis.

Step 4: For raw sequencing data, navigate to the DRA section and search by experiment or study accession.

Step 5: Use the FTP site for bulk downloads of large datasets.

## A4 – ExPASy (Expert Protein Analysis System)

**Official Website URL:** <https://www.expasy.org>

**Resource Type:** Portal

**Main Biological Domain:** Proteins / Systems biology

**What It Is Used For:** ExPASy is the bioinformatics resource portal of the Swiss Institute of Bioinformatics (SIB), providing access to scientific databases and software tools in proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics, and structural biology. It is the primary portal for accessing UniProtKB (the world's most comprehensive protein database), Swiss-Model (protein structure homology modeling), and many other SIB-developed resources. ExPASy is particularly valued for protein analysis, including sequence analysis, domain identification, post-translational modification prediction, and proteomics data interpretation.

**What Data It Contains:** ExPASy provides access to UniProtKB/Swiss-Prot and TrEMBL (protein sequences and annotations), PROSITE (protein domains and patterns), Swiss-Model Repository (homology models), HAMAP (microbial protein families), neXtProt (human protein knowledgebase), STRING (protein interactions, jointly), and over 150 other databases and tools developed by SIB groups. The portal serves as a gateway to the full SIB bioinformatics ecosystem.

**Main question it helps answer:** What is known about this protein's sequence, structure, function, domains, and interactions?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist (particularly in proteomics)

**Example scientific questions:**

- What domains and functional sites does this protein sequence contain?
- What is the predicted 3D structure of this protein based on homology modeling?
- What post-translational modifications are known for this protein?

**Example use cases:**

- Using Swiss-Model to generate a homology model of a protein with unknown structure
- Searching PROSITE to identify functional domains in a novel protein sequence
- Accessing neXtProt for comprehensive human protein annotations including tissue expression and disease associations

**Input Data Accepted:** Protein sequences, UniProt accession numbers, gene names, protein names, keywords

**Output Data Provided:** Protein annotations, domain predictions, homology models, interaction networks, proteomics data, functional site predictions

**Strengths:**

- Home of UniProtKB, the world's most comprehensive and well-curated protein database
- Strong focus on protein analysis with specialized tools not available elsewhere
- Swiss-Model provides high-quality homology modeling directly integrated with UniProt
- PROSITE provides expert-curated protein domain and pattern definitions
- SIB maintains high standards for data quality and curation

**Limitations:**

- Primarily focused on proteins; less comprehensive for nucleotide sequences, variants, or literature
- Some tools are less well-known than NCBI equivalents, leading to underutilization
- Interface has changed over the years; some older tutorials reference outdated tool locations
- Coverage of non-model organisms may be less comprehensive than NCBI for some data types

**Common beginner mistakes:**

- Confusing ExPASy (the portal) with UniProt (the database) — UniProt is accessible through ExPASy but is a separate resource
- Not distinguishing between Swiss-Prot (manually reviewed) and TrEMBL (computationally annotated) entries in UniProtKB
- Using PROSITE patterns without understanding their sensitivity/specificity tradeoffs
- Overlooking neXtProt as a human-specific protein resource with additional clinical annotations

**When to Use It:** Use ExPASy when your primary interest is protein analysis — function, domains, structure, post-translational modifications, or interactions. It is the best starting point for accessing UniProtKB and for protein-centric analyses.

**When NOT to Use It:** ExPASy is not the best choice for nucleotide sequence analysis (use NCBI or EMBL-EBI), literature search (use PubMed), or genome browsing (use Ensembl or UCSC). For variant interpretation, NCBI ClinVar or Ensembl VEP are more appropriate.

**Related databases / alternatives:**

- NCBI (<https://www.ncbi.nlm.nih.gov>): Broader scope including nucleotide sequences and literature
- EMBL-EBI (<https://www.ebi.ac.uk>): European portal with overlapping protein resources
- InterPro (<https://www.ebi.ac.uk/interpro>): Protein domain database at EBI, complementary to PROSITE

**How It Connects to Other Resources:** ExPASy/UniProt links to virtually every major biological database, including NCBI RefSeq, Ensembl, PDB, GO, KEGG, Reactome, InterPro, and many others. UniProt accession numbers are used as cross-references in hundreds of databases worldwide. Swiss-Model structures are deposited in the PDB and linked from UniProt entries.

**API / FTP / programmatic access:** UniProt REST API (<https://rest.uniprot.org>) provides programmatic access to all UniProt data in multiple formats (JSON, TSV, FASTA, GFF, XML). FTP access at <ftp://ftp.uniprot.org>. Swiss-Model API available at <https://swissmodel.expasy.org/docs/smtl>. PROSITE accessible via the ScanProsite web service.

**Evidence/curation level:** Mixed — UniProtKB/Swiss-Prot is manually reviewed (highest quality); UniProtKB/TrEMBL is computationally predicted; PROSITE patterns are manually curated

**Data Update Status:** UniProt updated approximately every 8 weeks with new releases; Swiss-Model Repository updated continuously; PROSITE updated with each UniProt release

Licensing / access restrictions: Open access; UniProt data available under Creative Commons Attribution 4.0 (CC BY 4.0) license

**Citation / Recommended Reference:** The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531. doi:10.1093/nar/gkac1052

**Beginner-Friendly Explanation:** ExPASy is Switzerland's main portal for protein biology, and it is home to UniProt — the most important protein database in the world. If you have a protein sequence and want to know what it does, what domains it contains, or what its 3D structure looks like, ExPASy is the place to start. The Swiss-Prot section of UniProt is especially valuable because every entry has been carefully checked by expert scientists.

**Advanced Technical Explanation:** ExPASy/UniProt implements a comprehensive cross-reference system linking each protein entry to over 200 external databases through standardized identifiers. The UniProt REST API supports complex queries using the UniProt query language (e.g., filtering by organism, reviewed status, GO term, or keyword) with output in multiple formats. UniProtKB/Swiss-Prot uses a structured evidence attribution system with ECO (Evidence and Conclusion Ontology) codes to distinguish experimental from inferred annotations at the level of individual annotation statements.

**One practical workflow example:**

Step 1: Navigate to <https://www.expasy.org> and click on "UniProt" to access the protein database.

Step 2: Search for your protein by name, gene symbol, or accession number; filter to "Reviewed (Swiss-Prot)" for manually curated entries.

Step 3: Examine the protein entry for function, subcellular localization, post-translational modifications, and disease associations.

Step 4: Click "Structure" to view available experimental structures (PDB) or generate a homology model via Swiss-Model.

Step 5: Use the "Sequences" tab to download the canonical sequence in FASTA format for further analysis.

## A5 – Ensembl

**Official Website URL:** <https://www.ensembl.org>

**Resource Type:** Genome Browser / Database / Portal

**Main Biological Domain:** DNA sequences / RNA/transcriptomics / Variants / Omics

**What It Is Used For:** Ensembl is a genome annotation and browser system developed jointly by EMBL-EBI and the Wellcome Sanger Institute, providing comprehensive genome annotation for vertebrates and selected other eukaryotes. It is used for browsing genome sequences, accessing gene and transcript annotations, retrieving variant data, analyzing regulatory features, and performing comparative genomics. Ensembl is the primary resource for Ensembl gene IDs (ENSG identifiers), which are widely used in RNA-seq analysis and other genomic workflows.

**What Data It Contains:** Ensembl contains genome assemblies and annotations for over 300 species, including gene models (protein-coding genes, non-coding RNAs, pseudogenes), transcript isoforms, protein sequences, regulatory features (promoters, enhancers, CTCF binding sites), genetic variants (from dbSNP, 1000 Genomes, gnomAD), comparative genomics data (gene trees, synteny, whole-genome alignments), and expression data. Ensembl Genomes extends coverage to plants, fungi, bacteria, and protists.

**Main question it helps answer:** What genes, transcripts, variants, and regulatory features are present at this genomic location or associated with this gene?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:**

- What transcripts are produced from the human TP53 gene, and what are their coordinates?
- What variants are present in the promoter region of a gene of interest?
- What are the orthologs of this human gene in mouse, zebrafish, and Drosophila?

**Example use cases:**

- Retrieving Ensembl gene IDs and transcript coordinates for RNA-seq analysis
- Browsing the genomic context of a variant identified in a GWAS study
- Downloading genome sequences and annotation files (GTF/GFF) for a reference genome

**Input Data Accepted:** Gene names, Ensembl IDs, genomic coordinates, variant IDs, species names, protein sequences (via BLAST)

**Output Data Provided:** Gene and transcript annotations, genomic sequences, variant data, regulatory features, comparative genomics data, protein sequences, GTF/GFF annotation files

**Strengths:**

- Comprehensive, regularly updated genome annotations for a wide range of species
- Powerful BioMart data mining tool for bulk data retrieval
- REST API (Ensembl REST API) provides programmatic access to all data
- Variant Effect Predictor (VEP) is a widely used tool for variant annotation.
- Strong comparative genomics resources including gene trees and synteny maps

**Limitations:**

- Ensembl gene IDs (ENSG) are version-specific; IDs can change between releases, causing compatibility issues
- Coverage of non-vertebrate species is less comprehensive than for human and mouse
- The web interface can be slow for large genomic regions or complex queries
- Annotation quality varies by species; human and mouse are best annotated
- Ensembl and NCBI RefSeq use different gene models, which can cause discrepancies

**Common beginner mistakes:**

- Using Ensembl gene IDs without recording the Ensembl release version, leading to irreproducibility
- Confusing Ensembl gene IDs (ENSG) with NCBI Gene IDs (integers) — these are different identifier systems
- Not realizing that Ensembl Genomes (plants, fungi, bacteria) is a separate portal from the main Ensembl site
- Downloading GTF files without checking that the genome assembly version matches the one used for alignment

**When to Use It:** Use Ensembl when you need comprehensive genome annotation for vertebrates, when you are working with RNA-seq data and need transcript coordinates, when you want to use the Variant Effect Predictor (VEP) for variant annotation, or when you need comparative genomics data.

**When NOT to Use It:** Ensembl is not the best choice for prokaryotic genome annotation (use NCBI RefSeq or specialized microbial databases), for literature search, or for protein functional annotation (use UniProt). For clinical variant interpretation, ClinVar provides more clinically relevant information than Ensembl's variant tracks.

**Related databases / alternatives:**

- UCSC Genome Browser (<https://genome.ucsc.edu>): Alternative genome browser with different annotation tracks
- NCBI Gene (<https://www.ncbi.nlm.nih.gov/gene>): Alternative gene annotation resource
- RefSeq (<https://www.ncbi.nlm.nih.gov/refseq>): NCBI's curated reference sequence database

**How It Connects to Other Resources:** Ensembl links to UniProt (protein sequences), NCBI Gene (cross-references), PDB (protein structures), GO (functional annotations), Reactome (pathways), OMIM (disease associations), and many other resources. Ensembl VEP integrates with ClinVar, dbSNP, gnomAD, and other variant databases. BioMart enables cross-database queries linking Ensembl data to external resources.

**API / FTP / programmatic access:** Ensembl REST API (<https://rest.ensembl.org>) provides programmatic access to sequences, annotations, variants, and comparative genomics data in JSON, XML, and other formats. FTP access at <ftp://ftp.ensembl.org> for bulk downloads of genome sequences, annotation files (GTF, GFF3, VCF), and other data. The biomaRt R package and pybiomart Python package provide convenient interfaces to BioMart.

**Evidence/curation level:** Mixed — gene models are computationally predicted using the Ensembl gene annotation pipeline, with manual curation for selected genes (GENCODE for human and mouse); variant data is imported from external sources with varying curation levels

**Data Update Status:** Regular numbered releases (approximately quarterly); human genome annotation updated with each release; archive releases available for reproducibility

**Licensing / access restrictions:** Open access; data available under Apache 2.0 license; genome sequences subject to the terms of the original genome project

**Citation / Recommended Reference:** Martin FJ et al. (2023) Ensembl 2023. Nucleic Acids Research, 51(D1):D933–D941. doi:10.1093/nar/gkac958

**Beginner-Friendly Explanation:** Ensembl is a genome browser and database that lets you explore the genomes of humans and hundreds of other species. You can look up any gene and see where it is in the genome, what different versions (transcripts) of it exist, and what variants have been found in it. Ensembl is especially important for RNA-seq analysis because it provides the gene coordinate files (GTF format) that are used to count reads mapping to each gene.

**Advanced Technical Explanation:** Ensembl uses an automated gene annotation pipeline that integrates evidence from protein alignments, cDNA/EST alignments, and ab initio gene predictions to generate gene models. The pipeline is run on each new genome assembly and produces Ensembl gene IDs (ENSG), transcript IDs (ENST), and protein IDs (ENSP) that are stable within a release but may change between releases. The Ensembl REST API implements a RESTful interface with content negotiation, supporting JSON, XML, FASTA, GFF3, and BED output formats. The Variant Effect Predictor (VEP) uses the Ensembl transcript database to annotate variants with predicted consequences (e.g., missense, synonymous, splice site) and integrates with multiple external databases for additional annotation.

**One practical workflow example:**

Step 1: Navigate to <https://www.ensembl.org> and search for your gene of interest (e.g., "BRCA2 human").

Step 2: Examine the gene summary page for transcript isoforms, protein domains, and variant data.

Step 3: Use BioMart (accessible from the top menu) to download a table of all transcripts with their coordinates, biotypes, and cross-references to UniProt and RefSeq.

Step 4: Download the GTF annotation file from the Ensembl FTP site for the appropriate genome assembly version.

Step 5: Use the Ensembl REST API to retrieve the sequence of a specific transcript: GET <https://rest.ensembl.org/sequence/id/ENST00000380152?type=cdna>

Step 6: Run VEP on a VCF file of variants to annotate them with predicted functional consequences.



## Beginner Example for category A

---

A student has just received the results of a Sanger sequencing experiment confirming the identity of a bacterial gene they cloned. They want to find out what is known about this gene and its protein product. The student navigates to NCBI (<https://www.ncbi.nlm.nih.gov>), selects the "Nucleotide" database, and searches for the gene name and organism. They find the GenBank record for the gene, which includes the sequence, annotation, and links to related records. From the GenBank record, they follow the link to the protein record in NCBI Protein, which links further to the UniProt entry for the protein.

From the UniProt entry (accessible via ExPASy at <https://www.expasy.org>), the student finds detailed information about the protein's function, subcellular localization, known post-translational modifications, and any disease associations. They also find links to the protein's 3D structure in the PDB and to pathway information in KEGG and Reactome. This single workflow — starting at NCBI, following links to UniProt via ExPASy — illustrates how integrated portals enable a beginner to navigate from a raw sequence to comprehensive biological knowledge without needing to know in advance which databases contain which information.

## Advanced Research Example for category A

---

An experienced bioinformatician is performing a comparative genomics analysis of a gene family across vertebrates. They begin at Ensembl (<https://www.ensembl.org>), using BioMart to retrieve all members of the gene family in human, mouse, zebrafish, and chicken, along with their Ensembl IDs, genomic coordinates, and UniProt cross-references. They download the protein sequences in FASTA format and use the Ensembl gene tree viewer to examine the evolutionary relationships among family members. They also retrieve the GTF annotation files for each species from the Ensembl FTP site to use in downstream RNA-seq analysis.

For protein-level analysis, the researcher switches to ExPASy/UniProt, using the REST API to retrieve all Swiss-Prot entries for the gene family with their manually reviewed functional annotations, domain structures, and known variants. They cross-reference these with EMBL-EBI's InterPro to identify conserved domains and with Reactome to map the family members to known pathways. The entire workflow is scripted in Python using the Biopython library for NCBI access, the UniProt REST API for protein data, and the Ensembl REST API for genomic coordinates, enabling fully reproducible analysis with explicit version tracking.

## Common Confusion Points for category A

---

NCBI, EMBL-EBI, and DDBJ all contain the same nucleotide sequences because they are INSDC partners — searching any one of them for a sequence accession number will return the same record. The choice between them is primarily one of interface preference and geographic proximity, not data content.



Ensembl and NCBI use different gene annotation systems and different gene identifiers. A gene may have an Ensembl ID (e.g., ENSG00000012048 for BRCA1) and an NCBI Gene ID (e.g., 672 for BRCA1) that are different numbers but refer to the same gene. Cross-referencing between these systems requires explicit ID mapping.

ExPASy is a portal, not a database. The databases accessible through ExPASy (UniProt, Swiss-Model, PROSITE) are separate resources with their own citations and update cycles. When citing protein data from ExPASy, cite the specific database (e.g., UniProtKB/Swiss-Prot) rather than ExPASy itself.

Ensembl gene IDs are release-specific. An ENSG ID that was valid in Ensembl release 95 may have been retired, merged, or split in a later release. Always record the Ensembl release version when using Ensembl IDs in published analyses.

The term "EMBL" is ambiguous: it can refer to the European Molecular Biology Laboratory (the research organization), the EMBL nucleotide sequence database (now called ENA), or EMBL-EBI (the bioinformatics institute). In modern usage, "ENA" refers to the nucleotide sequence database, and "EMBL-EBI" refers to the bioinformatics institute and its portal.

## Which One Should I Use?

---

For most bioinformatics tasks, NCBI is the best starting point due to its comprehensive coverage, familiar interface, and integration of sequence, literature, variant, and expression data. Use EMBL-EBI when you need resources that are primarily European (ENA for sequence submission, UniProt for protein annotation, Reactome for pathways, ChEMBL for chemical biology) or when you prefer the EBI's tools and interface. Use ExPASy specifically for protein analysis, particularly when you need UniProt's manually reviewed annotations or Swiss-Model homology modeling. Use Ensembl when you need genome annotation for vertebrates, transcript coordinates for RNA-seq analysis, or the Variant Effect Predictor for variant annotation. Use DDBJ when you are based in Japan or the Asia-Pacific region, or when you need to access Japanese-specific controlled-access genomic data through JGA.

## Category B: Literature and Scientific Publications

### Category Overview

Scientific literature databases are the primary tools for discovering, retrieving, and analyzing published research. In bioinformatics, literature search is not merely a preliminary step before "real" analysis — it is an integral part of the scientific process, used to identify known functions of genes and proteins, to find experimental evidence supporting computational predictions, to locate datasets for reanalysis, and to stay current with methodological advances. The quality and completeness of a literature search directly affect the quality of any biological interpretation, and choosing the right literature database for a given task can mean the difference between a comprehensive review and a significant gap in knowledge.

Literature databases differ substantially in their scope, indexing criteria, full-text availability, and analytical capabilities. PubMed, maintained by NCBI, is the gold standard for biomedical literature and indexes more than 40 million citations and abstracts from MEDLINE-indexed journals. PubMed Central (PMC) and Europe PMC extend this by providing full-text access to open-access articles. Google Scholar offers the broadest coverage of any literature search engine, including preprints, theses, and grey literature, but with less rigorous quality control. Semantic Scholar applies AI-based analysis to extract key findings and build citation networks. Commercial databases such as Scopus and Web of Science provide the most comprehensive citation analysis tools but require institutional subscriptions, limiting their accessibility.

A critical distinction in literature databases is between citation indexing (recording that a paper exists and who cited it) and full-text indexing (enabling search within the text of articles). PubMed indexes abstracts and metadata but not full text; PMC and Europe PMC index full text for open-access articles. This distinction matters for systematic reviews and meta-analyses, where comprehensive retrieval of all relevant studies is essential. Researchers conducting systematic reviews should search multiple databases, as no single database indexes all relevant literature. The choice of databases for a systematic review should be documented in the methods section, along with the search strategy and date of search.



## B1 – PubMed

**Official Website URL:** <https://pubmed.ncbi.nlm.nih.gov>

**Resource Type:** Literature Search Engine

**Main Biological Domain:** Literature

**What It Is Used For:** PubMed is the primary biomedical literature database maintained by the National Library of Medicine (NLM) at NCBI, providing free access to more than 40 million citations and abstracts and abstracts from MEDLINE, life science journals, and online books. It is used for literature searches, finding papers about specific genes, diseases, or methods, and identifying relevant studies for systematic reviews. PubMed is the standard starting point for any biomedical literature search.

**What Data It Contains:** PubMed contains citations and abstracts for articles indexed in MEDLINE (which covers approximately 5,200 journals in biomedicine, life sciences, behavioral sciences, chemical sciences, and bioengineering), plus additional life science journals and online books. It includes author information, MeSH (Medical Subject Headings) terms, publication types, funding information, and links to full text where available. PubMed does not store full-text articles itself but links to PMC and publisher websites.

**Main question it helps answer:** What published research exists on this gene, disease, protein, method, or biological topic?

**Typical user:** Beginner student / Researcher / Clinician / Bioinformatician / Wet-lab scientist

**Example Scientific Questions:** What studies have investigated the role of BRCA1 in DNA repair? What clinical trials have been conducted for a specific cancer treatment? What bioinformatics methods have been published for single-cell RNA-seq analysis?

**Example Use Cases:** Conducting a literature review on a gene or disease of interest Finding the original paper describing a bioinformatics tool or database Identifying datasets deposited in GEO or SRA that are associated with published studies

**Input Data Accepted:** Keywords, gene names, author names, journal names, MeSH terms, PubMed IDs (PMIDs), DOIs

**Output Data Provided:** Citation records (title, authors, abstract, journal, date, PMID), links to full text, MeSH terms, related articles, citation counts (via PubMed's "Cited by" feature for PMC articles)

**Strengths:** Free, comprehensive coverage of biomedical literature (35+ million citations) MeSH controlled vocabulary enables precise, consistent searching across synonyms Advanced search builder with Boolean operators, field tags, and filters E-utilities API enables programmatic access for large-scale literature mining Links to full text in PMC for open-access articles

**Limitations:** Does not index all biomedical journals; coverage limited to MEDLINE-indexed journals plus selected others Does not provide full-text search (only title, abstract, and MeSH terms are searchable) Preprints (bioRxiv, medRxiv) are not indexed in PubMed (though some appear in PMC) Citation analysis features are limited compared to Scopus or Web of Science MeSH indexing can lag behind publication by weeks to months

**Common Beginner Mistakes:** Assuming PubMed indexes all biomedical literature — some legitimate journals are not indexed Not using MeSH terms, leading to missed relevant articles that use different terminology Not using field tags (e.g., [tiab] for title/abstract, [mh] for MeSH) in complex searches Confusing PubMed (citations and abstracts) with PubMed Central (full-text articles)

**When to Use It:** Use PubMed as the primary resource for any biomedical literature search. It is the standard database for systematic reviews in medicine and life sciences, and it is the most reliable source for finding peer-reviewed biomedical publications.

**When NOT to Use It:** PubMed is not appropriate for finding preprints, theses, conference proceedings, or literature outside the biomedical domain. For comprehensive citation analysis, use Scopus or Web of Science. For broader scientific coverage, supplement with Google Scholar or Semantic Scholar.

**Related databases / alternatives:** PubMed Central (<https://www.ncbi.nlm.nih.gov/pmc>): Full-text repository for open-access articles Europe PMC (<https://europepmc.org>): European mirror with additional features Google Scholar (<https://scholar.google.com>): Broader coverage including preprints and grey literature Semantic Scholar (<https://www.semanticscholar.org>): AI-assisted literature discovery

**How It Connects to Other Resources:** PubMed records link to full text in PMC, to related sequences in GenBank, to gene records in NCBI Gene, and to clinical trial records in ClinicalTrials.gov. The LinkOut feature provides links to publisher websites and institutional library systems. PubMed IDs (PMIDs) are used as cross-references in many biological databases.

**API / FTP / programmatic access:** E-utilities API (<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/>) provides programmatic access; esearch for searching, efetch for retrieving records in XML or other formats. The PubMed API also supports citation retrieval. The Biopython Bio.Entrez module provides a Python interface. Bulk data available via FTP at <ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline/>

**Evidence/curation level:** Literature-curated (citations are from peer-reviewed journals indexed by NLM); MeSH indexing is performed by NLM indexers

**Data Update Status:** Daily updates; new citations added as articles are published and indexed

Licensing / access restrictions: Free and open access; no registration required. Full-text access depends on publisher agreements and institutional subscriptions.

**Citation / Recommended Reference:** Canese K, Weis S. PubMed: The Bibliographic Database. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK153387/>

**Beginner-Friendly Explanation:** PubMed is the world's largest free database of biomedical research papers. You can search for any topic in biology or medicine and find abstracts (summaries) of relevant papers. While PubMed doesn't always give you the full paper for free, it tells you where to find it. It is maintained by the US National Library of Medicine and is updated every day.

**Advanced Technical Explanation:** PubMed uses the MEDLINE database as its core, with NLM's Medical Subject Headings (MeSH) ontology for controlled vocabulary indexing. The E-utilities API implements the Entrez query syntax, supporting Boolean operators, field qualifiers, and date range filters. The esearch utility returns PMIDs



matching a query, which can then be fetched with efetch in XML, MEDLINE, or other formats. PubMed's relevance ranking algorithm considers term frequency, MeSH term matches, and recency.

**One practical workflow example:**

Step 1: Navigate to <https://pubmed.ncbi.nlm.nih.gov> and use the Advanced Search Builder.

Step 2: Add search terms with appropriate field tags: e.g., "BRCA1"[Gene/Protein Name] AND "breast cancer"[MeSH Terms] AND "2020:2024"[Date - Publication].

Step 3: Review the results, using filters (Article type, Species, Language) to narrow down.

Step 4: For papers of interest, click "Cite" to get formatted citations, or "Send to" to export to a reference manager.

Step 5: For programmatic access, use the E-utilities API: esearch to get PMIDs, then efetch to retrieve full records in XML format for parsing.

## B2 – PubMed Central (PMC)

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/pmc>

**Resource Type:** Literature Database / Repository

**Main Biological Domain:** Literature

**What It Is Used For:** PubMed Central is a free full-text archive of biomedical and life sciences journal literature maintained by NCBI. It provides open access to the complete text of articles, not just abstracts, enabling full-text search and text mining. PMC is the repository for articles published under open-access mandates from NIH, Wellcome Trust, and other funding agencies, and it is the primary resource for accessing the full text of open-access biomedical publications.

**What Data It Contains:** PMC contains full-text articles from journals that participate in the PMC archive, including articles published under open-access licenses and articles deposited under funder mandates. As of 2024, PMC contains over 9 million full-text articles. Articles are stored in XML format (JATS standard) and are available for download and text mining. PMC also hosts preprints from bioRxiv and medRxiv through the NIH Preprint Pilot.

**Main question it helps answer:** What is the full text of this open-access biomedical article, and what other articles cite or are cited by it?

**Typical user:** Researcher / Bioinformatician / Clinician / Data analyst (text mining)

**Example Scientific Questions:** What is the complete methods section of this paper describing a bioinformatics tool? What papers have cited this landmark study? What is the full text of all papers about a specific gene published in the last five years?

**Example Use Cases:** Accessing the full text of open-access articles without institutional subscription Text mining PMC articles to extract gene-disease associations or method descriptions Downloading bulk article data for natural language processing research

**Input Data Accepted:** Keywords, PMC IDs (PMCID), PubMed IDs (PMIDs), DOIs, author names, journal names

**Output Data Provided:** Full-text articles in HTML, PDF, and XML (JATS) formats; citation data; supplementary materials

**Strengths:** Free full-text access to millions of open-access biomedical articles Full-text search enables finding specific methods, reagents, or data within articles JATS XML format enables programmatic text mining Bulk download available for text mining research Includes preprints from bioRxiv and medRxiv (NIH Preprint Pilot)

**Limitations:** Does not contain all biomedical literature — only open-access articles and those deposited under funder mandates Coverage is biased toward NIH-funded research and open-access journals Full-text search is less precise than abstract-only search in PubMed due to noise from methods and references sections Some articles are available only as scanned PDFs without searchable text

**Common Beginner Mistakes:** Assuming PMC contains all articles indexed in PubMed — many PubMed articles are not in PMC Confusing PMCID (PMC identifier) with PMID (PubMed identifier) — these are different numbers Not using PMC for text mining when full-text data is needed

**When to Use It:** Use PMC when you need the full text of open-access articles, when you are conducting text mining research, or when you need to access supplementary materials from published papers.

**When NOT to Use It:** PMC is not appropriate for finding articles that are not open-access, for citation analysis (use Scopus or Web of Science), or for comprehensive literature searches (use PubMed as the primary search interface).

**Related databases / alternatives:** PubMed (<https://pubmed.ncbi.nlm.nih.gov>): Citation and abstract database; links to PMC for full text Europe PMC (<https://europepmc.org>): European equivalent with additional features bioRxiv (<https://www.biorxiv.org>): Preprint server for biology

**How It Connects to Other Resources:** PMC articles are linked to PubMed citations via PMID-PMCID mapping. PMC IDs are used as cross-references in NCBI databases. The PMC Open Access Subset is available for bulk download and text mining.

**API / FTP / programmatic access:** E-utilities API supports PMC access via efetch with db=pmc. Bulk download of the PMC Open Access Subset available at <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/> The BioC API (<https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC/>) provides structured text mining access.

**Evidence/curation level:** Literature-curated (peer-reviewed articles); preprints are not peer-reviewed

**Data Update Status:** Continuous updates as new articles are deposited

**Licensing / access restrictions:** Free access; individual article licenses vary (CC BY, CC BY-NC, etc.); the PMC Open Access Subset is available for text mining under open licenses

**Citation / Recommended Reference:** Roberts RJ (2001) PubMed Central: The GenBank of the published literature. Proceedings of the National Academy of Sciences, 98(2):381–382. doi:10.1073/pnas.98.2.381

**Beginner-Friendly Explanation:** PubMed Central is like PubMed but with the full text of articles, not just summaries. It is free to use and contains millions of complete scientific papers. If a paper is open-access (meaning the authors paid to make it freely available), you can usually find the full text in PMC. It is especially useful when you need to read the methods section of a paper to understand exactly how an experiment was done.

**Advanced Technical Explanation:** PMC stores articles in JATS (Journal Article Tag Suite) XML format, which provides structured markup for article sections, figures, tables, and references. The PMC Open Access Subset is available for bulk download in JATS XML, enabling large-scale text mining and NLP research. The BioC API provides a RESTful interface for retrieving PMC articles in BioC format, which is optimized for text mining applications. PMC implements the NCBI LinkOut system for linking to publisher websites and institutional access systems.

### One practical workflow example:

Step 1: Search PubMed for your topic and identify relevant papers with PMC full-text links.

Step 2: Click the "Free full text" link to access the article in PMC.

Step 3: For text mining, download the PMC Open Access Subset from the FTP site.

Step 4: Parse the JATS XML files to extract specific sections (methods, results) or entities (gene names, chemical names).

Step 5: Use the BioC API for structured access to individual articles by PMCID.



## B3 – Europe PMC

**Official Website URL:** <https://europepmc.org>

**Resource Type:** Literature Search Engine / Repository

**Main Biological Domain:** Literature

**What It Is Used For:** Europe PMC is a free, open-access literature database and full-text repository maintained by EMBL-EBI in partnership with a consortium of European funders. It provides access to biomedical literature including journal articles, preprints, patents, and grant information, with full-text search capabilities and text mining tools. Europe PMC is particularly valuable for European researchers and for accessing literature associated with European funding mandates (e.g., Wellcome Trust, MRC, BBSRC).

**What Data It Contains:** Europe PMC contains over 42 million abstracts and over 9 million full-text articles, including content from PubMed/MEDLINE, PubMed Central, preprint servers (bioRxiv, medRxiv, ChemRxiv, and others), patents, and agricultural literature. It also includes grant information from European funders and links between publications and associated datasets, protocols, and software.

**Main question it helps answer:** What published and preprint literature exists on this biomedical topic, including European-funded research?

**Typical user:** Researcher / Bioinformatician / Clinician

**Example Scientific Questions:** What preprints and published papers are available on a specific topic? What publications are associated with a specific European research grant? What datasets have been deposited in connection with a published study?

**Example Use Cases:** Searching for preprints alongside published literature in a single interface Finding publications linked to specific datasets in ENA or ArrayExpress Text mining Europe PMC for biological entity extraction

**Input Data Accepted:** Keywords, author names, journal names, grant IDs, PMIDs, PMCID, DOIs, preprint IDs

**Output Data Provided:** Citation records, full-text articles, preprint records, grant information, links to associated datasets and software

**Strengths:** Broader coverage than PubMed alone, including preprints and patents Links between publications and associated datasets (ENA, ArrayExpress, PDB, etc.) Full-text search across open-access articles REST API with rich query capabilities Annotations API for accessing text-mined biological entities

**Limitations:** Less widely used than PubMed in North America; some researchers may be unfamiliar with it Preprint coverage, while broader than PubMed, is still incomplete Quality control for preprints is lower than for peer-reviewed articles

**Common Beginner Mistakes:** Not realizing that Europe PMC includes preprints, which may not have been peer-reviewed Overlooking the dataset links feature, which can help find associated data for published studies

**When to Use It:** Use Europe PMC when you want to search preprints alongside published literature, when you need to find publications associated with European funding, or when you want to access the Annotations API for text-mined biological entities.



**When NOT to Use It:** Europe PMC is not the best choice for citation analysis (use Scopus or Web of Science) or for finding non-biomedical literature.

**Related databases / alternatives:**

- PubMed (<https://pubmed.ncbi.nlm.nih.gov>): US equivalent, standard for biomedical literature
- PubMed Central (<https://www.ncbi.nlm.nih.gov/pmc>): Full-text repository

**How It Connects to Other Resources:** Europe PMC links publications to associated datasets in ENA, ArrayExpress, PDB, and other EMBL-EBI databases. The Annotations API provides text-mined links between publications and biological entities (genes, proteins, diseases, chemicals).

**API / FTP / programmatic access:** Europe PMC REST API (<https://europepmc.org/RestfulWebService>) supports search, article retrieval, and citation queries. The Annotations API (<https://europepmc.org/AnnotationsApi>) provides access to text-mined biological entity annotations. Bulk data available via FTP.

**Evidence/curation level:** Literature-curated (peer-reviewed articles); preprints are not peer-reviewed; text-mined annotations are computationally predicted

**Data Update Status:** Daily updates

**Licensing / access restrictions:** Free and open access; individual article licenses vary

**Citation / Recommended Reference:** Europe PMC Consortium (2015) Europe PMC: a full-text literature database for life sciences and platform for innovation. *Nucleic Acids Research*, 43(D1):D1042–D1048. doi:10.1093/nar/gku1061

**Beginner-Friendly Explanation:** Europe PMC is similar to PubMed but is maintained by European research institutions and includes preprints (early versions of papers before peer review) alongside published articles. It is particularly useful if you want to find the very latest research, including papers that haven't been formally published yet. It also links papers to the datasets they describe, making it easier to find data associated with a study.

**Advanced Technical Explanation:** Europe PMC implements a RESTful API with support for complex queries using field-specific syntax, Boolean operators, and date range filters. The Annotations API provides access to text-mined annotations in JSON format, with entity types including genes, proteins, diseases, chemicals, and organisms, linked to specific text spans in articles. Europe PMC uses the same JATS XML format as PMC for full-text storage, enabling interoperability with PMC-based text mining pipelines.

**One practical workflow example:**

Step 1: Navigate to <https://europepmc.org> and search for your topic.

Step 2: Use the "Source" filter to include or exclude preprints, patents, or other source types.

Step 3: For a paper of interest, click "Data Links" to find associated datasets in EBI databases.

Step 4: Use the REST API to programmatically retrieve all papers matching a query: GET <https://www.ebi.ac.uk/europepmc/webservices/rest/search?query=BRCA1&format=json>

Step 5: Use the Annotations API to retrieve text-mined gene mentions from a specific article by PMCID.

## B4 – Google Scholar

**Official Website URL:** <https://scholar.google.com>

**Resource Type:** Literature Search Engine

**Main Biological Domain:** Literature (all scientific domains)

**What It Is Used For:** Google Scholar is a freely accessible web search engine that indexes the full text of scholarly literature across all disciplines, including journal articles, theses, books, conference papers, preprints, and technical reports. It is used for broad literature discovery, citation tracking, and finding grey literature not indexed in specialized databases.

**What Data It Contains:** Google Scholar indexes scholarly content from publishers, universities, preprint servers, and other sources across all scientific disciplines. It does not have a defined list of indexed sources; instead, it crawls the web for scholarly content. Coverage is broad but inconsistent, and the indexing criteria are not publicly documented. Google Scholar provides citation counts and "Cited by" links for most indexed papers.

**Main question it helps answer:** What scholarly literature exists on this topic across all scientific disciplines?

**Typical user:** Beginner student / Researcher / Clinician / Bioinformatician

**Example Scientific Questions:** What papers have cited this landmark bioinformatics paper? What theses or dissertations have been written on this topic? What conference papers or technical reports exist on this method?

**Example Use Cases:** Finding papers not indexed in PubMed (e.g., from non-MEDLINE journals, conference proceedings) Tracking citations to assess the impact of a paper Finding preprints and grey literature on a topic

**Input Data Accepted:** Keywords, author names, journal names, publication years, exact phrases

**Output Data Provided:** Citation records with links to full text (where available), citation counts, "Cited by" lists, related articles

**Strengths:** Broadest coverage of any literature search engine, including preprints, theses, and grey literature Free and accessible without institutional subscription Citation tracking ("Cited by") is useful for finding papers that build on a key study Covers all scientific disciplines, not just biomedicine

**Limitations:** Indexing criteria are not transparent; coverage is inconsistent and cannot be verified No controlled vocabulary (MeSH) for precise searching; relies on keyword matching Includes non-peer-reviewed content without clear labeling No API for programmatic access (as of 2024); scraping violates terms of service Citation counts may be inflated by self-citations and non-peer-reviewed sources Cannot be used for systematic reviews due to lack of reproducible search documentation

**Common Beginner Mistakes:** Using Google Scholar as the sole literature database for a systematic review — it is not appropriate for this purpose Trusting citation counts from Google Scholar as a measure of scientific impact without considering source quality Not realizing that some "papers" indexed by Google Scholar are preprints, theses, or non-peer-reviewed documents

**When to Use It:** Use Google Scholar for broad discovery searches, for finding papers not indexed in PubMed, for tracking citations, and for finding grey literature. It is a useful supplement to PubMed but should not replace it with systematic searches.

**When NOT to Use It:** Do not use Google Scholar as the primary database for systematic reviews or meta-analyses, as its indexing is not transparent and searches are not reproducible.

**Related databases / alternatives:** PubMed (<https://pubmed.ncbi.nlm.nih.gov>): More rigorous, reproducible biomedical literature search ; Semantic Scholar (<https://www.semanticscholar.org>): AI-assisted, with API access Scopus; (<https://www.scopus.com>): Comprehensive citation database (commercial)

**How It Connects to Other Resources:** Google Scholar links to publisher websites, institutional repositories, and preprint servers for full-text access. It does not have formal integration with biological databases.

**API / FTP / programmatic access:** No official API available. Third-party libraries (e.g., scholarly Python package) exist but may violate Google's terms of service and are unreliable.

**Evidence/curation level:** Not curated; automated web crawling with no quality filtering

**Data Update Status:** Continuous crawling; update frequency varies by source

**Licensing / access restrictions:** Free to use; no API; scraping prohibited by terms of service

**Citation / Recommended Reference:** Falagas ME et al. (2008) Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB Journal*, 22(2):338–342. doi:10.1096/fj.07-9492LSF

**Beginner-Friendly Explanation:** Google Scholar is like Google but specifically for scientific papers. It searches across all types of scholarly content — journal articles, theses, conference papers, and preprints — from all scientific fields. It is very easy to use and has the broadest coverage of any literature search tool, but it is less precise than PubMed and includes some non-peer-reviewed content. It is a good starting point for exploring a topic but should be supplemented with more rigorous databases for serious research.

**Advanced Technical Explanation:** Google Scholar uses Google's web crawling infrastructure to index scholarly content, applying heuristics to identify and extract bibliographic metadata from academic web pages. The lack of a public API and the prohibition on scraping make Google Scholar unsuitable for large-scale programmatic literature mining. Citation counts in Google Scholar are generally higher than in Scopus or Web of Science because Google Scholar indexes a broader range of sources, including preprints and grey literature that may not be peer-reviewed.

#### **One practical workflow example:**

Step 1: Navigate to <https://scholar.google.com> and enter your search terms.

Step 2: Use the "Since year" filter to restrict results to recent publications.

Step 3: Click "Cited by" on a key paper to find subsequent work that builds on it.

Step 4: Use the "Related articles" link to find papers on similar topics.

Step 5: For papers not freely available, click "All versions" to find preprint or institutional repository versions.

## B5 – Semantic Scholar

**Official Website URL:** <https://www.semanticscholar.org>

**Resource Type:** Literature Search Engine

**Main Biological Domain:** Literature (all scientific domains)

**What It Is Used For:** Semantic Scholar is a free, AI-powered academic literature search engine developed by the Allen Institute for AI. It uses natural language processing and machine learning to extract key findings, identify influential citations, and build semantic relationships between papers. It is used for literature discovery, citation analysis, and identifying the most influential papers in a field. Semantic Scholar is particularly valuable for its AI-generated paper summaries (TLDR) and its open API for programmatic access.

**What Data It Contains:** Semantic Scholar indexes over 200 million academic papers across all scientific disciplines, with particularly strong coverage of computer science and biomedical research. It provides citation graphs, author profiles, paper abstracts, AI-generated summaries (TLDR), and semantic similarity scores between papers. The Semantic Scholar Open Research Corpus (S2ORC) provides a large-scale dataset of full-text papers for research purposes.

**Main question it helps answer:** What are the most influential papers on this topic, and how do they relate to each other semantically?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example Scientific Questions:** What are the most highly cited papers on transformer models in bioinformatics? What papers are semantically similar to this key paper in my field? What is the citation network around a specific research topic?

**Example Use Cases:** Discovering influential papers in a new research area using AI-assisted recommendations  
Programmatic literature mining using the Semantic Scholar API  
Building citation networks for bibliometric analysis

**Input Data Accepted:** Keywords, author names, paper titles, DOIs, Semantic Scholar paper IDs

**Output Data Provided:** Citation records, AI-generated summaries (TLDR), citation counts, influential citation flags, semantic similarity scores, citation graphs

**Strengths:** Free and open API with generous rate limits for programmatic access  
AI-generated TLDR summaries help quickly assess paper relevance  
Identifies "influential citations"  
Open Research Corpus (S2ORC) available for large-scale research  
Covers all scientific disciplines

**Limitations:** Coverage of older literature and non-English papers may be incomplete  
AI-generated summaries can be inaccurate or misleading  
Less established than PubMed or Scopus for systematic reviews  
Citation counts may differ from other databases due to different indexing scope

**Common Beginner Mistakes:** Trusting AI-generated summaries without reading the original paper  
Using Semantic Scholar as the sole database for a systematic review

**When to Use It:** Use Semantic Scholar for AI-assisted literature discovery, for programmatic literature mining via the API, and for identifying influential papers in a field. It is a valuable supplement to PubMed.

**When NOT to Use It:** Do not use Semantic Scholar as the primary database for systematic reviews or for clinical literature searches where precision is critical.

**Related databases / alternatives:**

- PubMed (<https://pubmed.ncbi.nlm.nih.gov>): More rigorous biomedical literature search Google
- Scholar (<https://scholar.google.com>): Broader coverage, no API

**How It Connects to Other Resources:** Semantic Scholar links to publisher websites and preprint servers for full-text access. The API provides structured access to citation data that can be integrated with other analysis tools.

**API / FTP / programmatic access:** Semantic Scholar Academic Graph API (<https://api.semanticscholar.org>) provides programmatic access to paper metadata, citations, and author information in JSON format. The S2ORC dataset is available for download for research purposes.

**Evidence/curation level:** Automated indexing with AI-assisted metadata extraction; not manually curated

**Licensing / access restrictions:** Free to use; API available with registration

**Citation / Recommended Reference:** Kinney RM et al. (2023) The Semantic Scholar Open Data Platform. arXiv:2301.10140. Available at: <https://arxiv.org/abs/2301.10140>

**Beginner-Friendly Explanation:** Semantic Scholar is a free scientific paper search engine that uses artificial intelligence to help you find relevant papers and understand them quickly. It provides short AI-generated summaries of papers so you can quickly decide if a paper is relevant to your research. It also has a free API that programmers can use to automatically search for and retrieve paper information.

**Advanced Technical Explanation:** Semantic Scholar uses transformer-based NLP models (including SPECTER, a citation-informed document embedding model) to generate semantic representations of papers, enabling similarity-based retrieval and recommendation. The Academic Graph API provides access to a knowledge graph of papers, authors, and citations, with endpoints for paper search, batch retrieval, citation traversal, and author lookup. The S2ORC corpus provides full-text papers in a structured JSON format with section-level markup, suitable for large-scale NLP research.

**One practical workflow example:**

Step 1: Navigate to <https://www.semanticscholar.org> and search for your topic.

Step 2: Use the "Highly Influential Citations" filter to find papers that have had the most impact.

Step 3: Read the TLDR summary to quickly assess relevance, then click through to the full paper.

Step 4: Use the "Related Papers" feature to discover semantically similar work.

Step 5: For programmatic access, use the API: GET <https://api.semanticscholar.org/graph/v1/paper/search?query=BRCA1+DNA+repair&fields=title,abstract,citation>  
[Count](#)

## B6 – Scopus [COMMERCIAL — Institutional Access Required]

**Official Website URL:** <https://www.scopus.com>

**Resource Type:** Literature Search Engine / Citation Database

**Main Biological Domain:** Literature (all scientific domains)

**What It Is Used For:** Scopus is a comprehensive abstract and citation database operated by Elsevier, covering peer-reviewed literature across science, technology, medicine, social sciences, and arts and humanities. It is used for systematic literature searches, citation analysis, author and journal impact metrics, and research evaluation. Scopus provides more comprehensive coverage of non-English and non-biomedical literature than PubMed, making it valuable for interdisciplinary research and systematic reviews.

**What Data It Contains:** Scopus indexes over 90 million records from approximately 27,000 peer-reviewed journals, 200,000 books, and 9 million conference papers. It provides citation data, author profiles, journal metrics (CiteScore, SJR, SNIP), and affiliation information. Scopus does not provide full-text articles but links to publisher websites and institutional access systems.

**Main question it helps answer:** What is the comprehensive published literature on this topic, and what are the citation relationships between papers?

**Typical user:** Researcher / Clinician / Data analyst / Research administrator

**Example Scientific Questions:** What is the complete literature on a specific bioinformatics method, including non-English publications? What is the h-index of a specific researcher? What journals have the highest impact in a specific field?

**Example Use Cases:** Conducting a systematic review requiring comprehensive literature coverage Evaluating the citation impact of a research group or institution Identifying the most influential journals in a specific field

**Input Data Accepted:** Keywords, author names, journal names, affiliation names, DOIs, Scopus IDs

**Output Data Provided:** Citation records, citation counts, author profiles, journal metrics, citation networks, export to reference managers

**Strengths:** Comprehensive coverage of peer-reviewed literature across all disciplines Robust citation analysis tools (h-index, citation counts, citation networks) Author disambiguation and affiliation tracking Journal metrics (CiteScore, SJR) for evaluating journal quality API available for institutional subscribers

**Limitations:** Commercial database requiring institutional subscription — not freely accessible Does not index preprints or grey literature Coverage of older literature (pre-1996) is limited Full-text access requires separate publisher subscriptions

**Common Beginner Mistakes:** Assuming Scopus is freely accessible — it requires institutional subscription Using Scopus citation counts without noting that they differ from Web of Science or Google Scholar counts

**When to Use It:** Use Scopus for systematic reviews requiring comprehensive coverage, for citation analysis, and for research evaluation tasks. It is the standard tool for bibliometric analysis in many institutions.

**When NOT Use It:** Do not use Scopus if you do not have institutional access. For biomedical literature searches, PubMed is more appropriate and freely accessible.



**Related databases / alternatives:** Web of Science (<https://www.webofscience.com>): Competing commercial citation database PubMed (<https://pubmed.ncbi.nlm.nih.gov>): Free biomedical literature database Semantic Scholar (<https://www.semanticscholar.org>): Free alternative with API access

**How It Connects to Other Resources:** Scopus links to publisher websites for full-text access and integrates with reference management tools (Mendeley, RefWorks). The Scopus API enables integration with institutional research information systems.

**API / FTP / programmatic access:** Scopus API (<https://dev.elsevier.com>) available to institutional subscribers; supports search, abstract retrieval, and citation queries in JSON and XML formats.

**Evidence/curation level:** Literature-curated (peer-reviewed journals); Scopus Content Selection and Advisory Board reviews journals for inclusion

**Data Update Status:** Daily updates

**Licensing / access restrictions:** COMMERCIAL — requires institutional subscription through Elsevier

**Citation / Recommended Reference:** Burnham JF (2006) Scopus database: a review. Biomedical Digital Libraries, 3:1. doi:10.1186/1742-5581-3-1

**Beginner-Friendly Explanation:** Scopus is a large database of scientific papers from all fields of research, but it requires a paid subscription through your university or institution. It is particularly useful for finding papers from all over the world, including non-English publications, and for analyzing how many times a paper has been cited. If your institution has access, it is a powerful tool for comprehensive literature searches.

**Advanced Technical Explanation:** Scopus uses a structured metadata schema with author disambiguation algorithms to maintain consistent author profiles across name variations and institutional affiliations. The Scopus API implements the Elsevier Developer Portal standards, supporting complex queries with field-specific syntax, Boolean operators, and date range filters. Journal metrics (CiteScore, SJR, SNIP) are calculated annually using citation data from the Scopus database.

**One practical workflow example:**

- Step 1: Log in to Scopus through your institutional access portal.
- Step 2: Use the Advanced Search to construct a precise query with field tags (TITLE-ABS-KEY for title/abstract/keywords, AUTH for author, AFFIL for affiliation).
- Step 3: Apply filters for document type, date range, and subject area.
- Step 4: Export results to a reference manager (RIS, BibTeX, CSV) for systematic review management.
- Step 5: Use the "Analyze search results" feature to visualize publication trends, author networks, and country distributions.

## B7 – Web of Science [COMMERCIAL — Institutional Access Required]

**Official Website URL:** <https://www.webofscience.com>

**Resource Type:** Literature Search Engine / Citation Database

**Main Biological Domain:** Literature (all scientific domains)

**What It Is Used For:** Web of Science (WoS) is a comprehensive citation database operated by Clarivate, covering peer-reviewed literature in science, social sciences, arts, and humanities. It is the oldest and most established citation database, providing citation analysis tools including the Journal Impact Factor (JIF), h-index, and citation network analysis. Web of Science is widely used for systematic reviews, research evaluation, and bibliometric analysis.

**What Data It Contains:** Web of Science Core Collection indexes approximately 21,000 peer-reviewed journals, with citation data going back to 1900 for some collections. It includes the Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI), and other specialized collections. WoS provides citation data, author profiles, journal metrics (Impact Factor, Eigenfactor), and funding information.

**Main question it helps answer:** What is the citation history and impact of this paper, author, or journal?

**Typical user:** Researcher / Research administrator / Librarian

**Example Scientific Questions:** What is the Journal Impact Factor of a specific journal? What papers have cited this landmark study over the past decade? What is the h-index of a specific researcher?

**Example Use Cases:** Systematic reviews requiring comprehensive and reproducible literature searches Research evaluation and bibliometric analysis Identifying the most cited papers in a specific field

**Input Data Accepted:** Keywords, author names, journal names, DOIs, WoS accession numbers

**Output Data Provided:** Citation records, citation counts, Journal Impact Factors, h-indices, citation networks, export to reference managers

**Strengths:** Longest citation history of any database Journal Impact Factor is the most widely recognized journal metric Rigorous journal selection criteria ensure high-quality coverage Reproducible search documentation for systematic reviews API available for institutional subscribers

**Limitations:** Commercial database requiring institutional subscription Coverage of non-English and non-Western literature is less comprehensive than Scopus Does not index preprints or grey literature Full-text access requires separate publisher subscriptions

**Common Beginner Mistakes:** Assuming Web of Science is freely accessible Confusing the Journal Impact Factor (a journal-level metric) with the citation count of an individual paper

**When to Use It:** Use Web of Science for systematic reviews, for citation analysis, and for research evaluation. It is the standard tool for calculating Journal Impact Factors and for bibliometric analysis in many institutions.

**When NOT to Use It:** Do not use Web of Science if you do not have institutional access. For biomedical literature searches, PubMed is more appropriate and freely accessible.



**Related databases / alternatives:** Scopus (<https://www.scopus.com>): Competing commercial citation database with broader coverage, PubMed (<https://pubmed.ncbi.nlm.nih.gov>): Free biomedical literature database

**How It Connects to Other Resources:** Web of Science links to publisher websites for full-text access and integrates with reference management tools (EndNote, RefWorks). The WoS API enables integration with institutional research information systems.

**API / FTP / programmatic access:** Web of Science API (<https://developer.clarivate.com>) available to institutional subscribers; supports search and citation queries in JSON format.

**Evidence/curation level:** Literature-curated (peer-reviewed journals); Clarivate's editorial team reviews journals for inclusion

**Data Update Status:** Weekly updates

**Licensing / access restrictions:** COMMERCIAL — requires institutional subscription through Clarivate

**Citation / Recommended Reference:** Clarivate Analytics. Web of Science [Internet]. Philadelphia: Clarivate Analytics; 2024. Available from: <https://www.webofscience.com>

**Beginner-Friendly Explanation:** Web of Science is one of the oldest and most respected databases of scientific papers, but it requires a paid subscription. It is particularly known as the Journal Impact Factor, which is a measure of how often papers in a journal are cited. If your institution has access, it is a valuable tool for finding papers and understanding their scientific impact.

**Advanced Technical Explanation:** Web of Science uses a controlled indexing process with editorial review of journals for inclusion in the Core Collection. The citation index tracks all references cited in indexed articles, enabling forward and backward citation traversal. The Journal Impact Factor is calculated as the ratio of citations in a given year to citable items published in the previous two years, using only WoS-indexed citations. The WoS API supports the Expanded API for full record retrieval and the Starter API for basic search functionality.

#### **One practical workflow example:**

Step 1: Log in to Web of Science through your institutional access portal.

Step 2: Use the Advanced Search with field tags (TS= for topic, AU= for author, SO= for journal) to construct a precise query.

Step 3: Apply filters for document type, publication year, and Web of Science category.

Step 4: Export results in RIS or BibTeX format for import into a reference manager.

Step 5: Use the "Citation Report" feature to view citation trends and calculate h-index for a set of papers.

## Beginner Example for category B

---

A master's student is beginning a literature review on CRISPR-Cas9 applications in cancer therapy. They start with PubMed (<https://pubmed.ncbi.nlm.nih.gov>), using the Advanced Search to combine MeSH terms: "CRISPR-Cas Systems"[MeSH] AND "Neoplasms"[MeSH] AND "Therapy"[Subheading]. This returns several thousand results, which they filter to review articles published in the last five years. For each relevant paper, they check whether the full text is available in PubMed Central by looking for the "Free full text" link.

For papers not available in PMC, the student uses Google Scholar to search for the paper title and find preprint or institutional repository versions. They also use Semantic Scholar to find the most highly cited papers in the field and to read AI-generated summaries (TLDR) to quickly assess relevance. This combination of PubMed for systematic searching, PMC for full-text access, and Semantic Scholar for discovery represents an efficient literature review workflow for a beginner.

## Advanced Research Example for category B

---

A senior researcher is conducting a systematic review and meta-analysis of bioinformatics tools for single-cell RNA-seq analysis. They need a comprehensive, reproducible literature search across multiple databases. They search PubMed using a carefully constructed query with MeSH terms and free-text synonyms, documenting the exact search string and date. They then repeat the search in Scopus (through their institutional access) to capture papers not indexed in MEDLINE, and in Europe PMC to include preprints. The three searches are combined, duplicates are removed, and the resulting set is screened for inclusion. For citation analysis, the researcher uses Web of Science to identify the most influential papers in the field and to track how citation patterns have evolved over time. They use the Semantic Scholar API to programmatically retrieve citation networks for key papers, enabling automated identification of related work. The final systematic review documents the search strategy for all databases, including the specific query strings, databases searched, and dates of search, in accordance with PRISMA reporting guidelines.

## Common Confusion Points

---

PubMed and PubMed Central (PMC) are different resources. PubMed contains citations and abstracts; PMC contains full-text articles. Not all PubMed articles are in PMC, and not all PMC articles are indexed in PubMed. Google Scholar is not appropriate for systematic reviews because its indexing is not transparent, searches are not reproducible, and it cannot be queried programmatically. It is a useful discovery tool but not a rigorous search database. Scopus and Web of Science are commercial databases that require institutional subscriptions. Researchers without institutional access cannot use them. Many of their functions can be approximated using free alternatives (PubMed, Semantic Scholar, Europe PMC).

Citation counts differ between databases. A paper may have 500 citations in Google Scholar, 300 in Scopus, and 250 in Web of Science because each database indexes a different set of sources. Citation counts should always be reported with the source database specified. Preprints are not peer-reviewed. Europe PMC and Google Scholar index preprints from bioRxiv, medRxiv, and other servers. These papers have not undergone formal peer review and should be treated with appropriate caution, especially for clinical or policy-relevant findings.

## Which One Should I Use?

---

For most biomedical literature searches, start with PubMed — it is free, comprehensive for biomedical literature, and uses a controlled vocabulary (MeSH) that enables precise searching. Use PubMed Central when you need full-text access to open-access articles. Supplement with Europe PMC if you want to include preprints or find papers linked to specific datasets. Use Google Scholar for broad discovery and citation tracking, but not as a primary systematic search database. Use Semantic Scholar for AI-assisted discovery and programmatic access via its free API. If your institution provides access, use Scopus or Web of Science for comprehensive systematic reviews, citation analysis, and research evaluation tasks that require the most complete coverage and robust citation metrics.

## Category C: Nucleotide Sequence Databases

### Category Overview

Nucleotide sequence databases form the foundational layer of molecular biology data infrastructure. They store the raw sequence data — DNA and RNA sequences — that underlies virtually all of modern genomics, transcriptomics, and molecular biology. The three primary international nucleotide sequence databases — GenBank (NCBI), the European Nucleotide Archive (ENA, EMBL-EBI), and the DDBJ Sequence Database — are members of the International Nucleotide Sequence Database Collaboration (INSDC) and share all submitted sequences in real time, meaning that a sequence deposited in any one of the three is immediately available from all three. This tripartite collaboration, established in 1987, ensures global data availability and resilience.

Within the nucleotide sequence ecosystem, it is important to distinguish between primary sequence archives (GenBank, ENA, DDBJ) and curated reference sequence databases (RefSeq). Primary archives store sequences exactly as submitted by researchers, with minimal post-submission curation. RefSeq, by contrast, is a curated database of non-redundant reference sequences maintained by NCBI, where each entry represents the best available sequence for a given gene, transcript, or genome. RefSeq sequences are reviewed and updated by NCBI staff and are the preferred source for reference sequences in most bioinformatics analyses. The NCBI Nucleotide database is a search interface that provides unified access to both GenBank and RefSeq sequences.

Nucleotide sequence databases are used at multiple stages of a bioinformatics workflow: for retrieving reference sequences for alignment, for submitting newly generated sequences to the public record, for finding sequences from specific organisms or genes, and for downloading genome assemblies for annotation or comparative analysis. The choice between databases depends on the specific task: for sequence submission, any INSDC member is appropriate; for retrieving curated reference sequences, RefSeq is preferred; for downloading raw sequencing reads, the Sequence Read Archive (SRA) or ENA is the appropriate resource. Understanding these distinctions is essential for efficient and reproducible bioinformatics workflows.



## C1 – GenBank

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/genbank>

**Resource Type:** Database / Repository

**Main Biological Domain:** DNA sequences

**What It Is Used For:** GenBank is the primary US nucleotide sequence database maintained by NCBI and a founding member of the INSDC. It is used for depositing newly sequenced DNA and RNA sequences, retrieving sequences for analysis, and accessing the complete archive of publicly submitted nucleotide sequences. GenBank is the most widely used nucleotide sequence database in the world and is the standard repository for sequence data associated with published papers.

**What Data It Contains:** GenBank contains over 250 billion nucleotide bases from more than 2 billion sequences (as of 2024), spanning all domains of life and many viruses. Sequences are stored in GenBank flat file format with annotations including gene features, coding sequences (CDS), regulatory elements, and organism metadata. GenBank includes sequences from Sanger sequencing, next-generation sequencing assemblies, and whole-genome sequences, but not raw sequencing reads (which are stored in SRA).

**Main question it helps answer:** What nucleotide sequences have been submitted for this gene, organism, or genomic region?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:** What sequences are available for the 16S rRNA gene of a specific bacterial species? | What is the complete genome sequence of this pathogen? | What sequences have been submitted for this gene across different organisms?

**Example use cases:** Retrieving a reference sequence for PCR primer design; Submitting a newly sequenced gene or genome to the public record; Downloading sequences for phylogenetic analysis

**Input Data Accepted:** Accession numbers, gene names, organism names, keywords, sequence queries

**Output Data Provided:** Nucleotide sequences in GenBank flat file format or FASTA; sequence metadata

**Strengths:** Largest and most comprehensive nucleotide sequence archive; INSDC membership ensures data is mirrored to ENA and DDBJ; Integrated with NCBI BLAST for sequence similarity searching; Comprehensive annotation of sequences including gene features and CDS; Free and open access with no registration required

**Limitations:** Contains unreviewed, community-submitted sequences; quality varies widely; Redundant sequences (same gene from multiple submissions) can complicate analysis; Annotation quality varies by submitter; some entries have minimal annotation; Not appropriate for raw sequencing reads (use SRA); Large size makes comprehensive downloads challenging

**Common beginner mistakes:** Confusing GenBank (primary submissions) with RefSeq (curated reference sequences); Using GenBank sequences without checking annotation quality; Not recording accession numbers & versions for reproducibility; Downloading sequences without filtering for the appropriate organism or sequence type



**When to Use It:** Use GenBank when you need to retrieve sequences for a specific gene or organism, when you need to submit sequences to the public record, or when you need the most comprehensive coverage of available sequences for a given taxon.

**When NOT to Use It:** Do not use GenBank when you need curated, non-redundant reference sequences — use RefSeq instead. Do not use GenBank for raw sequencing reads — use SRA. For protein sequences, use UniProtKB or NCBI Protein.

**Related databases / alternatives:** RefSeq (<https://www.ncbi.nlm.nih.gov/refseq>): Curated reference sequences derived from GenBank; ENA (<https://www.ebi.ac.uk/ena>): European INSDC member, same data; DDBJ (<https://www.ddbj.nig.ac.jp>): Japanese INSDC member, same data

**How It Connects to Other Resources:** GenBank sequences are mirrored to ENA and DDBJ through INSDC. GenBank accession numbers are used as cross-references in NCBI Gene, RefSeq, UniProt, and many other databases. GenBank sequences can be searched using NCBI BLAST.

**API / FTP / programmatic access:** E-utilities API (efetch with db=nucleotide) for programmatic retrieval. FTP access at <ftp://ftp.ncbi.nlm.nih.gov/genbank/> for downloads. Biopython Bio.Entrez module provides Python access.

**Evidence/curation level:** Community-submitted; minimal post-submission curation (format validation only)

**Data Update Status:** Continuous updates; new sequences added daily; GenBank releases every 2 months

**Licensing / access restrictions:** Open access; sequences in the public domain

**Citation / Recommended Reference:** Sayers EW et al. (2022) GenBank. Nucleic Acids Research, 50(D1):D161–D164. doi:10.1093/nar/gkab1135

**Beginner-Friendly Explanation:** GenBank is the world's largest collection of DNA sequences, maintained by the US government. When scientists sequence a new gene or genome, they submit it to GenBank so that other researchers can access it. Every sequence gets a unique accession number (like a library catalog number) that you can use to find it again. GenBank is like a giant library of DNA sequences from all living things.

**Advanced Technical Explanation:** GenBank uses a flat file format with feature table annotations following the INSDC Feature Table specification. Each sequence record includes a LOCUS line (sequence name, length, molecule type, topology, division, date), DEFINITION, ACCESSION, VERSION (with GI number, now deprecated in favor of accession.version), FEATURES (annotated elements with qualifiers), and ORIGIN (sequence data). The GenBank division codes (BCT, VRL, PHG, SYN, etc.) classify sequences by organism type. Accession numbers follow the format: 1 letter + 5 digits (older), 2 letters + 6 digits (newer), or 2 letters + 8 digits (WGS).

### One practical workflow example:

Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/nucleotide> and search for your gene and organism

Step 2: Filter results by sequence length, date, and source (GenBank vs. RefSeq) as appropriate.

Step 3: Select sequences of interest and use "Send to" > "File" > "FASTA" to download sequences.

Step 4: Record the accession numbers and version numbers for reproducibility.

Step 5: For programmatic retrieval, use the E-utilities API: `efetch?db=nucleotide&id=ACCESSION&rettype=fasta&retmode=text`



## C2 – EMBL/ENA (European Nucleotide Archive)

**Official Website URL:** <https://www.ebi.ac.uk/ena>

**Resource Type:** Database / Repository

**Main Biological Domain:** DNA sequences

**What It Is Used For:** The European Nucleotide Archive (ENA) is the European member of the INSDC, maintained by EMBL-EBI. It provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing reads, sequence assemblies, and functional annotation. ENA is widely used for data submission (particularly by European researchers), sequence retrieval, and downloading raw sequencing data. ENA is often preferred over SRA for downloading raw sequencing reads due to its more user-friendly download interface.

**What Data It Contains:** ENA contains three main data classes: raw reads (equivalent to SRA), sequence assemblies (equivalent to GenBank/DDBJ), and annotated sequences (equivalent to GenBank/DDBJ). It holds the same assembled and annotated sequences as GenBank and DDBJ through INSDC data sharing, plus raw sequencing reads from next-generation sequencing experiments. ENA also maintains the European Genome-phenome Archive (EGA) for controlled-access human genomic data.

**Main question it helps answer:** What nucleotide sequences and raw sequencing data are available for this organism, gene, or experiment?

**Typical user:** Researcher / Bioinformatician (particularly in Europe)

**Example scientific questions:** What raw RNA-seq datasets are available for a specific tissue or disease? | What genome assemblies are available for a specific organism? | How do I submit my sequencing data to comply with European journal requirements?

**Example use cases:** Downloading raw FASTQ files for reanalysis of published RNA-seq experiments; Submitting genome assemblies and annotations to the public record; Accessing controlled-access human genomic data through EGA

**Input Data Accepted:** Accession numbers (ENA, SRA, GenBank), organism names, gene names, keywords, study/experiment/run identifiers

**Output Data Provided:** Nucleotide sequences in FASTA or EMBL flat file format, raw sequencing reads in FASTQ format, assembly files, metadata records

**Strengths:** User-friendly download interface for raw sequencing data (often preferred over SRA); Comprehensive coverage through INSDC data sharing; EGA provides controlled-access archive for sensitive human genomic data; REST API with rich query capabilities; Strong support for data submission from European researchers

**Limitations:** Interface may be less familiar to North American researchers than NCBI; Some ENA-specific tools are less well-documented than NCBI equivalents; EGA controlled-access data requires data access agreements

**Common beginner mistakes:** Not realizing that ENA contains the same assembled sequences as GenBank (INSDC sharing); Overlooking ENA as a source for raw sequencing data when SRA is slow or unavailable; Confusing ENA (nucleotide sequences) with EGA (controlled-access human genomic data)



**When to Use It:** Use ENA when you are based in Europe and need to submit data, when you prefer ENA's download interface for raw sequencing data, or when you need to access European-specific controlled-access data through EGA.

**When NOT to Use It:** ENA is not the best choice for accessing NCBI-specific resources (ClinVar, dbSNP, GEO) or for literature search.

**Related databases / alternatives:** GenBank (<https://www.ncbi.nlm.nih.gov/genbank>); SRA (<https://www.ncbi.nlm.nih.gov/sra>): NCBI raw sequencing read archive; DDBJ (<https://www.ddbj.nig.ac.jp>).

**How It Connects to Other Resources:** ENA is a full INSDC member; assembled sequences are mirrored to GenBank and DDBJ. ENA links to Ensembl for genome annotation, to UniProt for protein sequences, and to ArrayExpress/BioStudies for expression data. ENA accession numbers are recognized by all INSDC members.

**API / FTP / programmatic access:** ENA Portal API (<https://www.ebi.ac.uk/ena/portal/api>) provides programmatic access to metadata and download URLs. FTP access at <ftp://ftp.ebi.ac.uk/pub/databases/ena/>. The ENA browser API supports sequence retrieval in multiple formats.

**Evidence/curation level:** Community-submitted (primary sequences); INSDC standard curation

**Data Update Status:** Continuous updates; daily synchronization with INSDC partners

**Licensing / access restrictions:** Open access for most data;

**Citation / Recommended Reference:** Amid C et al. (2020) The European Nucleotide Archive in 2019. Nucleic Acids Research, 48(D1):D70–D76. doi:10.1093/nar/gkz1063

**Beginner-Friendly Explanation:** The European Nucleotide Archive (ENA) is Europe's main database for DNA sequences, and it contains the same sequences as GenBank in the US because they share data automatically. ENA is particularly popular for downloading raw sequencing data (the original files from a sequencing machine) because its download interface is easy to use. If you are in Europe and need to submit your sequencing data, ENA is the standard place to do it.

**Advanced Technical Explanation:** ENA implements a three-tier data model: raw reads (stored in FASTQ/BAM format with SRA XML metadata), assemblies (stored in FASTA/AGP format), and annotated sequences (stored in EMBL flat file format). The ENA Portal API supports complex queries using the ENA query syntax, with output in multiple formats (JSON, TSV, FASTA, EMBL). ENA accession number prefixes indicate data type: E/SRR/DRR for runs, E/SRP/DRP for studies, E/SRS/DRS for samples, E/SRX/DRX for experiments, and GCA/GCF for genome assemblies.

### One practical workflow example:

Step 1: Navigate to <https://www.ebi.ac.uk/ena> and search for a study of interest by accession number or keyword.

Step 2: Browse the study page to find associated runs (raw sequencing data).

Step 3: Use the "Download" button or the FTP links to download FASTQ files for reanalysis.

Step 4: For programmatic access, use the ENA Portal API: GET [https://www.ebi.ac.uk/ena/portal/api/filereport?accession=STUDY\\_ACCESSION&result=read\\_run&fields=run\\_accession,fastq ftp](https://www.ebi.ac.uk/ena/portal/api/filereport?accession=STUDY_ACCESSION&result=read_run&fields=run_accession,fastq ftp)

Step 5: Use the returned FTP URLs to download FASTQ files with wget or aspera.



## C3 – DDBJ Sequence Database

---

**Official Website URL:** <https://www.ddbj.nig.ac.jp>

**Resource Type:** Database / Repository

**Main Biological Domain:** DNA sequences

**What It Is Used For:** The DDBJ Sequence Database is the Japanese member of the INSDC, maintained by the National Institute of Genetics (NIG) in Japan. It collects, maintains, and distributes nucleotide sequence data, particularly from Japanese research groups, and provides the same INSDC sequence data as GenBank and ENA through data sharing. DDBJ also provides sequence analysis tools and submission services for Japanese researchers.

**What Data It Contains:** DDBJ contains the same assembled and annotated nucleotide sequences as GenBank and ENA through INSDC data sharing. It also maintains the DDBJ Sequence Read Archive (DRA) for raw next-generation sequencing data, the DDBJ BioProject and BioSample databases, and the Japanese Genotype-phenotype Archive (JGA) for controlled-access human genomic data.

**Main question it helps answer:** What nucleotide sequences are available for this gene or organism, particularly from Japanese research groups?

**Typical user:** Researcher / Bioinformatician (particularly in Japan and Asia-Pacific)

**Example scientific questions:**

- What sequences have been submitted by Japanese research groups for this organism?
- How do I submit my sequencing data to comply with Japanese journal requirements?
- What raw sequencing datasets are available in the DDBJ Sequence Read Archive?

**Example use cases:**

- Submitting sequences from a Japanese research project
- Accessing controlled-access Japanese human genomic data through JGA
- Downloading raw sequencing data from DRA

**Input Data Accepted:** Accession numbers, organism names, gene names, keywords

**Output Data Provided:** Nucleotide sequences in DDBJ flat file or FASTA format, raw sequencing reads, metadata records

**Strengths:**

- Full INSDC membership ensures access to the complete international nucleotide sequence database
- JGA provides controlled-access archive compliant with Japanese regulations
- Strong support for Japanese researchers

**Limitations:**

- Less comprehensive English documentation than NCBI or EBI
- Primarily serves the Japanese and Asia-Pacific research community
- Some tools are less well-known internationally

**Common beginner mistakes:**

- Not realizing that DDBJ contains the same INSDC sequences as GenBank and ENA
- Overlooking DDBJ as an alternative access point when NCBI or EBI is slow

**When to Use It:** Use DDBJ when you are based in Japan or the Asia-Pacific region, when you need to submit data to comply with Japanese requirements, or when you need to access JGA controlled-access data.

**When NOT to Use It:** DDBJ is not the best choice for non-sequence biological data or for researchers who need extensive English-language documentation.

**Related databases / alternatives:**

- GenBank (<https://www.ncbi.nlm.nih.gov/genbank>): US INSDC member
- ENA (<https://www.ebi.ac.uk/ena>): European INSDC member

**How It Connects to Other Resources:** DDBJ is a full INSDC member; sequences are mirrored to GenBank and ENA. DRA data is mirrored to SRA and ENA. DDBJ BioProject and BioSample records are synchronized with NCBI.

**API / FTP / programmatic access:** DDBJ Web API and FTP at <ftp://ftp.ddbj.nig.ac.jp>. ARSA search interface for advanced queries.

**Evidence/curation level:** Community-submitted; INSDC standard curation

**Data Update Status:** Continuous synchronization with INSDC partners

**Licensing / access restrictions:** Open access for most data; JGA requires data access agreements

**Citation / Recommended Reference:** Fukuda A et al. (2021) DDBJ update in 2021. Nucleic Acids Research, 49(D1):D71–D75. doi:10.1093/nar/gkaa982

**Beginner-Friendly Explanation:** DDBJ is Japan's national DNA sequence database and shares all its data with GenBank (US) and ENA (Europe) automatically. It is the main submission point for Japanese researchers and provides access to Japanese-specific controlled-access genomic data. The sequences in DDBJ are the same as those in GenBank and ENA.

**Advanced Technical Explanation:** DDBJ implements the INSDC Feature Table format for sequence annotation, identical to GenBank and ENA. The DRA uses the SRA XML metadata schema for interoperability with NCBI SRA and ENA. JGA implements a controlled-access model with data access committee review, analogous to dbGaP at NCBI and EGA at EMBL-EBI.

**One practical workflow example:**

Step 1: Navigate to <https://www.ddbj.nig.ac.jp> and use the ARSA search tool.

Step 2: Search for sequences by organism or gene name.

Step 3: Download sequences in FASTA or DDBJ flat file format.

Step 4: For raw sequencing data, navigate to DRA and search by study accession.

Step 5: Use FTP for bulk downloads.

## C4 – RefSeq

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/refseq>

**Resource Type:** Database (Curated)

**Main Biological Domain:** DNA sequences / RNA/transcriptomics / Proteins

**What It Is Used For:** RefSeq is a curated, non-redundant database of reference sequences for genomes, transcripts, and proteins, maintained by NCBI. It provides a stable, well-annotated reference sequence for each gene, transcript, and protein, derived from GenBank submissions but with additional curation and quality control. RefSeq sequences are the standard reference for genome annotation, RNA-seq analysis, variant calling, and many other bioinformatics workflows. RefSeq accession numbers (NM\_, NR\_, NP\_, NC\_, NG\_, etc.) are widely used in literature and in clinical genomics.

**What Data It Contains:** RefSeq contains reference sequences for genomic DNA (NC\_, NG\_, NT\_, NW\_ accessions), mRNA transcripts (NM\_ accessions), non-coding RNA (NR\_ accessions), proteins (NP\_ accessions), and whole genome shotgun sequences (NZ\_ accessions). RefSeq covers over 100,000 organisms and includes manually curated sequences for model organisms and computationally generated sequences for others. RefSeq also provides genome annotation files (GFF, GTF) for use in bioinformatics pipelines.

**Main question it helps answer:** What is the authoritative reference sequence for this gene, transcript, or protein?

**Typical user:** Bioinformatician / Researcher / Clinician

**Example scientific questions:**

- What is the reference mRNA sequence for the human TP53 gene?
- What is the current genome assembly and annotation for the mouse reference genome?
- What RefSeq accession should I use as the reference for variant calling?

**Example use cases:** Downloading the human reference genome (GRCh38) annotation in GTF format for RNA-seq analysis; Using RefSeq mRNA sequences as references for primer design; Reporting variants using RefSeq transcript accessions (e.g., NM\_000546.6:c.817C>T for TP53)

**Input Data Accepted:** RefSeq accession numbers, gene names, organism names, keywords

**Output Data Provided:** Reference sequences in FASTA format, GenBank flat file format, GFF/GTF annotation files, protein sequences

**Strengths:** Curated, non-redundant reference sequences with stable accession numbers; Standard reference for clinical variant reporting (HGVS nomenclature uses RefSeq accessions); Comprehensive genome annotation files for major model organisms; Regularly updated with new assemblies and annotations; Integrated with NCBI Gene, ClinVar, and other NCBI databases

**Limitations:** Annotation quality varies by organism; model organisms are well-annotated, others less so RefSeq and Ensembl use different gene models, which can cause discrepancies in transcript counts Some RefSeq entries are computationally predicted (XM\_, XR\_, XP\_ accessions) with lower confidence Accession version numbers change when sequences are updated, requiring version tracking

**Common beginner mistakes:** Confusing RefSeq (curated reference sequences) with GenBank (primary submissions); Not distinguishing between curated (NM\_, NP\_) and predicted (XM\_, XP\_) RefSeq accessions; Using an outdated RefSeq version without recording the version number; Mixing RefSeq and Ensembl gene models in the same analysis

**When to Use It:** Use RefSeq when you need a stable, curated reference sequence for a gene, transcript, or protein. It is the standard reference for clinical variant reporting, RNA-seq analysis, and genome annotation.

**When NOT to Use It:** Do not use RefSeq when you need the most comprehensive coverage of all submitted sequences (use GenBank), when you need Ensembl-specific gene models (use Ensembl), or when you need protein functional annotations (use UniProt).

**Related databases / alternatives:** GenBank (<https://www.ncbi.nlm.nih.gov/genbank>): Primary sequence archive from which RefSeq is derived; Ensembl (<https://www.ensembl.org>): Alternative genome annotation system; UniProt (<https://www.uniprot.org>): Curated protein sequence and function database

**How It Connects to Other Resources:** RefSeq accession numbers are used as cross-references in NCBI Gene, ClinVar, dbSNP, UniProt, and many other databases. RefSeq genome assemblies are the basis for NCBI's genome annotation pipeline. RefSeq transcript accessions are used in HGVS variant nomenclature.

**API / FTP / programmatic access:** E-utilities API (efetch with db=nucleotide or db=protein) for programmatic retrieval. FTP access at <ftp://ftp.ncbi.nlm.nih.gov/refseq/> for bulk downloads of genome sequences, annotation files, and release notes.

**Evidence/curation level:** Mixed — manually curated for model organisms (NM\_, NP\_ accessions); computationally predicted for others (XM\_, XP\_ accessions)

**Data Update Status:** Regular releases (approximately monthly); continuous updates for individual records

**Licensing / access restrictions:** Open access; sequences in the public domain

**Citation / Recommended Reference:** O'Leary NA et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745. doi:10.1093/nar/gkv1189

**Beginner-Friendly Explanation:** RefSeq is NCBI's collection of carefully checked reference sequences for genes, RNA molecules, and proteins. Unlike GenBank, which stores everything that researchers submit, RefSeq contains only one high-quality reference sequence for each gene in each organism. When you are doing RNA-seq analysis or looking up a gene's sequence, RefSeq is usually the best place to get the reference sequence because it has been reviewed and is regularly updated.

**Advanced Technical Explanation:** RefSeq implements a hierarchical accession number system: NC\_ (complete genomic molecules), NG\_ (incomplete genomic regions), NM\_ (mRNA), NR\_ (non-coding RNA), NP\_ (protein), NT\_ (contig), NW\_ (WGS contig), NZ\_ (WGS scaffold/chromosome). Predicted records use X prefix (XM\_, XR\_, XP\_). RefSeq annotation is generated by the NCBI Eukaryotic Genome Annotation Pipeline (EGAP) for eukaryotes and the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) for prokaryotes. RefSeq release notes document changes between releases, enabling version tracking for reproducibility.



**One practical workflow example:**

Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/refseq> and search for your gene (e.g., "TP53 Homo sapiens mRNA").

Step 2: Select the NM\_ accession for the curated mRNA reference sequence.

Step 3: Download the sequence in FASTA format for use as a reference in alignment or primer design.

Step 4: For genome annotation, download the GFF3 or GTF file from the RefSeq FTP site for the appropriate genome assembly.

Step 5: Record the RefSeq accession number and version (e.g., NM\_000546.6) for reproducibility.

## C5 – NCBI Nucleotide

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/nucleotide>

**Resource Type:** Database (Search Interface)

**Main Biological Domain:** DNA sequences

**What It Is Used For:** NCBI Nucleotide is a unified search interface that provides access to nucleotide sequences from multiple NCBI databases, including GenBank, RefSeq, and the Third Party Annotation (TPA) database. It is the primary web interface for searching and retrieving nucleotide sequences at NCBI, combining the content of GenBank and RefSeq into a single searchable resource.

**What Data It Contains:** NCBI Nucleotide provides access to all nucleotide sequences in GenBank (community-submitted), RefSeq (curated reference sequences), and TPA (third-party annotations). It does not store sequences separately from these underlying databases but provides a unified search interface across them. The total content includes billions of sequences from all organisms.

**Main question it helps answer:** What nucleotide sequences are available for this gene, organism, or accession number across all NCBI nucleotide databases?

**Typical user:** Beginner student / Researcher / Bioinformatician

**Example scientific questions:**

- What sequences are available for the human GAPDH gene?
- What is the sequence associated with this accession number?
- What complete genome sequences are available for this bacterial species?

**Example use cases:**

- Searching for sequences to use as BLAST queries
- Retrieving a sequence by accession number
- Browsing available sequences for a specific organism and gene

**Input Data Accepted:** Accession numbers, gene names, organism names, keywords, sequence queries

**Output Data Provided:** Nucleotide sequences in GenBank flat file or FASTA format, with annotations and links to related records

**Strengths:** Unified search across GenBank and RefSeq; Familiar NCBI interface with advanced search options; Direct integration with BLAST for sequence similarity searching; Links to related records in NCBI Gene, Protein, and other databases

**Limitations:** Does not distinguish clearly between GenBank and RefSeq results without filtering; Large result sets can be difficult to navigate; Not appropriate for raw sequencing reads (use SRA)

**Common beginner mistakes:** Not filtering by database (GenBank vs. RefSeq) when a specific type of sequence is needed; Confusing NCBI Nucleotide (search interface) with GenBank (the underlying database)

**When to Use It:** Use NCBI Nucleotide as the primary search interface for nucleotide sequences at NCBI. It is the most convenient starting point for finding sequences by gene name, organism, or accession number.

**When NOT to Use It:** For raw sequencing reads, use SRA. For protein sequences, use NCBI Protein or UniProt. For genome assemblies, use NCBI Assembly.

**Related databases / alternatives:** GenBank (<https://www.ncbi.nlm.nih.gov/genbank>): Underlying primary sequence database; RefSeq (<https://www.ncbi.nlm.nih.gov/refseq>): Underlying curated reference database; ENA (<https://www.ebi.ac.uk/ena>): European equivalent

**How It Connects to Other Resources:** NCBI Nucleotide links to NCBI Gene, Protein, Structure, PubMed, and other NCBI databases. Sequences can be sent directly to BLAST for similarity searching.

**API / FTP / programmatic access:** E-utilities API (esearch and efetch with db=nucleotide) for programmatic access. Biopython Bio.Entrez module provides Python interface.

**Evidence/curation level:** Mixed — includes both community-submitted (GenBank) and curated (RefSeq) sequences

**Data Update Status:** Continuous updates reflecting changes in GenBank and RefSeq

**Licensing / access restrictions:** Open access

**Citation / Recommended Reference:** Sayers EW et al. (2022) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 50(D1):D20–D26. doi:10.1093/nar/gkab1112

**Beginner-Friendly Explanation:** NCBI Nucleotide is the main search page for finding DNA sequences at NCBI. It searches across both GenBank (all submitted sequences) and RefSeq (carefully checked reference sequences) at the same time. If you have an accession number or a gene name, this is the place to go to find the sequence. Think of it as the search engine for NCBI's DNA sequence collections.

**Advanced Technical Explanation:** NCBI Nucleotide implements the Entrez search system with support for field-specific queries using tags such as [Organism], [Gene Name], [Accession], [Sequence Length], and [Modification Date]. The search results can be filtered by source database (GenBank, RefSeq, TPA) using the "Source databases" filter. Programmatic access via E-utilities supports complex queries with Boolean operators and field tags, with output in multiple formats (GenBank flat file, FASTA, XML).

#### **One practical workflow example:**

Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/nucleotide> and enter your search query.

Step 2: Use the "Filters" panel to restrict to RefSeq sequences if you need curated references.

Step 3: Select sequences of interest and use "Send to" > "File" > "FASTA" to download.

Step 4: For a specific accession, enter it directly in the search box to retrieve the record.

Step 5: Use the "Run BLAST" button to search for similar sequences directly from the results page.



## Beginner Example for category C

---

A student has just completed a PCR experiment and Sanger-sequenced a fragment of the 16S rRNA gene from a bacterial isolate. They want to identify the organism. They navigate to NCBI Nucleotide (<https://www.ncbi.nlm.nih.gov/nucleotide>) and use the BLAST link to submit their sequence for similarity searching against the 16S rRNA database. The top BLAST hits show high identity (>99%) to *Escherichia coli* sequences in GenBank, suggesting the isolate is *E. coli*. The student records the accession numbers of the top hits and the BLAST parameters used.

To find the reference 16S rRNA sequence for *E. coli* for comparison, the student searches NCBI Nucleotide for "16S ribosomal RNA *Escherichia coli*" and filters to RefSeq sequences. They download the RefSeq reference sequence in FASTA format and record the accession number (e.g., NR\_024570.1) for their lab notebook. This workflow — BLAST for identification, RefSeq for reference sequences — is a standard approach for sequence-based organism identification.

## Advanced Research Example for category C

---

A bioinformatician is building a phylogenetic tree of a bacterial gene family across 50 species. They use the NCBI Nucleotide E-utilities API to programmatically retrieve all RefSeq sequences for the gene family across the target organisms, filtering by organism taxonomy and sequence type. They download the sequences in FASTA format and align them using MUSCLE or MAFFT. For organisms where RefSeq sequences are not available, they search GenBank for community-submitted sequences, carefully evaluating annotation quality before inclusion.

For the genome-level analysis, the researcher downloads complete genome assemblies from the NCBI Assembly database (linked from RefSeq) and uses the GFF3 annotation files to extract the gene coordinates. They also check ENA for any additional sequences from European research groups that may not yet be in GenBank, using the ENA Portal API to query by gene name and organism. The final dataset is assembled from multiple INSDC sources, with all accession numbers and database versions recorded for reproducibility.

## Common Confusion Points

---

GenBank, ENA, and DDBJ contain the same assembled and annotated sequences because they are INSDC partners. Searching any one of them for a specific accession number will return the same record. The choice between them is primarily one of interface preference and geographic proximity.

RefSeq is not the same as GenBank. GenBank contains all submitted sequences; RefSeq contains only curated reference sequences derived from GenBank. For most bioinformatics analyses, RefSeq is preferred because it provides non-redundant, well-annotated reference sequences.





RefSeq accession prefixes indicate the type of sequence: NM\_ (mRNA), NR\_ (non-coding RNA), NP\_ (protein), NC\_ (complete genomic molecule), NG\_ (incomplete genomic region). Predicted sequences use X prefixes (XM\_, XR\_, XP\_) and have lower confidence than curated sequences.

Raw sequencing reads (FASTQ files) are not stored in GenBank, ENA assembled sequences, or DDBJ assembled sequences. Raw reads are stored in SRA (NCBI), ENA raw reads section, or DRA (DDBJ). These are separate databases from the assembled sequence archives.

NCBI Nucleotide is a search interface, not a separate database. It searches across GenBank and RefSeq simultaneously. When citing sequences retrieved through NCBI Nucleotide, cite the specific database (GenBank or RefSeq) and the accession number, not "NCBI Nucleotide."

### Which One Should I Use?

---

For retrieving reference sequences for bioinformatics analysis (RNA-seq, variant calling, primer design), use RefSeq — it provides curated, non-redundant reference sequences with stable accession numbers. For comprehensive sequence searches across all submitted sequences, use GenBank (via NCBI Nucleotide as the search interface). For downloading raw sequencing reads, use ENA (preferred for its user-friendly download interface) or SRA. For sequence submission, use any INSDC member (GenBank, ENA, or DDBJ) — the choice depends on your geographic location and institutional requirements. For European researchers, ENA is the standard submission point; for Japanese researchers, DDBJ; for North American researchers, GenBank.

## Category D: Sequence Similarity and Search Tools

### Category Overview

Sequence similarity search is one of the most fundamental operations in bioinformatics. The principle underlying these tools is that sequence similarity implies evolutionary relationship (homology), and that homologous sequences tend to share biological function. By comparing an unknown sequence to a database of known sequences, researchers can infer the function, structure, and evolutionary history of the unknown sequence. This approach, formalized by the development of BLAST in 1990, has become the most widely used computational method in biology, with billions of BLAST searches performed annually.

The landscape of sequence similarity tools has diversified considerably since the original BLAST algorithm. Different tools are optimized for different use cases: BLAST remains the standard for general-purpose similarity searching; PSI-BLAST extends BLAST's sensitivity for detecting distant homologs using iterative profile-based searching; HMMER uses profile hidden Markov models (HMMs) for even more sensitive detection of remote homologs, particularly for protein family analysis; DIAMOND provides BLAST-like sensitivity at orders-of-magnitude faster speeds for large-scale metagenomic and proteomics analyses; and FASTA provides an alternative alignment algorithm with different sensitivity/speed tradeoffs. Understanding when to use each tool is essential for efficient and accurate sequence analysis.

A critical concept in sequence similarity searching is the distinction between sensitivity and specificity. Sensitivity refers to the ability to detect true homologs, including distant ones; specificity refers to the ability to avoid false positives (non-homologous sequences that appear similar by chance). BLAST is fast and specific but may miss distant homologs. HMMER and PSI-BLAST are more sensitive but slower and may produce more false positives if not used carefully. The E-value (expected value) is the standard statistical measure of significance in sequence similarity searches: it represents the number of hits with a given score that would be expected by chance in a database of the given size. An E-value of 0.001 means that approximately 1 in 1,000 random database searches would produce a hit of this quality by chance. Lower E-values indicate more significant hits.

## D1 – NCBI BLAST (Basic Local Alignment Search Tool)

---

**Official Website URL:** <https://blast.ncbi.nlm.nih.gov>

**Resource Type:** Tool (Sequence Similarity Search)

**Main Biological Domain:** DNA sequences / Proteins

**What It Is Used For:** NCBI BLAST is the most widely used sequence similarity search tool in bioinformatics, enabling researchers to compare a query sequence (nucleotide or protein) against databases of known sequences to identify similar sequences. It is used for sequence identification, functional annotation, homolog detection, primer checking, and many other applications. NCBI BLAST provides a web interface for interactive searches and an API for programmatic access.

**What Data It Contains:** NCBI BLAST is a tool, not a database — it searches against NCBI's sequence databases including the non-redundant nucleotide database (nt), the non-redundant protein database (nr), RefSeq, UniProt, PDB, and others. The databases searched are maintained by NCBI and updated regularly.

**Main question it helps answer:** What known sequences are similar to my query sequence, and what can I infer about its function or identity?

**Typical user:** Beginner student / Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:**

- What organism does this unknown DNA sequence come from?
- What is the function of this novel protein sequence?
- Are there any known sequences similar to my PCR product?

**Example use cases:**

- Identifying an unknown sequence from a sequencing experiment
- Checking PCR primers for off-target binding sites
- Finding homologs of a protein of interest across different organisms

**Input Data Accepted:** Nucleotide or protein sequences in FASTA format or plain text; accession numbers

- **Output Data Provided:** Alignment results with E-values, percent identity, alignment length, and bit scores; links to matching database records; graphical alignment overview

**Strengths:**

- Most widely used and well-documented sequence similarity tool
- Multiple BLAST programs for different query/database combinations (BLASTn, BLASTp, BLASTx, tBLASTn, tBLASTx)
- Web interface is beginner-friendly with sensible defaults
- Searches against comprehensive NCBI databases (nr, nt, RefSeq, UniProt, PDB)
- API and command-line versions available for large-scale analyses

**Limitations:**

- Web interface can be slow for large queries or during peak usage
- Standard BLAST may miss very distant homologs (use PSI-BLAST or HMMER for sensitive searches)



- E-value interpretation requires understanding of database size effects
- Not suitable for very large-scale searches (millions of sequences) — use DIAMOND instead
- Results depend heavily on the database searched; different databases give different results

**Common beginner mistakes:**

- Using BLASTn (nucleotide) when BLASTp (protein) would be more appropriate for functional annotation
- Interpreting any BLAST hit as proof of functional equivalence
- Not checking the E-value threshold; accepting hits with E-values > 0.01 without caution
- Not specifying the organism filter when looking for hits in a specific taxon
- Forgetting to record the database version and search date

**When to Use It:** Use NCBI BLAST for standard sequence similarity searches, for identifying unknown sequences, for finding homologs in NCBI databases, and for checking primer specificity. It is the appropriate first tool for most sequence analysis tasks.

**When NOT to Use It:** Do not use NCBI BLAST for very large-scale searches (use DIAMOND or local BLAST), for detecting very distant homologs (use PSI-BLAST or HMMER), or for searching non-NCBI databases (use EBI BLAST or local BLAST with custom databases).

**Related databases / alternatives:**

- EBI BLAST (<https://www.ebi.ac.uk/Tools/sss/ncbiblast>): BLAST against EBI databases
- PSI-BLAST: More sensitive iterative BLAST for distant homologs
- HMMER (<https://hmmer.org>): Profile HMM-based search for remote homologs
- DIAMOND (<https://github.com/bbuchfink/diamond>): Fast BLAST-like search for large datasets

**How It Connects to Other Resources:** NCBI BLAST results link directly to GenBank, RefSeq, UniProt, and PDB records. BLAST can be run against any NCBI database, and results include cross-references to related databases.

**API / FTP / programmatic access:** NCBI BLAST API (<https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi>) supports programmatic submission and result retrieval. The BLAST+ command-line suite (<https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html>) enables local BLAST searches. Biopython Bio.Blast module provides Python interface.

**Evidence/curation level:** Tool (not a database); results depend on the curation level of the database searched

**Data Update Status:** NCBI BLAST databases updated regularly (nr and nt updated approximately weekly)

**Licensing / access restrictions:** Free and open access; BLAST+ software available under public domain license

**Citation / Recommended Reference:** Camacho C et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics, 10:421. doi:10.1186/1471-2105-10-421

**Beginner-Friendly Explanation:** BLAST is like a search engine for DNA and protein sequences. You give it a sequence, and it searches through millions of known sequences to find the ones that are most similar. If your sequence is similar to a known gene, BLAST will find it and tell you what that gene does. It is one of the most important tools in bioinformatics and is used by scientists all over the world every day.

**Advanced Technical Explanation:** BLAST uses a heuristic algorithm based on finding short exact matches (words) between the query and database sequences, then extending these matches into local alignments using a scoring matrix (BLOSUM62 for proteins, nucleotide substitution matrices for DNA). The statistical significance of alignments is assessed using the Karlin-Altschul statistics, which model the distribution of alignment scores for random sequences. The E-value is calculated as  $E = Kmne^{(-\lambda S)}$ , where  $K$  and  $\lambda$  are statistical parameters,  $m$  and  $n$  are the query and database lengths, and  $S$  is the raw alignment score. BLAST+ implements multiple programs: blastn (nucleotide-nucleotide), blastp (protein-protein), blastx (translated nucleotide vs. protein), tblastn (protein vs. translated nucleotide), and tblastx (translated nucleotide vs. translated nucleotide).

**One practical workflow example:**

Step 1: Navigate to <https://blast.ncbi.nlm.nih.gov> and select the appropriate BLAST program (BLASTp for protein, BLASTn for nucleotide).

Step 2: Paste your query sequence in FASTA format into the search box.

Step 3: Select the appropriate database (nr for comprehensive search, RefSeq for curated references, UniProtKB/Swiss-Prot for manually reviewed proteins).

Step 4: Set the E-value threshold (default 0.05; use  $1e-5$  for more stringent searches) and submit.

Step 5: Examine the results: check E-values, percent identity, and alignment coverage; follow links to database records for functional information.

Step 6: For programmatic access, use the BLAST API: submit with PUT, retrieve results with GET using the RID (Request ID).

## D2 – EBI BLAST (NCBI-BLAST at EMBL-EBI)

**Official Website URL:** <https://www.ebi.ac.uk/Tools/sss/ncbiblast>

**Resource Type:** Tool (Sequence Similarity Search)

**Main Biological Domain:** DNA sequences / Proteins

**What It Is Used For:** EBI BLAST is the EMBL-EBI implementation of the NCBI BLAST algorithm, providing sequence similarity searching against EBI-hosted databases including UniProtKB, ENA, Ensembl, and others. It uses the same BLAST algorithm as NCBI BLAST but searches against different databases, making it particularly useful for searching against UniProt or ENA sequences. EBI BLAST is part of the EBI's suite of sequence analysis tools (EBI Tools).

**What Data It Contains:** EBI BLAST is a tool that searches against EBI-hosted databases including UniProtKB (Swiss-Prot and TrEMBL), ENA nucleotide sequences, Ensembl genomes, PDB structures, and others. The databases are maintained by EMBL-EBI and updated regularly.

**Main question it helps answer:** What sequences in EBI databases are similar to my query sequence?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What UniProt proteins are similar to my query protein sequence?
- What ENA nucleotide sequences match my query?
- How does my sequence compare to Ensembl gene sequences?

**Example use cases:**

- Searching against UniProtKB/Swiss-Prot for manually reviewed protein annotations
- Searching against ENA for nucleotide sequences from specific organisms
- Comparing results from EBI BLAST and NCBI BLAST to assess database coverage differences

**Input Data Accepted:** Nucleotide or protein sequences in FASTA format; accession numbers

**Output Data Provided:** Alignment results with E-values, percent identity, and bit scores; links to EBI database records

**Strengths:**

- Searches against EBI-specific databases (UniProtKB, ENA, Ensembl) not available at NCBI BLAST
- REST API (EBI Tools API) enables programmatic access
- Same BLAST algorithm as NCBI BLAST, enabling direct comparison of results
- Integrated with other EBI tools and databases

**Limitations:**

- Less widely used than NCBI BLAST; fewer tutorials and community resources
- EBI Tools web services have undergone restructuring; some older URLs may be outdated
- Not as comprehensive as NCBI BLAST for searching against all available sequences

**Common beginner mistakes:**

- Not realizing that EBI BLAST and NCBI BLAST search different databases, so results may differ
- Using EBI BLAST when NCBI BLAST would be more appropriate for the specific database needed

**When to Use It:** Use EBI BLAST when you specifically want to search against UniProtKB, ENA, or other EBI-hosted databases, or when you prefer the EBI interface and tools.

**When NOT to Use It:** For comprehensive searches against all available sequences, NCBI BLAST against the nr database is more appropriate. For very large-scale searches, use DIAMOND.

**Related databases / alternatives:**

- NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>): US equivalent, searches NCBI databases
- HMMER (<https://www.ebi.ac.uk/Tools/hmmer>): More sensitive profile-based search

**How It Connects to Other Resources:** EBI BLAST results link to UniProt, ENA, PDB, and other EBI databases. The EBI Tools API enables integration with other EBI services.

**API / FTP / programmatic access:** EBI Tools REST API (<https://www.ebi.ac.uk/Tools/common/toolbox/help/>) supports programmatic submission and result retrieval in multiple formats.

**Evidence/curation level:** Tool; results depend on the database searched

**Data Update Status:** Databases updated regularly by EMBL-EBI

**Licensing / access restrictions:** Free and open access

**Citation / Recommended Reference:** Madeira F et al. (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*, 50(W1):W276–W279. doi:10.1093/nar/gkac240

**Beginner-Friendly Explanation:** EBI BLAST is the European version of BLAST, maintained by EMBL-EBI. It works the same way as NCBI BLAST but searches against European databases like UniProt (for protein information) and ENA (for DNA sequences). If you want to search specifically against UniProt's carefully reviewed protein database, EBI BLAST is a good choice.

**Advanced Technical Explanation:** EBI BLAST implements the NCBI BLAST+ algorithm through the EBI Tools web services infrastructure, which provides a RESTful API for job submission, status checking, and result retrieval. The API supports asynchronous job submission with polling for completion, enabling integration into automated pipelines. EBI BLAST can search against multiple databases simultaneously, and results are returned in standard BLAST output formats (XML, JSON, plain text).

**One practical workflow example:**

- Step 1: Navigate to <https://www.ebi.ac.uk/Tools/sss/ncbiblast> and select the program type.
- Step 2: Paste your sequence and select the target database (e.g., UniProtKB/Swiss-Prot).
- Step 3: Submit the search and wait for results.
- Step 4: Examine alignments and follow links to UniProt entries for functional information.
- Step 5: For programmatic access, use the EBI Tools REST API to submit jobs and retrieve results.



## D3 – HMMER

**Official Website URL:** <https://www.ebi.ac.uk/Tools/hmmer/> / <https://hmmer.org>

**Resource Type:** Tool (Sequence Similarity Search / Profile HMM)

**Main Biological Domain:** Proteins / DNA sequences

**What It Is Used For:** HMMER is a software suite for sequence analysis using profile hidden Markov models (profile HMMs). It is used for sensitive detection of remote protein homologs, protein family classification, and domain identification. HMMER builds statistical models of protein families from multiple sequence alignments and uses these models to search sequence databases for new members of the family. It is significantly more sensitive than BLAST for detecting distant homologs and is the standard tool for protein family analysis.

**What Data It Contains:** HMMER is a tool, not a database. It searches against sequence databases (UniProtKB, PDB, ENA, or user-provided databases) using profile HMMs. The Pfam database (now part of InterPro) provides a library of pre-built HMMs for known protein families, which can be used with HMMER for domain annotation.

**Main question it helps answer:** Does my protein sequence belong to a known protein family, and what distant homologs exist in sequence databases?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- What protein family does this novel sequence belong to?
- What are the most distant homologs of this protein in the UniProt database?
- What domains are present in this protein sequence?

**Example use cases:** Annotating protein domains in a newly sequenced genome using Pfam HMMs; Detecting remote homologs of a protein of interest that BLAST misses; Building a profile HMM for a protein family and searching for new members

**Input Data Accepted:** Protein sequences in FASTA format; multiple sequence alignments (for building HMMs); pre-built HMM profiles

**Output Data Provided:** Alignment results with E-values and bit scores; domain annotations; HMM profiles

**Strengths:** Significantly more sensitive than BLAST for detecting distant homologs; Profile HMMs capture the statistical properties of protein families, enabling sensitive and specific searches; Pfam/InterPro HMM library provides pre-built models for thousands of protein families; Web interface at EBI and command-line version available; Standard tool for protein family annotation in genome projects

**Limitations:** Slower than BLAST for single-sequence searches; Requires a multiple sequence alignment or pre-built HMM profile for maximum sensitivity; More complex to use than BLAST; requires understanding of HMM concepts; Web interface at EBI may be slow for large queries

**Common beginner mistakes:** Using HMMER when BLAST would be sufficient for closely related sequences; Not understanding the difference between `hmmsearch` (search a database with an HMM) and `hmmsearch` (search a sequence against an HMM database)

- Not filtering results by E-value and domain coverage



**When to Use It:** Use HMMER when BLAST fails to find significant hits for a protein of interest, when you need to classify a protein into a known family, or when you need to annotate protein domains in a genome.

**When NOT to Use It:** For routine sequence identification where BLAST is sufficient, HMMER adds unnecessary complexity. For nucleotide sequence searches, BLAST is generally more appropriate.

**Related databases / alternatives:** NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>): Faster, less sensitive alternative; PSI-BLAST: Iterative BLAST with intermediate sensitivity; InterPro (<https://www.ebi.ac.uk/interpro>): Uses HMMER internally for domain annotation

**How It Connects to Other Resources:** HMMER is used internally by InterPro, Pfam, and other domain databases. HMMER results link to Pfam, UniProt, and PDB records. The Pfam HMM library is available from InterPro.

**API / FTP / programmatic access:** EBI HMMER web service API (<https://www.ebi.ac.uk/Tools/hmmer/>) supports programmatic access. HMMER command-line suite available at <https://hmmer.org> for local installation. Pfam HMM library available from InterPro FTP.

**Data Update Status:** HMMER software updated periodically; Pfam/InterPro HMM library updated with each InterPro release

**Licensing / access restrictions:** Free and open source (BSD license)

**Citation / Recommended Reference:** Eddy SR (2011) Accelerated Profile HMM Searches. PLOS Computational Biology, 7(10):e1002195. doi:10.1371/journal.pcbi.1002195

**Beginner-Friendly Explanation:** HMMER is a more powerful but more complex alternative to BLAST for finding related protein sequences. Instead of comparing your sequence directly to database sequences, HMMER builds a statistical model of a protein family and uses that model to search for new members. This makes it much better at finding distantly related proteins that BLAST might miss. It is widely used for identifying what family a protein belongs to.

**Advanced Technical Explanation:** HMMER implements profile hidden Markov models (profile HMMs) using the Forward/Backward algorithm for probabilistic alignment and the Viterbi algorithm for optimal alignment. The MSV (Multiple Segment Viterbi) filter provides a fast pre-screening step that eliminates most non-homologous sequences before the full HMM comparison. HMMER3 achieves near-BLAST speeds while maintaining HMM sensitivity through the use of SIMD vectorization and the MSV/Viterbi/Forward filter cascade. The hmmbuild program constructs profile HMMs from multiple sequence alignments; hmmsearch searches a sequence database with a profile HMM; hmmscan searches a sequence against a profile HMM database (e.g., Pfam).

#### One practical workflow example:

- Step 1: Navigate to <https://www.ebi.ac.uk/Tools/hmmer> and select "hmmscan" to search your sequence against the Pfam database.
- Step 2: Paste your protein sequence in FASTA format and submit.
- Step 3: Examine the domain hits, noting E-values and domain coverage.
- Step 4: Follow links to Pfam entries for detailed domain descriptions and family alignments.
- Step 5: For command-line use: install HMMER, download the Pfam HMM database, and run: `hmmscan --domtblout results.txt Pfam-A.hmm query.fasta`

## D4 – PSI-BLAST (Position-Specific Iterated BLAST)

**Official Website URL:** <https://blast.ncbi.nlm.nih.gov> (part of BLAST suite)

**Resource Type:** Tool (Sequence Similarity Search)

**Main Biological Domain:** Proteins

**What It Is Used For:** PSI-BLAST (Position-Specific Iterated BLAST) is an iterative sequence similarity search tool that builds a position-specific scoring matrix (PSSM) from the results of an initial BLAST search and uses this matrix to perform subsequent, more sensitive searches. Each iteration can detect more distant homologs than the previous one, making PSI-BLAST significantly more sensitive than standard BLASTp for finding remote protein homologs. It is used for detecting distant evolutionary relationships and for building protein family profiles.

**What Data It Contains:** PSI-BLAST is a tool that searches against NCBI protein databases (nr, RefSeq, UniProt, etc.). It does not contain data itself.

**Main question it helps answer:** What distant protein homologs exist in the database that standard BLAST cannot detect?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- Are there any distant homologs of this protein in organisms where BLAST finds no hits?
- What is the evolutionary relationship between this protein and other protein families?
- Can I build a sensitive profile for this protein family?

**Example use cases:**

- Detecting remote homologs of a protein of interest
- Building a PSSM for use in downstream analyses
- Identifying structural homologs when sequence identity is below 30%

**Input Data Accepted:** Protein sequences in FASTA format

**Output Data Provided:** Alignment results with E-values; PSSM (position-specific scoring matrix) for use in subsequent searches

**Strengths:** More sensitive than standard BLASTp for detecting distant homologs; Iterative approach progressively builds a more sensitive profile; Integrated into the NCBI BLAST web interface and BLAST+ command-line suite; PSSM can be saved and reused for subsequent searches

**Limitations:** Risk of "profile drift" — if false positives are included in early iterations, the profile can diverge from the true family; Requires careful monitoring of each iteration to avoid including non-homologous sequences; Slower than standard BLAST due to iterative nature; Less sensitive than HMMER for very distant homologs

**Common beginner mistakes:** Running too many iterations without checking for false positives; Not setting an appropriate E-value threshold for including sequences in the profile; Using PSI-BLAST when HMMER would be more appropriate for very distant homologs

**When to Use It:** Use PSI-BLAST when standard BLASTp finds no significant hits or only a few hits, and you want to detect more distant homologs. It is intermediate in sensitivity between BLASTp and HMMER.

**When NOT to Use It:** For routine sequence identification, standard BLASTp is sufficient. For very sensitive searches, HMMER is more appropriate. PSI-BLAST is not appropriate for nucleotide sequences.

**Related databases / alternatives:**

- NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>): Standard BLAST, less sensitive
- HMMER (<https://hmmer.org>): More sensitive profile-based search

**How It Connects to Other Resources:** PSI-BLAST is part of the NCBI BLAST suite and searches against NCBI databases. Results link to GenBank, RefSeq, and UniProt records.

**API / FTP / programmatic access:** Available through NCBI BLAST API and BLAST+ command-line suite (psiblast program).

**Evidence/curation level:** Tool; results depend on the database searched

**Data Update Status:** Part of BLAST+ suite; updated with BLAST+ releases

**Licensing / access restrictions:** Free and open access; BLAST+ software in public domain

**Citation / Recommended Reference:** Altschul SF et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402. doi:10.1093/nar/25.17.3389

**Beginner-Friendly Explanation:** PSI-BLAST is an advanced version of BLAST that is better at finding distantly related proteins. It works by doing multiple rounds of searching: in the first round, it finds close relatives; in subsequent rounds, it uses what it learned to find more distant relatives. This makes it much more powerful than regular BLAST for discovering evolutionary relationships, but it requires more careful use to avoid finding false matches.

**Advanced Technical Explanation:** PSI-BLAST constructs a position-specific scoring matrix (PSSM) from the multiple sequence alignment of sequences found in the first BLAST iteration. The PSSM captures position-specific amino acid preferences across the aligned sequences, providing a more sensitive model of the protein family than a single query sequence. In each subsequent iteration, the PSSM is updated with newly found sequences (those meeting the inclusion E-value threshold), progressively refining the family model. The risk of profile drift is managed by setting conservative inclusion thresholds (typically  $E < 0.001$ ) and manually reviewing sequences included in each iteration.

**One practical workflow example:**

- Step 1: Navigate to <https://blast.ncbi.nlm.nih.gov> and select "Protein BLAST" (BLASTp).
- Step 2: Paste your protein sequence and select "PSI-BLAST" from the algorithm options.
- Step 3: Run the first iteration and review the results; note the E-value threshold for inclusion.
- Step 4: Click "Run PSI-BLAST iteration 2" to perform the next iteration with the updated PSSM.
- Step 5: Continue for 3-5 iterations, monitoring for false positives (sequences from unrelated families).
- Step 6: Save the final PSSM for use in downstream analyses.

## D5 – DIAMOND (Double Index Alignment of Next-generation sequencing Data)

**Official Website URL:** <https://github.com/bbuchfink/diamond>

**Resource Type:** Tool (Sequence Similarity Search)

**Main Biological Domain:** Proteins / DNA sequences

**What It Is Used For:** DIAMOND is a high-performance sequence alignment tool designed for searching protein and translated DNA sequences against large databases at speeds up to 100 times faster than BLAST, with comparable sensitivity. It is primarily used for large-scale metagenomic analyses, proteomics database searches, and any application where BLAST is too slow due to the volume of query sequences. DIAMOND is a command-line tool and does not have a web interface.

**What Data It Contains:** DIAMOND is a tool, not a database. It searches against user-provided protein databases (typically formatted from FASTA files using the `diamond makedb` command). It can search against any protein database, including NCBI nr, UniProtKB, or custom databases.

**Main question it helps answer:** What proteins in a large database are similar to my large set of query sequences, at BLAST-like sensitivity but much faster?

**Typical user:** Bioinformatician / Data analyst

**Example scientific questions:**

- What are the taxonomic and functional annotations of millions of reads from a metagenomic sample?
- What proteins in UniProt are similar to all predicted proteins in a newly sequenced genome?
- How can I perform BLAST-like searches on a computing cluster efficiently?

**Example use cases:** Taxonomic classification of metagenomic reads using DIAMOND against the NCBI nr database; Functional annotation of predicted proteins in a genome assembly; Large-scale proteomics database searching

**Input Data Accepted:** Protein sequences or nucleotide sequences (translated on-the-fly) in FASTA format; pre-formatted DIAMOND database

**Output Data Provided:** Alignment results in BLAST tabular format (m8), SAM format, or other formats; taxonomic classification (with `--taxonmap` option)

**Strengths:** 100-500x faster than BLAST for large-scale searches; Comparable sensitivity to BLAST for most applications; Supports taxonomic classification with NCBI taxonomy integration; Memory-efficient; can run on standard workstations for large databases; Actively maintained with regular updates

**Limitations:** Command-line only; no web interface; Requires local installation and database formatting; Slightly lower sensitivity than BLAST for some edge cases; Not appropriate for single-sequence searches where BLAST's web interface is more convenient

**Common beginner mistakes:** Not formatting the database with `diamond makedb` before searching; Using DIAMOND for single-sequence searches where BLAST's web interface is more convenient; Not specifying the output format (default is BLAST tabular, which may not include all desired fields)

**When to Use It:** Use DIAMOND when you have large numbers of query sequences (thousands to millions) and need BLAST-like results quickly, particularly for metagenomic analyses or large-scale genome annotation.

**When NOT to Use It:** For single-sequence searches or small query sets, NCBI BLAST's web interface is more convenient. For very sensitive searches of distant homologs, HMMER is more appropriate.

**Related databases / alternatives:** NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>): Standard tool, slower but with web interface; HMMER (<https://hmmmer.org>): More sensitive for distant homologs

**How It Connects to Other Resources:** DIAMOND output can be used with tools like MEGAN for taxonomic and functional analysis of metagenomic data. DIAMOND databases can be built from any FASTA file, including NCBI nr and UniProtKB.

**API / FTP / programmatic access:** Command-line tool; available via conda (conda install -c bioconda diamond), GitHub (<https://github.com/bbuchfink/diamond>), or pre-compiled binaries.

**Evidence/curation level:** Tool; results depend on the database searched

**Data Update Status:** Actively maintained; regular releases on GitHub

**Licensing / access restrictions:** Free and open source (GPL v3)

**Citation / Recommended Reference:** Buchfink B, Reuter K, Drost HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods, 18:366–368. doi:10.1038/s41592-021-01101-x

**Beginner-Friendly Explanation:** DIAMOND is a very fast version of BLAST designed for analyzing large amounts of sequencing data, like the millions of reads from a metagenomics experiment. It gives results similar to BLAST but can be 100 times faster. It is a command-line tool, so it requires some programming knowledge to use, but it is essential for large-scale bioinformatics analyses.

**Advanced Technical Explanation:** DIAMOND uses a double-index approach that indexes both the query sequences and the database, enabling rapid seed-and-extend alignment without the need to scan the entire database for each query. It implements a reduced amino acid alphabet for seeding (reducing the 20-letter alphabet to 11 letters) and uses SIMD vectorization for the extension step. DIAMOND2 (version 2.0+) introduced a more sensitive mode (--sensitive, --more-sensitive, --very-sensitive, --ultra-sensitive) that approaches BLAST sensitivity at the cost of speed. The --taxonmap option enables taxonomic classification of hits using the NCBI taxonomy database.

#### **One practical workflow example:**

Step 1: Install DIAMOND via conda: conda install -c bioconda diamond

Step 2: Download the NCBI nr database in FASTA format from the NCBI FTP site.

Step 3: Format the database: diamond makedb --in nr.faa -d nr

Step 4: Run the search: diamond blastx -d nr -q reads.fasta -o results.m8 --outfmt 6 qseqid sseqid pident length evalue bitscore staxids

Step 5: Parse the output file (tab-separated) for downstream analysis or import into MEGAN for visualization.

## D6 – FASTA (EBI Sequence Similarity Search)

**Official Website URL:** <https://www.ebi.ac.uk/Tools/sss/fasta>

**Resource Type:** Tool (Sequence Similarity Search)

**Main Biological Domain:** DNA sequences / Proteins

**What It Is Used For:** FASTA is a sequence similarity search tool developed by William Pearson and David Lipman, predating BLAST. The EBI provides a web interface for FASTA searches against EBI-hosted databases. FASTA uses a different algorithm than BLAST (based on the ktup word method and the Smith-Waterman algorithm for final alignment) and can provide different sensitivity/specificity tradeoffs. It is used as an alternative to BLAST, particularly when Smith-Waterman-quality alignments are desired.

**What Data It Contains:** FASTA is a tool that searches against EBI-hosted databases including UniProtKB, ENA, and others.

**Main question it helps answer:** What sequences in EBI databases are similar to my query, using the FASTA algorithm?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- How do FASTA results compare to BLAST results for this query?
- What sequences are similar to my query using Smith-Waterman alignment?

**Example use cases:**

- Cross-validating BLAST results with an independent algorithm
- Searching for sequences when BLAST sensitivity is insufficient

**Input Data Accepted:** Nucleotide or protein sequences in FASTA format

**Output Data Provided:** Alignment results with E-values and scores; Smith-Waterman alignments

**Strengths:**

- Uses Smith-Waterman algorithm for final alignments (more accurate than BLAST's heuristic)
- Alternative algorithm provides independent validation of BLAST results
- Available through EBI web interface

**Limitations:**

- Slower than BLAST for large databases
- Less widely used than BLAST; fewer tutorials and community resources
- EBI Tools web services have undergone restructuring; check current availability

**Common beginner mistakes:**

- Confusing the FASTA file format (a sequence format) with the FASTA search tool (an algorithm)
- Not realizing that FASTA and BLAST use different algorithms and may give different results

**When to Use It:** Use FASTA when you want to cross-validate BLAST results with an independent algorithm, or when you specifically need Smith-Waterman-quality alignments.



**When NOT to Use It:** For routine sequence similarity searches, BLAST is faster and more widely used. For very sensitive searches, HMMER is more appropriate.

**Related databases / alternatives:**

- NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>): Faster, more widely used alternative
- HMMER (<https://hmmer.org>): More sensitive for distant homologs

**How It Connects to Other Resources:** FASTA at EBI links to UniProt, ENA, and other EBI databases.

**API / FTP / programmatic access:** EBI Tools REST API supports FASTA searches programmatically.

**Evidence/curation level:** Tool; results depend on the database searched

**Data Update Status:** Part of EBI Tools suite; updated with EBI database releases

**Licensing / access restrictions:** Free and open access; FASTA software available under open source license

**Citation / Recommended Reference:** Pearson WR (2013) An Introduction to Sequence Similarity ("Homology") Searching. Current Protocols in Bioinformatics, 42:3.1.1–3.1.8. doi:10.1002/0471250953.bi0301s42

**Beginner-Friendly Explanation:** FASTA is one of the original sequence similarity search tools, developed before BLAST. It uses a slightly different method to find similar sequences and can sometimes find things that BLAST misses. The name "FASTA" is also used for the most common DNA/protein sequence file format, which can be confusing — the tool and the file format share the same name but are different things.

**Advanced Technical Explanation:** The FASTA algorithm uses a two-phase approach: first, it identifies regions of high similarity using a rapid word-matching method (ktup), then it uses the Smith-Waterman dynamic programming algorithm to compute optimal local alignments for the top-scoring regions. This provides more accurate alignments than BLAST's heuristic extension, at the cost of speed. The Smith-Waterman algorithm guarantees the optimal local alignment, whereas BLAST's heuristic may miss some optimal alignments.

**One practical workflow example:**

- Step 1: Navigate to <https://www.ebi.ac.uk/Tools/sss/fasta> and select the program type.
- Step 2: Paste your sequence and select the target database.
- Step 3: Submit the search and compare results to a parallel BLAST search.
- Step 4: For sequences where BLAST and FASTA give different results, examine the alignments carefully.
- Step 5: Use the EBI Tools REST API for programmatic access to FASTA searches.



## Beginner Example for category D

---

A student has sequenced a novel gene from a soil bacterium and wants to know what it does. They navigate to NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>) and select BLASTp (protein BLAST) after translating their nucleotide sequence. They paste the protein sequence into the search box, select the nr (non-redundant) database, and submit the search. The results show several hits with E-values below  $1e-10$  and percent identities above 40%, all annotated as "beta-lactamase" in various bacterial species. This strongly suggests that the novel gene encodes a beta-lactamase enzyme.

To confirm this finding and get more detailed functional information, the student follows the link to the top BLAST hit in UniProt, which provides a detailed description of beta-lactamase function, mechanism, and clinical significance. They also note the Pfam domain (PF00144, Beta-lactamase) listed in the UniProt entry, which they can use with HMMER to search for more distant homologs. This workflow — BLAST for initial identification, UniProt for functional details — is the standard approach for characterizing a novel sequence.

## Advanced Research Example for category D

---

A bioinformatician is performing functional annotation of a newly assembled metagenome from a deep-sea sediment sample, containing approximately 5 million predicted protein sequences. Running NCBI BLAST against the nr database for all 5 million sequences would take weeks on a standard server. Instead, they use DIAMOND (<https://github.com/bbuchfink/diamond>) with the `--more-sensitive` flag against a local copy of the NCBI nr database, completing the search in approximately 12 hours on a 32-core server. The DIAMOND output is in BLAST tabular format and is imported into MEGAN for taxonomic and functional classification.

For protein families where DIAMOND finds no significant hits (approximately 30% of predicted proteins), the researcher uses HMMER with the Pfam HMM library to search for domain annotations. This two-step approach — DIAMOND for broad coverage, HMMER for sensitive domain detection — is a standard workflow for metagenomic functional annotation. The researcher records the DIAMOND version, database version, and search parameters for reproducibility, and deposits the annotated metagenome in the ENA with the analysis accession number.

## Common Confusion Points

---

The FASTA file format (a text format for storing sequences, with a header line starting with ">") is completely different from the FASTA search tool (an algorithm for sequence similarity searching). Both are named "FASTA" but are unrelated. When someone says "FASTA format," they mean the file format; when they say "FASTA search," they mean the algorithm.

BLAST E-values depend on database size. The same alignment will have a lower (more significant) E-value when searched against a small database than against a large one, because the probability of finding a match by chance is lower in a smaller database. This means that E-values from different BLAST searches (against different databases) are not directly comparable.

PSI-BLAST and HMMER are both more sensitive than standard BLAST, but they work differently. PSI-BLAST builds a profile iteratively from BLAST results; HMMER builds a profile from a pre-existing multiple sequence alignment. HMMER is generally more sensitive and less prone to profile drift, but requires a pre-existing alignment or HMM profile.

DIAMOND is not a web tool — it is a command-line program that must be installed locally. It is not appropriate for single-sequence searches where BLAST's web interface is more convenient. DIAMOND is designed for large-scale analyses with thousands to millions of query sequences.

Percent identity alone is not a reliable indicator of functional equivalence. Two proteins with 40% identity may have the same function (if the identity is in functionally important regions) or different functions (if the identity is in structurally conserved but functionally variable regions). Always examine the alignment coverage and the functional annotations of the matched sequences, not just the percent identity.

## Which One Should I Use?

---

For most routine sequence identification tasks, start with NCBI BLAST — it is fast, well-documented, and searches comprehensive NCBI databases. Use EBI BLAST when you specifically want to search against UniProtKB or other EBI databases. Use PSI-BLAST when standard BLASTp finds no significant hits and you want to detect more distant protein homologs through iterative searching. Use HMMER when you need the most sensitive detection of remote homologs, when you want to classify a protein into a known family using Pfam, or when you need to annotate protein domains in a genome. Use DIAMOND when you have large numbers of query sequences (thousands to millions) and need BLAST-like results quickly, particularly for metagenomic analyses. Use FASTA (EBI) when you want to cross-validate BLAST results with an independent algorithm or when you need Smith-Waterman-quality alignments.

## Category E: Genome Browsers and Genome Annotation

### Category Overview

Genome browsers are specialized visualization tools that display genomic features along chromosomal coordinates, enabling researchers to explore the genomic context of genes, variants, regulatory elements, and other features. They are essential for interpreting genomic data in its chromosomal context, for integrating multiple data types visually, and for generating hypotheses about gene regulation and function. Unlike sequence databases, which store and retrieve sequences, genome browsers provide an interactive graphical interface for navigating the genome and overlaying multiple annotation tracks.

The major genome browsers — Ensembl, UCSC Genome Browser, NCBI Genome Data Viewer, JBrowse, and IGV — differ in their scope, annotation tracks, customization options, and deployment models. Ensembl and UCSC are web-based browsers with pre-loaded annotation tracks for many species; they are the standard tools for exploring publicly available genome annotations. NCBI Genome Data Viewer provides access to NCBI's genome assemblies and annotations. JBrowse is an open-source, embeddable browser designed for deployment with custom datasets. IGV (Integrative Genomics Viewer) is a desktop application optimized for visualizing personal genomic data, including BAM files from sequencing experiments.

A critical concept in genome browsing is the distinction between the genome assembly (the physical sequence of the genome) and the genome annotation (the features annotated on that sequence). The same genome assembly may have multiple annotation tracks from different sources (e.g., Ensembl gene models, RefSeq gene models, GENCODE annotations), and these may differ in the number and boundaries of annotated genes and transcripts. When using a genome browser for analysis, it is essential to record both the genome assembly version (e.g., GRCh38/hg38) and the annotation version (e.g., Ensembl release 110, GENCODE v44) to ensure reproducibility.

## E1 – Ensembl Genome Browser

**Official Website URL:** <https://www.ensembl.org>

**Resource Type:** Genome Browser / Database

**Main Biological Domain:** DNA sequences / RNA/transcriptomics / Variants

**What It Is Used For:** The Ensembl Genome Browser provides an interactive web interface for exploring genome sequences, gene annotations, transcript isoforms, regulatory features, genetic variants, and comparative genomics data for vertebrates and selected other eukaryotes. It is used for browsing genomic regions, retrieving gene and transcript information, visualizing variants in genomic context, and accessing comparative genomics data. Ensembl is the primary genome browser for researchers working with Ensembl gene IDs and GENCODE annotations.

**What Data It Contains:** Ensembl contains genome assemblies and annotations for over 300 species, including gene models, transcript isoforms, regulatory features, genetic variants, comparative genomics data (gene trees, synteny, whole-genome alignments), and expression data. The human and mouse genomes use GENCODE annotations, which are the most comprehensive and manually curated gene annotations available.

**Main question it helps answer:** What genes, transcripts, variants, and regulatory features are present at this genomic location?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:**

- What transcripts are produced from this gene, and what are their exon structures?
- What variants are present in the coding region of this gene?
- What are the orthologs of this gene in other vertebrates?

**Example use cases:** Visualizing the exon structure of a gene for primer design; Checking variant consequences using the Variant Effect Predictor (VEP); Downloading GTF annotation files for RNA-seq analysis

**Input Data Accepted:** Gene names, Ensembl IDs, genomic coordinates, variant IDs, species names

**Output Data Provided:** Gene and transcript annotations, genomic sequences, variant data, regulatory features, comparative genomics data, GTF/GFF annotation files

**Strengths:** Comprehensive, regularly updated genome annotations for many species; GENCODE annotations for human and mouse (most comprehensive available); Variant Effect Predictor (VEP) for variant annotation; BioMart for bulk data retrieval; REST API for programmatic access

**Limitations:** Ensembl gene IDs change between releases; version tracking is essential; Web interface can be slow for large genomic regions; Annotation quality varies by species

**Common beginner mistakes:** Not recording the Ensembl release version when using Ensembl IDs; Confusing Ensembl gene IDs with NCBI Gene IDs; Not checking that the genome assembly version matches the one used for alignment

**When to Use It:** Use Ensembl when you need comprehensive genome annotation for vertebrates, when you are working with RNA-seq data and need transcript coordinates, or when you want to use VEP for variant annotation.



**When NOT to Use It:** For prokaryotic genomes, use NCBI RefSeq. For clinical variant interpretation, ClinVar provides more clinically relevant information.

**Related databases / alternatives:**

- UCSC Genome Browser (<https://genome.ucsc.edu>): Alternative browser with different annotation tracks
- NCBI Genome Data Viewer (<https://www.ncbi.nlm.nih.gov/genome/gdv>): NCBI's genome browser

**How It Connects to Other Resources:** Ensembl links to UniProt, NCBI Gene, PDB, GO, Reactome, OMIM, and many other resources. VEP integrates with ClinVar, dbSNP, and gnomAD.

**API / FTP / programmatic access:** Ensembl REST API (<https://rest.ensembl.org>); BioMart; FTP at <ftp://ftp.ensembl.org>; biomaRt R package; pybiomart Python package.

**Evidence/curation level:** Mixed — computationally predicted gene models with manual curation for human/mouse (GENCODE)

**Data Update Status:** Regular numbered releases (approximately quarterly)

**Licensing / access restrictions:** Open access; Apache 2.0 license

**Citation / Recommended Reference:** Martin FJ et al. (2023) Ensembl 2023. Nucleic Acids Research, 51(D1):D933–D941. doi:10.1093/nar/gkac958

**Beginner-Friendly Explanation:** Ensembl is a genome browser that lets you explore the genomes of humans and hundreds of other species. You can look up any gene and see where it is in the genome, what different versions (transcripts) of it exist, and what variants have been found in it. It is especially important for RNA-seq analysis because it provides the gene coordinate files used to count reads.

**Advanced Technical Explanation:** Ensembl uses an automated gene annotation pipeline integrating protein alignments, cDNA/EST alignments, and ab initio predictions. The REST API implements content negotiation supporting JSON, XML, FASTA, GFF3, and BED output. VEP uses the Ensembl transcript database to annotate variants with predicted consequences and integrates with multiple external databases.

**One practical workflow example:**

Step 1: Navigate to <https://www.ensembl.org> and search for your gene.

Step 2: Examine transcript isoforms and exon structures in the gene view.

Step 3: Use BioMart to download transcript coordinates and cross-references.

Step 4: Download the GTF annotation file from the FTP site for RNA-seq analysis.

Step 5: Run VEP on a VCF file to annotate variants with predicted consequences.

Step 6: Record the Ensembl release version for reproducibility.

## E2 – UCSC Genome Browser

**Official Website URL:** <https://genome.ucsc.edu>

**Resource Type:** Genome Browser

**Main Biological Domain:** DNA sequences / Variants / Epigenomics

**What It Is Used For:** The UCSC Genome Browser is a web-based genome browser developed at the University of California Santa Cruz, providing an interactive interface for exploring the human and many other genomes with hundreds of pre-loaded annotation tracks. It is used for visualizing genomic features, overlaying multiple data tracks, exploring regulatory elements, and accessing ENCODE and other large-scale genomics datasets.

**What Data It Contains:** The UCSC Genome Browser hosts genome assemblies and annotation tracks for over 100 species, including gene annotations (RefSeq, GENCODE, Ensembl), conservation scores, regulatory elements (ENCODE tracks), genetic variants (dbSNP, ClinVar), epigenomic data (ChIP-seq, ATAC-seq, Hi-C), and many other data types. The ENCODE data portal is tightly integrated with the UCSC Browser.

**Main question it helps answer:** What genomic features and regulatory elements are present at this chromosomal location?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:** What regulatory elements (enhancers, promoters) are near this gene? | What is the conservation score of this genomic region across vertebrates? | What ENCODE ChIP-seq peaks overlap with this variant?

**Example use cases:** Visualizing ENCODE regulatory data in the context of a gene of interest; Checking conservation of a genomic region across species; Extracting sequence or annotation data using the Table Browser.

**Input Data Accepted:** Genomic coordinates (chr:start-end), gene names, accession numbers, variant IDs; custom tracks in BED, BedGraph, BAM, VCF formats

**Output Data Provided:** Interactive genome visualization with multiple annotation tracks; sequence and annotation data via Table Browser; custom track display

**Strengths:** Extensive collection of pre-loaded annotation tracks (hundreds of tracks for human genome); ENCODE data integration provides comprehensive regulatory annotation; Table Browser enables extraction of annotation data in tabular format; Custom track support for uploading personal data.

**Limitations:** Interface can be overwhelming due to the large number of tracks; Some tracks are only available for human and mouse; coverage of other species is more limited; Table Browser queries can be slow for large genomic regions; Coordinate system differences between assemblies (hg19 vs. hg38) require careful attention

**Common beginner mistakes:** Not checking which genome assembly version (hg19/GRCh37 vs. hg38/GRCh38) is being used; Being overwhelmed by the number of tracks and not knowing which ones to enable; Confusing UCSC gene annotations with Ensembl or RefSeq annotations

**When to Use It:** Use the UCSC Genome Browser when you need to visualize regulatory elements, conservation data, or ENCODE tracks in genomic context, or when you need to extract annotation data using the Table Browser.





**When NOT to Use It:** For variant effect prediction, Ensembl VEP is more appropriate. For downloading genome annotation files for bioinformatics pipelines, Ensembl or RefSeq FTP sites are more convenient.

**Related databases / alternatives:** Ensembl (<https://www.ensembl.org>): Alternative browser with different annotation focus; NCBI Genome Data Viewer (<https://www.ncbi.nlm.nih.gov/genome/gdv>): NCBI's browser

**How It Connects to Other Resources:** UCSC Browser links to NCBI, Ensembl, ENCODE, dbSNP, ClinVar, and many other resources. The Table Browser can export data in formats compatible with downstream analysis tools.

**API / FTP / programmatic access:** UCSC DAS server; REST API (<https://api.genome.ucsc.edu>); Table Browser for data extraction; FTP at <ftp://hgdownload.soe.ucsc.edu> for bulk downloads.

**Evidence/curation level:** Mixed — tracks from multiple sources with varying curation levels

**Data Update Status:** Continuous updates; new tracks added regularly; genome assemblies updated as new versions are released

**Licensing / access restrictions:** Free and open access; data from individual tracks subject to their own licenses

**Citation / Recommended Reference:** Nassar LR et al. (2023) The UCSC Genome Browser database: 2023 update. Nucleic Acids Research, 51(D1):D1188–D1195. doi:10.1093/nar/gkac1072

**Beginner-Friendly Explanation:** The UCSC Genome Browser is like a map of the human genome that you can zoom in and out of. It shows you where genes are, what regulatory elements are nearby, how conserved a region is across different species, and much more. It has hundreds of different "tracks" (layers of information) that you can turn on and off. It is one of the most powerful tools for exploring the genome visually.

**Advanced Technical Explanation:** The UCSC Genome Browser uses a MySQL database backend with a custom binary format (bigBed, bigWig) for efficient storage and retrieval of large genomic datasets. The REST API (<https://api.genome.ucsc.edu>) provides programmatic access to track data, sequence, and annotation in JSON format. The Table Browser implements a SQL-like query interface for extracting annotation data from any track in tabular format. Custom tracks can be uploaded in BED, BedGraph, BAM, CRAM, VCF, and other formats for visualization alongside pre-loaded tracks.

### One practical workflow example:

Step 1: Navigate to <https://genome.ucsc.edu> and select the genome assembly (e.g., Human GRCh38/hg38).

Step 2: Enter a gene name or genomic coordinates in the search box.

Step 3: Enable relevant annotation tracks (e.g., GENCODE genes, ENCODE regulatory elements, conservation).

Step 4: Use the Table Browser (Tools > Table Browser) to extract annotation data for a genomic region.

Step 5: Upload a custom BED or VCF file as a custom track to visualize your own data alongside pre-loaded tracks.

Step 6: Use the REST API for programmatic access: GET <https://api.genome.ucsc.edu/getData/track?genome=hg38;track=knownGene;chrom=chr17;start=7668421;end=7687490>



## E3 – NCBI Genome Data Viewer

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/genome/gdv>

**Resource Type:** Genome Browser

**Main Biological Domain:** DNA sequences / Variants

**What It Is Used For:** The NCBI Genome Data Viewer (GDV) is NCBI's web-based genome browser, providing visualization of NCBI genome assemblies and annotations. It is used for exploring RefSeq gene annotations, viewing variants from dbSNP and ClinVar, and visualizing genomic features in the context of NCBI's reference assemblies. GDV is particularly useful for accessing NCBI-specific annotations and for researchers who prefer to work within the NCBI ecosystem.

**What Data It Contains:** GDV provides access to NCBI genome assemblies and RefSeq annotations, including gene models, transcripts, variants (dbSNP, ClinVar), and other genomic features. It covers all organisms with genome assemblies in the NCBI Assembly database.

**Main question it helps answer:** What RefSeq genes and variants are present at this genomic location in the NCBI reference assembly?

**Typical user:** Researcher / Bioinformatician / Clinician

**Example scientific questions:**

- What RefSeq transcripts are annotated at this genomic location?
- What ClinVar variants are present in this gene?
- What is the genomic context of this variant in the NCBI reference assembly?

**Example use cases:**

- Visualizing RefSeq gene annotations for a genomic region
- Checking ClinVar variant classifications in genomic context
- Exploring genome assemblies for non-model organisms

**Input Data Accepted:** Genomic coordinates, gene names, accession numbers, variant IDs

**Output Data Provided:** Interactive genome visualization with RefSeq annotations, variant tracks, and other NCBI data

**Strengths:**

- Direct access to NCBI RefSeq annotations and variant databases
- Covers all organisms with NCBI genome assemblies
- Integrated with ClinVar and dbSNP for variant visualization
- Part of the NCBI ecosystem with links to other NCBI databases

**Limitations:**

- Fewer pre-loaded annotation tracks than UCSC Browser
- Less widely used than Ensembl or UCSC; fewer tutorials
- Interface is less feature-rich than Ensembl or UCSC

**Common beginner mistakes:**

- Not realizing that GDV uses RefSeq annotations, which may differ from Ensembl annotations
- Overlooking GDV for non-model organisms where UCSC and Ensembl have limited coverage

**When to Use It:** Use GDV when you need to visualize RefSeq annotations or NCBI variant data, or when you are working with organisms not covered by Ensembl or UCSC.

**When NOT to Use It:** For comprehensive regulatory annotation, UCSC Browser is more appropriate. For Ensembl-specific annotations, use the Ensembl Browser.

**Related databases / alternatives:** Ensembl (<https://www.ensembl.org>): More comprehensive for vertebrates; UCSC Genome Browser (<https://genome.ucsc.edu>): More annotation tracks

**How It Connects to Other Resources:** GDV links to NCBI Gene, RefSeq, ClinVar, dbSNP, and other NCBI databases.

**API / FTP / programmatic access:** NCBI E-utilities API for programmatic access to underlying data; FTP for genome assembly downloads.

**Evidence/curation level:** Mixed — RefSeq annotations (mixed curation); variant data from dbSNP and ClinVar (varying curation levels)

**Data Update Status:** Updated with RefSeq releases and NCBI database updates

**Licensing / access restrictions:** Free and open access

**Citation / Recommended Reference:** Sayers EW et al. (2022) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 50(D1):D20–D26. doi:10.1093/nar/gkab1112

**Beginner-Friendly Explanation:** The NCBI Genome Data Viewer is NCBI's own genome browser, similar to the UCSC Genome Browser but using NCBI's own gene annotations (RefSeq). It is useful for looking at genes and variants in the context of NCBI's reference genome sequences. If you are already working with NCBI data, GDV provides a convenient way to visualize it.

**Advanced Technical Explanation:** GDV uses NCBI's genome assembly database as its backbone, displaying RefSeq annotations alongside variant data from dbSNP and ClinVar. It supports custom track upload in standard formats (BED, VCF, BAM) and provides links to NCBI's sequence viewer for detailed sequence-level analysis. GDV is particularly useful for organisms with NCBI genome assemblies but limited coverage in Ensembl or UCSC.

**One practical workflow example:**

- Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/genome/gdv> and select an organism.
- Step 2: Enter a gene name or genomic coordinates to navigate to the region of interest.
- Step 3: Enable ClinVar and dbSNP tracks to visualize variants.
- Step 4: Click on individual features to access linked NCBI database records.
- Step 5: Use the "Download" option to export sequence or annotation data.

## E4 – JBrowse

**Official Website URL:** <https://jbrowse.org>

**Resource Type:** Genome Browser (Open-source, deployable)

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** JBrowse is an open-source, JavaScript-based genome browser designed for deployment with custom datasets. It is used by research groups, databases, and genome projects to provide interactive genome visualization for their own data. JBrowse 2 (the current version) supports a wide range of genomic data types and can be deployed as a standalone web application or embedded in other websites. It is the browser of choice for many specialized genome databases and for researchers who need to share their own genomic data interactively.

**What Data It Contains:** JBrowse is a browser framework, not a database. It displays data provided by the deploying organization or user, which can include genome sequences, gene annotations, BAM/CRAM alignments, VCF variants, BED features, BigWig signal tracks, and many other formats.

**Main question it helps answer:** How can I visualize and share my own genomic data interactively?

**Typical user:** Bioinformatician / Database developer / Researcher (advanced)

**Example scientific questions:**

- How can I create an interactive genome browser for my newly sequenced organism?
- How can I share my RNA-seq alignment data with collaborators in a browser interface?
- What does my genome assembly look like with my custom annotations?

**Example use cases:**

- Deploying a genome browser for a newly sequenced organism
- Sharing BAM alignment files with collaborators through a web interface
- Embedding a genome browser in a database or publication website

**Input Data Accepted:** FASTA genome sequences, GFF/GTF annotations, BAM/CRAM alignments, VCF variants, BED features, BigWig signal tracks

**Output Data Provided:** Interactive genome visualization; data export in various formats

**Strengths:** Open-source and freely deployable; Supports a wide range of genomic data formats; JBrowse 2 has a modern, extensible architecture; Can be embedded in other websites; Active development community

**Limitations:** Requires technical expertise to deploy and configure; Not a hosted service with pre-loaded data (unlike Ensembl or UCSC); Requires a web server for deployment

**Common beginner mistakes:**

- Confusing JBrowse (a deployable browser) with hosted browsers like Ensembl or UCSC
- Not realizing that JBrowse requires data to be provided by the user

**When to Use It:** Use JBrowse when you need to deploy a genome browser for your own data, when you want to share genomic data interactively with collaborators, or when you are building a genome database.

**When NOT to Use It:** For exploring publicly available genome annotations, use Ensembl or UCSC. JBrowse is not appropriate for users who need a pre-loaded browser with public data.

**Related databases / alternatives:**

- IGV (<https://igv.org>): Desktop application for personal genomic data visualization
- Ensembl (<https://www.ensembl.org>): Hosted browser with pre-loaded data
- UCSC Genome Browser (<https://genome.ucsc.edu>): Hosted browser with pre-loaded data

**How It Connects to Other Resources:** JBrowse can display data from any source that provides standard genomic data formats. JBrowse 2 supports remote data sources via URLs, enabling integration with cloud-hosted data.

**API / FTP / programmatic access:** JBrowse 2 has a plugin API for extending functionality; data is accessed via standard file formats and URLs.

**Evidence/curation level:** Tool/framework; data quality depends on the deploying organization

**Data Update Status:** Actively maintained; regular releases on GitHub

**Licensing / access restrictions:** Free and open source (Apache 2.0 license)

**Citation / Recommended Reference:** Diesh C et al. (2023) JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biology*, 24:74. doi:10.1186/s13059-023-02914-z

**Beginner-Friendly Explanation:** JBrowse is a genome browser that you can install and run yourself, rather than using a website like Ensembl or UCSC. It is used by research groups and databases to create their own genome browsers for their own data. If you have sequenced a new organism and want to share the genome with others in an interactive way, JBrowse is a good tool to use.

**Advanced Technical Explanation:** JBrowse 2 is built on a React/TypeScript framework with a plugin-based architecture that enables extensible data adapters, view types, and track types. It supports lazy loading of remote data via HTTP range requests, enabling efficient visualization of large datasets without downloading entire files. JBrowse 2 supports multiple view types including linear genome view, circular genome view, dotplot view (for synteny), and breakpoint split view (for structural variants). The configuration is JSON-based and can be generated programmatically.

**One practical workflow example:**

- Step 1: Install JBrowse 2 CLI: `npm install -g @jbrowse/cli`
- Step 2: Create a new JBrowse instance: `jbrowse create /path/to/jbrowse`
- Step 3: Add a genome assembly: `jbrowse add-assembly genome.fasta --load copy`
- Step 4: Add annotation tracks: `jbrowse add-track genes.gff3 --load copy`
- Step 5: Add alignment tracks: `jbrowse add-track alignments.bam --load copy`
- Step 6: Start a web server and open JBrowse in a browser to visualize the data.

## E5 – IGV (Integrative Genomics Viewer)

---

**Database Name:** IGV (Integrative Genomics Viewer)

**Official Website URL:** <https://igv.org>

**Resource Type:** Genome Browser (Desktop Application)

**Main Biological Domain:** DNA sequences / Variants / Epigenomics / RNA/transcriptomics

**What It Is Used For:** IGV is a high-performance desktop genome browser developed at the Broad Institute, designed for interactive visualization of large genomic datasets including BAM/CRAM alignment files, VCF variant files, BED feature files, and BigWig signal tracks. It is the standard tool for visualizing personal genomic data from sequencing experiments, including RNA-seq alignments, ChIP-seq peaks, and variant calls. IGV is particularly valued for its ability to handle large files efficiently and for its integration with cloud storage.

**What Data It Contains:** IGV is a browser application, not a database. It displays data loaded by the user from local files or remote URLs. It includes built-in reference genomes (human, mouse, and others) and can load annotation tracks from UCSC, Ensembl, and other sources.

**Main question it helps answer:** What do my sequencing alignments, variants, and other genomic data look like in the context of the reference genome?

**Typical user:** Bioinformatician / Researcher / Clinician

**Example scientific questions:**

- Do my RNA-seq reads align correctly to the exons of this gene?
- What does the variant call look like in the raw alignment data?
- Are there any alignment artifacts near this variant?

**Example use cases:** Visualizing BAM files from RNA-seq or WGS experiments; Manually inspecting variant calls to distinguish true variants from artifacts; Visualizing ChIP-seq or ATAC-seq signal tracks alongside gene annotations

**Input Data Accepted:** BAM/CRAM alignment files, VCF variant files, BED feature files, BigWig/BigBed signal tracks, GFF/GTF annotation files, FASTA sequences

**Output Data Provided:** Interactive genome visualization; screenshots; data export

**Strengths:** Handles large BAM/CRAM files efficiently; Supports cloud storage (Google Cloud, AWS S3, Azure) for remote file access; Available as desktop application, web application (igv.js), and command-line tool; Widely used and well-documented; Supports many genomic data formats

**Limitations:** Desktop application requires local installation (though web version available); Not appropriate for sharing data publicly (use JBrowse for that); Large files may be slow to load on computers with limited memory

**Common beginner mistakes:** Not indexing BAM files (with samtools index) before loading in IGV; Not ensuring that the reference genome in IGV matches the genome used for alignment; Confusing IGV (for personal data) with Ensembl/UCSC (for public data)

**When to Use It:** Use IGV when you need to visualize your own sequencing data (BAM files, VCF files), when you need to manually inspect variant calls, or when you need to check alignment quality.

**When NOT to Use It:** For exploring publicly available genome annotations, use Ensembl or UCSC. For sharing data with others, use JBrowse or a hosted browser.

**Related databases / alternatives:** JBrowse (<https://jbrowse.org>): Open-source deployable browser; Ensembl (<https://www.ensembl.org>): Hosted browser for public data; UCSC Genome Browser (<https://genome.ucsc.edu>): Hosted browser for public data

**How It Connects to Other Resources:** IGV can load annotation tracks from UCSC, Ensembl, and other sources via URLs. It integrates with cloud storage for remote file access.

**API / FTP / programmatic access:** igv.js (<https://github.com/igvteam/igv.js>) provides a JavaScript library for embedding IGV in web applications. IGV command-line batch mode enables automated screenshot generation.

**Evidence/curation level:** Tool; data quality depends on the user's data

**Data Update Status:** Actively maintained; regular releases

**Licensing / access restrictions:** Free and open source (MIT license)

**Citation / Recommended Reference:** Robinson JT et al. (2011) Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26. doi:10.1038/nbt.1754

**Beginner-Friendly Explanation:** IGV is a program you install on your computer to look at your own sequencing data. If you have done RNA-seq or whole-genome sequencing, IGV lets you see exactly where your reads aligned to the genome, which is essential for checking the quality of your data and for manually inspecting variants. It is one of the most important tools for anyone working with sequencing data.

**Advanced Technical Explanation:** IGV uses a tile-based rendering approach that enables efficient visualization of large genomic datasets by loading only the data visible in the current view. It supports indexed file formats (BAM/CRAM with .bai/.crai index, VCF with .tbi index, BigBed/BigWig) for random access to specific genomic regions. The igv.js library implements the same rendering engine as the desktop application in JavaScript, enabling embedding in web applications. IGV supports the GA4GH htsget protocol for streaming genomic data from cloud storage.

#### **One practical workflow example:**

- Step 1: Download and install IGV from <https://igv.org/doc/desktop/>
- Step 2: Select the reference genome matching your alignment (e.g., Human (hg38)).
- Step 3: Load your BAM file (File > Load from File); ensure the BAM index (.bai) is in the same directory.
- Step 4: Navigate to a gene or genomic region of interest using the search box.
- Step 5: Load a VCF file to overlay variant calls on the alignment.
- Step 6: Zoom in to individual reads to inspect alignment quality and variant evidence.



## Beginner Example (Category E)

---

A student has just completed an RNA-seq experiment and wants to check whether their reads are aligning correctly to the gene of interest. They download IGV (<https://igv.org>) and load their BAM file along with the human reference genome (hg38). They navigate to the gene of interest and can see the read alignments across the exons, confirming that the reads are mapping to the expected locations. They notice that one exon has very few reads, which they investigate further.

To understand the gene structure better, the student also opens the Ensembl Genome Browser (<https://www.ensembl.org>) and searches for the same gene. They can see all the annotated transcript isoforms and compare the exon structure to what they observed in IGV. They download the GTF annotation file from Ensembl for use in their differential expression analysis, recording the Ensembl release version (e.g., release 110) for their methods section.

## Advanced Research Example (Category E)

---

A bioinformatician is analyzing whole-genome sequencing data from a cancer patient and needs to visualize somatic variants in their genomic context. They use IGV to load the tumor and normal BAM files alongside a VCF file of somatic variant calls. For each candidate variant, they manually inspect the alignment in IGV to distinguish true somatic mutations from alignment artifacts, strand bias, or low-quality reads. They use the UCSC Genome Browser to check whether any variants overlap with known regulatory elements (ENCODE tracks) or conserved regions (PhyloP conservation scores).

For variants in coding regions, they use the Ensembl Variant Effect Predictor (VEP) to annotate predicted functional consequences and check ClinVar for clinical significance. They also use the UCSC Table Browser to extract all ENCODE ChIP-seq peaks overlapping a specific genomic region, which they then visualize as a custom track in IGV. This multi-browser workflow — IGV for personal data, UCSC for regulatory context, Ensembl for variant annotation — is a standard approach for comprehensive genomic variant analysis.

## Common Confusion Points (Category E)

---

Ensembl and UCSC use different gene annotation systems. The same gene may have different transcript boundaries, different numbers of isoforms, and different exon structures in Ensembl vs. UCSC (RefSeq) annotations. When comparing results from different analyses, always check that the same annotation source was used.

The genome assembly version (e.g., GRCh38/hg38 vs. GRCh37/hg19) is different from the annotation version (e.g., Ensembl release 110, GENCODE v44). Both must be recorded for reproducibility. A common error is using hg38 coordinates with hg19 annotations, which produces incorrect results.



IGV is for visualizing your own data; Ensembl and UCSC are for exploring publicly available annotations. IGV does not contain pre-loaded annotation data (beyond basic reference genomes); it displays data that you provide. Ensembl and UCSC contain extensive pre-loaded annotation tracks but are not designed for visualizing personal BAM files.

JBrowse is a browser framework, not a hosted service. Unlike Ensembl and UCSC, JBrowse does not have a central website with pre-loaded data. It must be deployed by the user with their own data. Many specialized genome databases use JBrowse as their browser, so you may encounter JBrowse instances at various URLs.

Genome browsers display annotations; they do not perform analysis. A genome browser shows you what is known about a genomic region, but it does not perform statistical analysis, variant calling, or differential expression analysis. For analysis, you need dedicated bioinformatics tools; genome browsers are for visualization and exploration.

### Which One Should I Use? (Category E)

---

For exploring publicly available genome annotations for vertebrates, use Ensembl (particularly for Ensembl/Gencode annotations, VEP, and BioMart) or the UCSC Genome Browser (particularly for regulatory annotation, ENCODE data, and the Table Browser). Use NCBI Genome Data Viewer when you need to visualize RefSeq annotations or NCBI variant data, or when working with organisms not covered by Ensembl or UCSC. Use IGV for visualizing your own sequencing data (BAM files, VCF files) — it is the standard tool for manual inspection of alignment data. Use JBrowse when you need to deploy a genome browser for your own data or when building a genome database. For most researchers, the combination of Ensembl (for annotation) and IGV (for personal data) covers the majority of genome browsing needs.

## Category F: Gene and Genome Databases

### Category Overview

Gene and genome databases provide structured information about individual genes — their identifiers, nomenclature, genomic locations, functional annotations, disease associations, and cross-references to other resources. While genome browsers display genes in their chromosomal context, gene databases provide gene-centric records that aggregate information from multiple sources into a single, searchable entry. These databases are essential for standardizing gene identifiers across analyses, for retrieving comprehensive gene information, and for linking genes to diseases, pathways, and other biological entities.

The gene database landscape includes resources with different scopes and emphases. NCBI Gene provides comprehensive gene records for all organisms with sequenced genomes, integrating information from RefSeq, UniProt, GO, and many other sources. GeneCards provides an integrated summary of human gene information from dozens of sources, making it a convenient starting point for human gene research. HGNC (HUGO Gene Nomenclature Committee) is the authoritative source for human gene names and symbols, providing the standardized nomenclature used in publications and databases worldwide. OMIM (Online Mendelian Inheritance in Man) is the definitive resource for Mendelian disease genetics, providing manually curated entries for genes and phenotypes. Ensembl Genes provides genome-wide gene annotations for vertebrates and other eukaryotes, with a focus on computational annotation and comparative genomics.

A critical issue in gene databases is identifier consistency. Different databases use different gene identifiers: NCBI Gene uses integer IDs (e.g., 672 for human BRCA1), Ensembl uses ENSG IDs (e.g., ENSG00000012048), HGNC uses HGNC IDs (e.g., HGNC:1100), and UniProt uses accession numbers (e.g., P38398). These identifiers are not interchangeable, and converting between them requires explicit ID mapping. Tools such as BioMart, the HGNC Multi-Symbol Checker, and the UniProt ID mapping service provide this functionality. When reporting results, always specify which identifier system was used and provide cross-references to other systems where possible.

## F1 – NCBI Gene

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/gene>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** NCBI Gene is a comprehensive gene database maintained by NCBI, providing gene-centric records that integrate information from RefSeq, UniProt, GO, OMIM, dbSNP, GEO, and many other sources. It is used for retrieving comprehensive gene information, finding gene identifiers, accessing gene annotations, and navigating between related resources. NCBI Gene is the primary resource for NCBI Gene IDs, which are widely used in bioinformatics tools and databases.

**What Data It Contains:** NCBI Gene contains records for genes from all organisms with sequenced genomes, including gene symbols, names, aliases, genomic locations, RefSeq sequences, protein products, GO annotations, pathway associations, disease associations (OMIM), variant data (dbSNP, ClinVar), expression data (GEO), and cross-references to dozens of other databases. As of 2024, NCBI Gene contains records for genes from over 30,000 organisms.

**Main question it helps answer:** What is known about this gene across all levels of biological data?

**Typical user:** Beginner student / Researcher / Clinician / Bioinformatician

**Example scientific questions:**

- What is the official symbol and name for this gene?
- What RefSeq transcripts are associated with this gene?
- What diseases are associated with mutations in this gene?

**Example use cases:** Looking up the NCBI Gene ID for a gene to use in bioinformatics tools; Finding all RefSeq transcripts for a gene for RNA-seq analysis; Checking disease associations and OMIM entries for a gene of interest

**Input Data Accepted:** Gene names, symbols, NCBI Gene IDs, organism names, keywords

**Output Data Provided:** Gene records with genomic location, RefSeq sequences, protein products, GO annotations, disease associations, expression data, and cross-references

**Strengths:** Comprehensive integration of information from many sources; Covers all organisms with sequenced genomes; NCBI Gene IDs are stable and widely used; Links to RefSeq, UniProt, OMIM, GO, and many other resources; E-utilities API for programmatic access

**Limitations:** Annotation quality varies by organism; model organisms are well-annotated; NCBI Gene IDs differ from Ensembl IDs; conversion requires explicit mapping; Some information is automatically imported and may not be manually reviewed

**Common beginner mistakes:** Confusing NCBI Gene IDs with Ensembl IDs or other identifier systems; Not checking the organism when searching for a gene; Assuming all information in a Gene record is manually reviewed

**When to Use It:** Use NCBI Gene as the primary resource for gene information within the NCBI ecosystem, for finding NCBI Gene IDs, and for accessing integrated gene annotations.



**When NOT to Use It:** For Ensembl-specific annotations, use Ensembl. For detailed protein functional annotations, use UniProt. For disease genetics, OMIM provides more detailed clinical information.

**Related databases / alternatives:** Ensembl (<https://www.ensembl.org>): Alternative gene annotation system; GeneCards (<https://www.genecards.org>): Integrated human gene information; HGNC (<https://www.genenames.org>): Authoritative human gene nomenclature

**How It Connects to Other Resources:** NCBI Gene links to RefSeq, UniProt, OMIM, GO, dbSNP, ClinVar, GEO, Reactome, KEGG, and many other databases. Gene IDs are used as cross-references in hundreds of bioinformatics tools.

**API / FTP / programmatic access:** E-utilities API (esearch and efetch with db=gene) for programmatic access. FTP at <ftp://ftp.ncbi.nlm.nih.gov/gene/> for bulk downloads including gene2refseq, gene2go, and gene\_info files.

**Evidence/curation level:** Mixed — automatically imported from RefSeq, UniProt, GO, and other sources; some manual curation for high-priority genes

**Data Update Status:** Continuous updates reflecting changes in source databases

**Licensing / access restrictions:** Open access

**Citation / Recommended Reference:** Sayers EW et al. (2022) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 50(D1):D20–D26. doi:10.1093/nar/gkab1112

**Beginner-Friendly Explanation:** NCBI Gene is a database that collects everything known about a gene in one place. If you search for a gene by name, you will find its official symbol, where it is in the genome, what protein it makes, what diseases it is associated with, and links to many other databases. It is a great starting point for learning about any gene.

**Advanced Technical Explanation:** NCBI Gene implements a gene-centric data model that aggregates information from multiple NCBI databases and external sources through automated import pipelines. The gene2refseq file (available via FTP) provides mappings between Gene IDs and RefSeq accession numbers, enabling programmatic cross-referencing. The gene2go file provides GO annotations for all genes. The E-utilities API supports complex queries with field tags including [Gene/Protein Name], [Organism], and [Gene ID].

#### **One practical workflow example:**

Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/gene> and search for your gene (e.g., "BRCA1 Homo sapiens").

Step 2: Examine the gene record for official symbol, genomic location, and RefSeq transcripts.

Step 3: Click "OMIM" to access disease associations, or "GO" for functional annotations.

Step 4: Use the "RefSeq" link to access curated reference sequences.

Step 5: For programmatic access, download the gene\_info and gene2refseq files from the NCBI FTP site for bulk ID mapping.

## F2 – GeneCards

**Official Website URL:** <https://www.genecards.org>

**Resource Type:** Knowledgebase / Integrated Portal

**Main Biological Domain:** DNA sequences / Proteins / Diseases / Omics

**What It Is Used For:** GeneCards is an integrated human gene database that aggregates information from over 150 databases and resources into a single, comprehensive gene card for each human gene. It is used for quickly accessing a broad overview of a gene's function, expression, disease associations, variants, pathways, and interactions. GeneCards is particularly valued for its comprehensive integration of diverse data sources and its user-friendly interface.

**What Data It Contains:** GeneCards contains information for all human genes (protein-coding, non-coding RNA, pseudogenes), including gene function, protein information, expression data (tissue and cell type), disease associations, genetic variants, pathways, protein interactions, drug associations, and cross-references to hundreds of databases. Data is automatically imported from sources including UniProt, Ensembl, NCBI Gene, OMIM, ClinVar, GTEx, and many others.

**Main question it helps answer:** What is the comprehensive biological and clinical profile of this human gene?

**Typical user:** Researcher / Clinician / Wet-lab scientist / Beginner student

**Example scientific questions:**

- What diseases are associated with this gene?
- In what tissues is this gene expressed?
- What drugs target the protein encoded by this gene?

**Example use cases:** Getting a quick overview of a gene before designing experiments; Finding disease associations for a gene identified in a GWAS study; Identifying drug targets based on gene function and disease associations

**Input Data Accepted:** Gene names, symbols, HGNC IDs, Ensembl IDs, keywords

**Output Data Provided:** Comprehensive gene cards with function, expression, disease associations, variants, pathways, interactions, and cross-references

**Strengths:** Comprehensive integration of 150+ data sources; User-friendly interface with well-organized information; Covers all human genes including non-coding RNAs; Regularly updated with new data sources

**Limitations:** Human-specific; not useful for non-human organisms; Automatically imported data may include errors or outdated information; Some advanced features require a subscription (GeneCards Suite); Not appropriate for primary data analysis; use source databases for specific data types

**Common beginner mistakes:** Using GeneCards as a primary data source without checking the original source databases; Not realizing that some GeneCards features require a paid subscription; Assuming all information is manually curated (most is automatically imported)

**When to Use It:** Use GeneCards for a quick, comprehensive overview of a human gene, for finding disease associations, and for identifying relevant databases to consult for specific data types.



**When NOT to Use It:** For non-human organisms, use NCBI Gene or Ensembl. For primary data analysis, use the source databases directly. For clinical variant interpretation, use ClinVar.

**Related databases / alternatives:**

- NCBI Gene (<https://www.ncbi.nlm.nih.gov/gene>): More comprehensive for non-human organisms
- OMIM (<https://www.omim.org>): More detailed for disease genetics
- UniProt (<https://www.uniprot.org>): More detailed for protein function

**How It Connects to Other Resources:** GeneCards aggregates data from 150+ databases and provides links to all source databases. It is a hub for navigating the gene information landscape.

**API / FTP / programmatic access:** GeneCards API available for institutional subscribers (GeneCards Suite). Basic web access is free.

**Evidence/curation level:** Mixed — automatically imported from multiple sources; some manual curation

**Data Update Status:** Regular updates as source databases are updated

**Licensing / access restrictions:** Basic access free; advanced features require GeneCards Suite subscription

**Citation / Recommended Reference:** Stelzer G et al. (2016) The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, 54:1.30.1–1.30.33. doi:10.1002/cpbi.5

**Beginner-Friendly Explanation:** GeneCards is like a Wikipedia page for every human gene, but with information automatically collected from over 150 scientific databases. If you want to quickly learn everything about a gene — what it does, where it is expressed, what diseases it is linked to, what drugs target it — GeneCards is the place to start. It is especially useful for getting an overview before diving into more specialized databases.

**Advanced Technical Explanation:** GeneCards implements an automated data integration pipeline that imports and harmonizes data from over 150 source databases, using gene symbol and identifier mapping to link records across sources. The GeneCards Suite provides API access and advanced filtering capabilities for programmatic use. GeneCards uses a scoring system (GIFtS — GeneCards Inferred Functionality Score) to rank genes by the amount of available information, which can be used to prioritize genes for further investigation.

**One practical workflow example:**

Step 1: Navigate to <https://www.genecards.org> and search for your gene of interest.

Step 2: Review the gene summary, including function, aliases, and genomic location.

Step 3: Scroll to the "Diseases" section to find disease associations with evidence scores.

Step 4: Check the "Expression" section for tissue-specific expression data from GTEx and other sources.

Step 5: Follow links to source databases (UniProt, OMIM, ClinVar) for more detailed information on specific aspects.

## F3 – HGNC (HUGO Gene Nomenclature Committee)

---

**Official Website URL:** <https://www.genenames.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences (gene nomenclature)

**What It Is Used For:** HGNC is the authoritative source for human gene names and symbols, providing standardized nomenclature for all human genes. It is used for resolving gene name ambiguities, finding the official symbol for a gene, converting between different identifier systems, and ensuring consistent gene naming in publications and databases. HGNC assigns unique HGNC IDs to all human genes and maintains a curated database of approved gene symbols, names, and aliases.

**What Data It Contains:** HGNC contains records for all human genes (approximately 43,000 as of 2024), including approved gene symbols, full names, aliases, previous symbols, chromosomal location, gene family assignments, and cross-references to NCBI Gene, Ensembl, UniProt, OMIM, and other databases. HGNC also maintains gene family and group classifications.

**Main question it helps answer:** What is the official, approved symbol and name for this human gene?

**Typical user:** Researcher / Bioinformatician / Clinician / Database developer

**Example scientific questions:**

- What is the official HGNC symbol for this gene?
- What are the aliases and previous symbols for this gene?
- What HGNC ID corresponds to this Ensembl or NCBI Gene ID?

**Example use cases:**

- Resolving gene name ambiguities in a literature review
- Converting between HGNC IDs, NCBI Gene IDs, and Ensembl IDs
- Ensuring consistent gene naming in a manuscript or database

**Input Data Accepted:** Gene symbols, names, aliases, HGNC IDs, NCBI Gene IDs, Ensembl IDs

**Output Data Provided:** Approved gene symbols, names, aliases, HGNC IDs, cross-references to other databases

**Strengths:** Authoritative source for human gene nomenclature; Resolves ambiguities between aliases and official symbols; Provides cross-references to all major gene databases; Multi-Symbol Checker tool for batch ID conversion; REST API for programmatic access

**Limitations:**

- Human-specific; not applicable to non-human organisms
- Gene nomenclature can change over time; older symbols may be deprecated
- Not a source of functional gene information (use NCBI Gene or UniProt for that)

**Common beginner mistakes:**

- Using gene aliases or previous symbols instead of the current approved symbol
- Not checking HGNC when gene symbols appear inconsistent across databases



- Confusing HGNC IDs with NCBI Gene IDs or Ensembl IDs

**When to Use It:** Use HGNC when you need the official human gene symbol, when you need to resolve gene name ambiguities, or when you need to convert between gene identifier systems.

**When NOT to Use It:** HGNC is not appropriate for non-human organisms or for functional gene information. For gene function, use NCBI Gene or UniProt.

**Related databases / alternatives:** NCBI Gene (<https://www.ncbi.nlm.nih.gov/gene>): Broader scope including non-human organisms; Ensembl (<https://www.ensembl.org>): Alternative gene annotation system

**How It Connects to Other Resources:** HGNC provides cross-references to NCBI Gene, Ensembl, UniProt, OMIM, and many other databases. HGNC IDs are used as stable identifiers in many bioinformatics tools and databases.

**API / FTP / programmatic access:** HGNC REST API (<https://rest.genenames.org>) provides programmatic access to gene records. Bulk downloads available at <https://www.genenames.org/download/> Multi-Symbol Checker for batch processing.

**Evidence/curation level:** Manually curated by HGNC nomenclature committee

**Data Update Status:** Continuous updates as new genes are approved or nomenclature is revised

**Licensing / access restrictions:** Open access; data available under Creative Commons Attribution 4.0 license

**Citation / Recommended Reference:** Tweedie S et al. (2021) Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Research, 49(D1):D939–D946. doi:10.1093/nar/gkaa980

**Beginner-Friendly Explanation:** HGNC is the official naming authority for human genes. Every human gene has an official symbol (like BRCA1 or TP53) that is approved by HGNC, and this is the name that should be used in scientific publications. HGNC also keeps track of old names and aliases, so if you encounter an unfamiliar gene name, you can check HGNC to find the current official symbol.

**Advanced Technical Explanation:** HGNC implements a controlled nomenclature system with rules for gene symbol construction (typically 1-6 uppercase letters, sometimes followed by numbers). The HGNC REST API supports search by symbol, name, alias, or cross-reference ID, returning records in JSON format. The Multi-Symbol Checker accepts lists of gene symbols (including aliases and previous symbols) and returns the current approved symbols, enabling batch normalization of gene lists. HGNC IDs are stable identifiers that persist even when gene symbols change.

### One practical workflow example:

Step 1: Navigate to <https://www.genenames.org> and search for your gene by symbol or name.

Step 2: Verify the approved symbol and check for aliases or previous symbols.

Step 3: Note the HGNC ID for use as a stable cross-reference.

Step 4: Use the cross-reference links to find the corresponding NCBI Gene ID and Ensembl ID.

Step 5: For batch processing, use the Multi-Symbol Checker to convert a list of gene symbols to approved HGNC symbols.

## F4 – OMIM (Online Mendelian Inheritance in Man) — Cross-reference Entry

---

**Official Website URL:** <https://www.omim.org>

**Entry type:** Cross-reference entry — full database card provided in Category M, M1.

OMIM is mentioned in Category F because it links human genes to Mendelian disease phenotypes and is useful when a gene-level question becomes a disease-gene question. For full coverage of OMIM, including clinical use, disease-gene relationships, curation, access restrictions, limitations, and workflow examples, see Category M: Disease and Clinical Genomics Databases, M1 – OMIM.

## F5 – Ensembl Genes

---

**Official Website URL:** <https://www.ensembl.org>

**Resource Type:** Database / Genome Browser

**Main Biological Domain:** DNA sequences / RNA/transcriptomics

**What It Is Used For:** Ensembl Genes provides comprehensive genome-wide gene annotations for vertebrates and selected other eukaryotes, including protein-coding genes, non-coding RNAs, and pseudogenes. It is used for retrieving gene and transcript annotations, accessing Ensembl gene IDs (ENSG identifiers), downloading annotation files (GTF/GFF3) for bioinformatics pipelines, and performing comparative genomics analyses. Ensembl Genes is the primary source of gene annotations for RNA-seq analysis workflows.

**What Data It Contains:** Ensembl Genes contains gene models for over 300 species, including protein-coding genes, non-coding RNA genes (lncRNA, miRNA, snoRNA, etc.), pseudogenes, and readthrough transcripts. For human and mouse, Ensembl uses GENCODE annotations, which are the most comprehensive and manually curated gene annotations available. Each gene entry includes transcript isoforms, exon structures, protein sequences, GO annotations, and cross-references to other databases.

**Main question it helps answer:** What genes and transcripts are annotated in this genome, and what are their coordinates and sequences?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- What transcripts are annotated for this gene, and what are their exon coordinates?
- What is the Ensembl gene ID for this gene?
- What non-coding RNA genes are annotated in this genomic region?

**Example use cases:** Downloading GTF annotation files for RNA-seq read counting; Retrieving Ensembl gene IDs for use in differential expression analysis; Accessing GENCODE annotations for human genome analysis

**Input Data Accepted:** Gene names, Ensembl IDs, genomic coordinates, species names

**Output Data Provided:** Gene and transcript annotations, Ensembl IDs, GTF/GFF3 files, protein sequences, GO annotations

**Strengths:** Comprehensive gene annotations for 300+ species; GENCODE annotations for human and mouse; Regular releases with version tracking; BioMart for bulk data retrieval; REST API for programmatic access

**Limitations:** Ensembl gene IDs change between releases; version tracking is essential – Annotation quality varies by species – Ensembl and RefSeq gene models differ; mixing them causes errors

**Common beginner mistakes:** Not recording the Ensembl release version when using Ensembl IDs; Mixing Ensembl and RefSeq annotations in the same analysis; Not checking that the GTF file matches the genome assembly used for alignment

**When to Use It:** Use Ensembl Genes when you need genome-wide gene annotations for vertebrates, when you are performing RNA-seq analysis and need transcript coordinates, or when you need GENCODE annotations for human or mouse.

**When NOT to Use It:** For prokaryotic genomes, use NCBI RefSeq. For protein functional annotations, use UniProt.

**Related databases / alternatives:** NCBI Gene (<https://www.ncbi.nlm.nih.gov/gene>): Alternative gene annotation system; RefSeq (<https://www.ncbi.nlm.nih.gov/refseq>): NCBI's curated reference sequences

**How It Connects to Other Resources:** Ensembl links to UniProt, NCBI Gene, PDB, GO, Reactome, OMIM, and many other resources. BioMart enables cross-database queries.

**API / FTP / programmatic access:** Ensembl REST API (<https://rest.ensembl.org>); BioMart; FTP at <ftp://ftp.ensembl.org>; biomaRt R package.

**Evidence/curation level:** Mixed — computationally predicted with manual curation for human/mouse.

**Data Update Status:** Regular numbered releases (approximately quarterly)

**Licensing / access restrictions:** Open access; Apache 2.0 license

**Citation / Recommended Reference:** Martin FJ et al. (2023) Ensembl 2023. Nucleic Acids Research, 51(D1):D933–D941. doi:10.1093/nar/gkac958

**Beginner-Friendly Explanation:** Ensembl Genes is the part of Ensembl that tells you where all the genes are in a genome and what they look like. For each gene, it shows you all the different versions (transcripts) that can be made from it, including the exact positions of all the exons. This information is essential for RNA-seq analysis, where you need to know the gene coordinates to count how many reads map to each gene.

**Advanced Technical Explanation:** Ensembl uses an automated gene annotation pipeline that integrates protein alignments, cDNA/EST alignments, & ab initio gene predictions to generate gene models. For human and mouse, GENCODE annotations are used, which include extensive manual curation by the HAVANA team. Ensembl gene IDs (ENSG) are stable within a release but may change between releases due to gene merges, splits, or retirements. The Ensembl REST API provides access to gene annotations in JSON, GFF3, & other formats.

#### One practical workflow example:

- Step 1: Navigate to <https://www.ensembl.org> and search for your gene.
- Step 2: Note the Ensembl gene ID (ENSG) and the current Ensembl release version.
- Step 3: Use BioMart to download a table of all transcripts with their coordinates and biotypes.
- Step 4: Download the GTF annotation file from the Ensembl FTP site for the appropriate genome assembly.
- Step 5: Use the GTF file with featureCounts or HTSeq for RNA-seq read counting.

## Beginner Example for category F

---

A student has identified a gene (CFTR) from a literature search on cystic fibrosis and wants to learn more about it. They start with GeneCards (<https://www.genecards.org>), which provides a comprehensive overview of CFTR including its function, expression pattern, disease associations, and links to other databases. From GeneCards, they follow the link to OMIM to read the detailed clinical and molecular description of cystic fibrosis (OMIM #219700) and the CFTR gene (OMIM \*602421).

To find the official gene symbol and cross-references, the student checks HGNC (<https://www.genenames.org>) and confirms that "CFTR" is the approved symbol, with HGNC ID 1884. They also find the NCBI Gene ID (1080) and Ensembl ID (ENSG00000001626) for use in bioinformatics tools. This workflow — GeneCards for overview, OMIM for disease details, HGNC for nomenclature — is a standard approach for initial gene characterization.

## Advanced Research Example for category F

---

A bioinformatician is performing a genome-wide analysis of genes associated with a specific disease pathway. They use the NCBI Gene E-utilities API to programmatically retrieve all human genes annotated with a specific GO term, downloading the gene2go file from the NCBI FTP site and filtering for the relevant GO term. They then use the HGNC REST API to convert NCBI Gene IDs to approved HGNC symbols and Ensembl IDs, enabling integration with RNA-seq data analyzed using Ensembl annotations. For each candidate gene, they query OMIM to check for known disease associations and retrieve MIM numbers for use in clinical reporting. They also use the Ensembl BioMart to retrieve transcript coordinates and protein sequences for all candidate genes, enabling downstream structural and functional analysis. The entire workflow is scripted in Python, with all database versions and API query parameters recorded for reproducibility.

## Common Confusion Points

---

NCBI Gene IDs, Ensembl IDs, HGNC IDs, and UniProt accession numbers are all different identifier systems for the same gene. They are not interchangeable. Converting between them requires explicit ID mapping using tools like BioMart, the HGNC Multi-Symbol Checker, or the UniProt ID mapping service. OMIM covers only Mendelian (single-gene) inherited disorders. It is not appropriate for complex diseases (diabetes, hypertension, schizophrenia) that are influenced by many genes and environmental factors. For complex disease genetics, use the GWAS Catalog or DisGeNET. GeneCards aggregates data from many sources automatically, which means it may contain outdated or inconsistent information. Always verify important findings by checking the original source databases (UniProt, OMIM, ClinVar) rather than relying solely on GeneCards. Ensembl gene IDs (ENSG) are release-specific. An ENSG ID that was valid in Ensembl release 95 may have been retired or changed in a later release. Always record the Ensembl

release version when using Ensembl IDs in published analyses. HGNC is the authoritative source for human gene nomenclature, but gene symbols can change over time. A gene may have been known by a different symbol in older literature. When searching for a gene, check HGNC for current and previous symbols to ensure you are finding all relevant literature.

### Which One Should I Use?

---

For a quick, comprehensive overview of a human gene, start with GeneCards — it aggregates information from 150+ sources into a single page. For the official human gene symbol and ID conversion, use HGNC. For detailed information about Mendelian disease genetics, use OMIM. For gene information across all organisms and for NCBI Gene IDs, use NCBI Gene. For genome-wide gene annotations for RNA-seq analysis or comparative genomics, use Ensembl Genes. In practice, most researchers use a combination of these resources: GeneCards for initial exploration, HGNC for nomenclature, OMIM for disease genetics, NCBI Gene for cross-references, and Ensembl for genomic coordinates.

## Category G: Transcriptomics and Gene Expression Databases

### Category Overview

Transcriptomics databases store, curate, and provide access to gene expression data from a wide range of experimental platforms and biological contexts. The transcriptomics data landscape has evolved dramatically over the past two decades, from early microarray experiments to RNA-seq, single-cell RNA-seq, and spatial transcriptomics. The databases that store this data have evolved correspondingly, from simple repositories of microarray files to sophisticated platforms that provide uniform reprocessing, interactive visualization, and integration with other omics data types.

The transcriptomics database ecosystem can be divided into three tiers. The first tier consists of primary data repositories — GEO (Gene Expression Omnibus) and ArrayExpress/BioStudies — which store raw and processed expression data as submitted by researchers, with minimal post-submission curation. The second tier consists of curated expression resources — Expression Atlas and GTEx — which provide uniformly processed, quality-controlled expression data with rich metadata and interactive visualization tools. The third tier consists of raw sequencing read archives — SRA (Sequence Read Archive) and ENA — which store the raw FASTQ files from RNA-seq experiments, enabling independent reanalysis with different pipelines.

Understanding which tier of the transcriptomics database ecosystem is appropriate for a given task is essential for efficient data access. For finding published expression datasets associated with a specific biological condition, GEO or ArrayExpress is the starting point. For accessing uniformly processed expression data across many experiments, Expression Atlas is more appropriate. For tissue-specific expression in humans, GTEx is the definitive resource. For downloading raw sequencing reads for reanalysis, SRA or ENA is required. Many analyses require data from multiple tiers: for example, finding a relevant dataset in GEO, downloading the raw reads from SRA, reprocessing with a standardized pipeline, and comparing results to Expression Atlas data.

## G1 – GEO (Gene Expression Omnibus)

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/geo>

**Resource Type:** Repository / Database

**Main Biological Domain:** RNA/transcriptomics / Epigenomics / Omics

**What It Is Used For:** GEO is the primary public repository for high-throughput functional genomics data, maintained by NCBI. It stores data from microarray, RNA-seq, ChIP-seq, ATAC-seq, methylation, and other high-throughput experiments. GEO is used for depositing expression data associated with publications, for finding publicly available datasets for reanalysis, and for accessing processed expression matrices and metadata. GEO is the most widely used expression data repository in the world.

**What Data It Contains:** GEO contains over 5,000 organisms and more than 4 million samples (as of 2024), organized into GEO Series (GSE, a complete experiment), GEO Samples (GSM, individual samples), GEO Platforms (GPL, the array or sequencing platform), and GEO DataSets (GDS, curated datasets). Data types include microarray expression data, RNA-seq count matrices, ChIP-seq peak files, methylation arrays, and many other high-throughput data types. Raw sequencing reads associated with GEO submissions are stored in SRA.

**Main question it helps answer:** What publicly available expression datasets exist for this biological condition, tissue, or disease?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What RNA-seq datasets are available for breast cancer vs. normal tissue?
- What microarray datasets have been published for this specific cell type?
- What is the expression profile of this gene across different tissues?

**Example use cases:**

- Finding publicly available datasets for meta-analysis or reanalysis
- Depositing expression data associated with a publication
- Accessing processed expression matrices for downstream analysis

**Input Data Accepted:** Keywords, organism names, experiment types, GEO accession numbers (GSE, GSM, GPL, GDS)

**Output Data Provided:** Expression data matrices, metadata, processed data files, links to raw reads in SRA

**Strengths:**

- Largest public repository for functional genomics data
- Covers microarray, RNA-seq, ChIP-seq, and many other data types
- GEO2R tool for basic differential expression analysis within GEO
- Links to raw reads in SRA for reanalysis
- Free and open access

**Limitations:**

- Data quality and processing vary widely between submissions



- Metadata quality is inconsistent; finding relevant datasets requires careful searching
- No uniform reprocessing; each dataset uses the submitter's pipeline
- GEO2R is limited in functionality compared to dedicated DE analysis tools
- Large datasets can be slow to download

#### Common beginner mistakes:

- Downloading processed data without checking the processing pipeline used
- Not downloading the raw reads from SRA when reanalysis with a standardized pipeline is needed
- Not recording the GEO accession number for reproducibility
- Confusing GEO Series (GSE) with GEO DataSets (GDS) — GDS are curated subsets of GEO data

**When to Use It:** Use GEO when you need to find publicly available expression datasets for a specific biological condition, when you need to deposit expression data associated with a publication, or when you need to access processed expression matrices for downstream analysis.

**When NOT to Use It:** For uniformly processed expression data, use Expression Atlas. For tissue-specific expression in humans, use GTEx. For raw sequencing reads, use SRA or ENA.

#### Related databases / alternatives:

- ArrayExpress/BioStudies (<https://www.ebi.ac.uk/arrayexpress>): European equivalent
- Expression Atlas (<https://www.ebi.ac.uk/gxa>): Uniformly processed expression data
- SRA (<https://www.ncbi.nlm.nih.gov/sra>): Raw sequencing reads

**How It Connects to Other Resources:** GEO links to SRA for raw sequencing reads, to PubMed for associated publications, and to NCBI Gene for gene-level expression profiles. GEO accession numbers are widely used in publications and are cross-referenced in many databases.

**API / FTP / programmatic access:** GEO E-utilities API (esearch and efetch with db=gds) for programmatic access. FTP at <ftp://ftp.ncbi.nlm.nih.gov/geo/> for bulk downloads. The GEOquery R package provides convenient access to GEO data in R. The GEO2R web tool provides basic differential expression analysis.

**Evidence/curation level:** Community-submitted; minimal post-submission curation (format validation); GEO DataSets are curated subsets

**Data Update Status:** Continuous updates as new datasets are submitted

**Licensing / access restrictions:** Open access; individual datasets subject to submitter's terms

**Citation / Recommended Reference:** Barrett T et al. (2013) NCBI GEO: archive for functional genomics data sets — update. Nucleic Acids Research, 41(D1):D991–D995. doi:10.1093/nar/gks1193

**Beginner-Friendly Explanation:** GEO is the world's largest collection of gene expression data. When scientists do experiments to measure which genes are active in different conditions — like comparing cancer cells to normal cells — they deposit their data in GEO so other researchers can use it. You can search GEO to find datasets relevant to your research question and download the data for your own analysis. It is like a library of gene expression experiments.

**Advanced Technical Explanation:** GEO implements a hierarchical data model with four entity types: Platform (GPL, describing the measurement technology), Sample (GSM, individual biological samples), Series (GSE,

complete experiments), and DataSet (GDS, curated subsets). The GEO SOFT format (Simple Omnibus Format in Text) is the standard submission format, containing metadata and data matrices. The GEOquery R package implements functions for downloading and parsing GEO SOFT files, returning ExpressionSet or SummarizedExperiment objects for downstream analysis. GEO accession numbers follow the format GSE + integer (Series), GSM + integer (Sample), GPL + integer (Platform), GDS + integer (DataSet).

### One practical workflow example:

- Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/geo> and search for your condition (e.g., "breast cancer RNA-seq Homo sapiens").
- Step 2: Filter results by organism, experiment type, and date.
- Step 3: Examine the Series (GSE) record for experimental design, sample metadata, and data availability.
- Step 4: Download the processed data matrix (Series Matrix file) for quick analysis, or follow the SRA link to download raw reads for reanalysis.
- Step 5: In R, use GEOquery: `gse <- getGEO("GSE12345"); exprs(gse[[1]])` to access the expression matrix.
- Step 6: Record the GSE accession number and data processing details for your methods section.

## G2 – ArrayExpress / BioStudies [Note: ArrayExpress has been integrated into BioStudies at EMBL-EBI]

**Official Website URL:** <https://www.ebi.ac.uk/arrayexpress> (redirects to BioStudies: <https://www.ebi.ac.uk/biostudies/arrayexpress>)

**Resource Type:** Repository / Database

**Main Biological Domain:** RNA/transcriptomics / Epigenomics / Omics

**What It Is Used For:** ArrayExpress was the European equivalent of GEO, storing functional genomics data from microarray and sequencing experiments. It has been integrated into BioStudies, EMBL-EBI's broader data repository, while maintaining the ArrayExpress data collection and accession numbers. ArrayExpress/BioStudies is used for depositing expression data to comply with European journal and funder requirements, and for accessing publicly available expression datasets from European research groups.

**What Data It Contains:** ArrayExpress contains data from microarray, RNA-seq, ChIP-seq, and other high-throughput experiments, organized by experiment accession (E-MTAB-, E-GEOD-, etc.). Many GEO datasets are mirrored in ArrayExpress (with E-GEOD- accessions). The collection contains data from thousands of experiments across many organisms and biological conditions.

**Main question it helps answer:** What publicly available expression datasets are available in the European data archive for this biological condition?

**Typical user:** Researcher / Bioinformatician (particularly in Europe)

**Example scientific questions:**

- What expression datasets are available for this disease in the European archive?
- How do I submit my expression data to comply with European journal requirements?
- What microarray datasets are available for this organism?

**Example use cases:** Depositing expression data for European-funded research; Finding expression datasets not in GEO; Accessing data from European research groups

**Input Data Accepted:** Keywords, organism names, experiment types, ArrayExpress accession numbers

**Output Data Provided:** Expression data matrices, metadata, processed data files, links to raw reads in ENA

**Strengths:** European data archive with strong compliance with MIAME/MINSEQE standards; Many GEO datasets mirrored for European access; Integration with ENA for raw sequencing reads; Part of the EMBL-EBI ecosystem

**Limitations:** Integration into BioStudies has caused some confusion about current URLs and interfaces; Less widely used than GEO in North America; Some older ArrayExpress features have changed with the BioStudies integration

**Common beginner mistakes:** Using the old ArrayExpress URL without realizing it now redirects to BioStudies; Not realizing that many GEO datasets are also available in ArrayExpress with different accession numbers

**When to Use It:** Use ArrayExpress/BioStudies when you are based in Europe and need to deposit data, when you need to find datasets not in GEO, or when you prefer the EBI interface.

**When NOT to Use It:** For the most comprehensive expression dataset search, GEO has broader coverage. For uniformly processed data, use Expression Atlas.

**Related databases / alternatives:** GEO (<https://www.ncbi.nlm.nih.gov/geo>): US equivalent, broader coverage; Expression Atlas (<https://www.ebi.ac.uk/gxa>): Uniformly processed expression data; ENA (<https://www.ebi.ac.uk/ena>): Raw sequencing reads

**How It Connects to Other Resources:** ArrayExpress/BioStudies links to ENA for raw sequencing reads, to Expression Atlas for uniformly processed data, and to Europe PMC for associated publications.

**API / FTP / programmatic access:** BioStudies API (<https://www.ebi.ac.uk/biostudies/api>) for programmatic access. FTP at <ftp://ftp.ebi.ac.uk/pub/databases/arrayexpress/> for bulk downloads.

**Evidence/curation level:** Community-submitted; MIAME/MINSEQE compliance checking

**Data Update Status:** Continuous updates as new datasets are submitted

**Licensing / access restrictions:** Open access for most data

**Citation / Recommended Reference:** Athar A et al. (2019) ArrayExpress update — from bulk to single-cell expression data. Nucleic Acids Research, 47(D1):D711–D715. doi:10.1093/nar/gky964

**Beginner-Friendly Explanation:** ArrayExpress is Europe's main database for gene expression data, similar to GEO in the US. It has been recently integrated into a broader database called BioStudies, but the ArrayExpress data collection is still accessible. If you are in Europe and need to submit your expression data, or if you are looking for datasets from European research groups, ArrayExpress/BioStudies is the place to go.

**Advanced Technical Explanation:** ArrayExpress implements the MIAME (Minimum Information About a Microarray Experiment) and MINSEQE (Minimum Information about a high-throughput SEQuencing Experiment) standards for metadata. The BioStudies integration provides a unified submission interface for all study types. ArrayExpress accession numbers follow the format E-[source]-[number], where source indicates the origin (MTAB for direct submissions, GEOD for GEO-mirrored data, TABM for tab2mage submissions).

#### **One practical workflow example:**

- Step 1: Navigate to <https://www.ebi.ac.uk/biostudies/arrayexpress> and search for your condition.
- Step 2: Filter by organism, experiment type, and technology.
- Step 3: Examine the experiment record for design, samples, and data files.
- Step 4: Download processed data files or follow ENA links for raw reads.
- Step 5: Record the ArrayExpress accession number for your methods section.

## G3 – Expression Atlas

**Official Website URL:** <https://www.ebi.ac.uk/gxa>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** RNA/transcriptomics

**What It Is Used For:** Expression Atlas is a curated resource at EMBL-EBI that provides uniformly processed and quality-controlled gene expression data across a wide range of species, tissues, cell types, and experimental conditions. Unlike GEO and ArrayExpress, which store data as submitted, Expression Atlas reprocesses raw data using standardized pipelines, enabling meaningful comparisons across experiments. It is used for exploring gene expression patterns, comparing expression across tissues and conditions, and accessing pre-computed differential expression results.

**What Data It Contains:** Expression Atlas contains uniformly processed RNA-seq and microarray data from thousands of experiments across many organisms, including baseline expression data (expression levels in different tissues and cell types) and differential expression data (expression changes between conditions). Data is organized by experiment and by gene, enabling both experiment-centric and gene-centric queries. Expression Atlas also includes single-cell RNA-seq data through the Single Cell Expression Atlas.

**Main question it helps answer:** In what tissues, cell types, and conditions is this gene expressed, based on uniformly processed data?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:**

- In what tissues is this gene most highly expressed?
- Is this gene differentially expressed in this disease condition compared to normal?
- What genes are differentially expressed in this experiment?

**Example use cases:** Checking tissue-specific expression of a gene before designing experiments; Finding pre-computed differential expression results for a published dataset; Comparing expression patterns across species

**Input Data Accepted:** Gene names, Ensembl IDs, experiment accession numbers, organism names, tissue names

**Output Data Provided:** Baseline expression heatmaps, differential expression results, expression profiles across tissues and conditions

**Strengths:** Uniformly processed data enables meaningful cross-experiment comparisons; Pre-computed differential expression results save analysis time; Interactive visualization tools; Covers both bulk RNA-seq and single-cell RNA-seq (Single Cell Expression Atlas); REST API for programmatic access

**Limitations:** Coverage is limited to experiments that have been reprocessed by the Expression Atlas team; Not all GEO/ArrayExpress datasets are available in Expression Atlas; Uniform processing may not be optimal for all data types

**Common beginner mistakes:** Assuming Expression Atlas contains all GEO/ArrayExpress datasets — it covers only a curated subset; Not distinguishing between baseline expression (absolute levels) and differential expression (relative changes)

**When to Use It:** Use Expression Atlas when you want uniformly processed expression data for cross-experiment comparisons, when you want pre-computed differential expression results, or when you want to explore tissue-specific expression patterns.

**When NOT Use It:** For the most comprehensive dataset search, use GEO or ArrayExpress. For raw data reanalysis, use SRA or ENA.

**Related databases / alternatives:** GEO (<https://www.ncbi.nlm.nih.gov/geo/>): Broader coverage, less uniform processing; GTEx (<https://gtexportal.org/>): Human tissue-specific expression; ArrayExpress (<https://www.ebi.ac.uk/arrayexpress/>): Source data for Expression Atlas

**How It Connects to Other Resources:** Expression Atlas links to ArrayExpress/BioStudies for source data, to Ensembl for gene annotations, and to Reactome for pathway context. The Single Cell Expression Atlas links to the Human Cell Atlas.

**API / FTP / programmatic access:** Expression Atlas REST API (<https://www.ebi.ac.uk/gxa/api>) provides programmatic access to expression data. FTP at <ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/atlas/> for bulk downloads.

**Evidence/curation level:** Curated — data is selected and uniformly reprocessed by the Expression Atlas team

**Data Update Status:** Regular updates as new experiments are processed

**Licensing / access restrictions:** Open access; data available under Creative Commons licenses

**Citation / Recommended Reference:** Moreno P et al. (2022) Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Research*, 50(D1):D129–D140. doi:10.1093/nar/gkab1030

**Beginner-Friendly Explanation:** Expression Atlas is a database that shows you where and when genes are active (expressed) in different organisms, tissues, and conditions. Unlike other expression databases that just store data as scientists submitted, Expression Atlas reprocesses all the data using the same methods, so you can fairly compare expression levels across different experiments.

**Advanced Technical Explanation:** Expression Atlas uses standardized analysis pipelines for RNA-seq (iRAP pipeline using STAR for alignment and DESeq2 for differential expression) and microarray data (limma for differential expression). Baseline expression data is presented as TPM (transcripts per million) values, enabling cross-sample comparisons. The REST API supports queries by gene, experiment, or condition, returning data in JSON or TSV format. The Single Cell Expression Atlas uses the SCE (SingleCellExperiment) data model and provides UMAP visualizations of cell type clusters.

#### One practical workflow example:

- Step 1: Navigate to <https://www.ebi.ac.uk/gxa> and search for your gene of interest.
- Step 2: View the baseline expression heatmap showing expression across tissues and organisms.
- Step 3: Click on a specific experiment to view differential expression results.
- Step 4: Use the REST API to download expression data programmatically: GET [https://www.ebi.ac.uk/gxa/genes/ENSG00000012048/baseline\\_experiments](https://www.ebi.ac.uk/gxa/genes/ENSG00000012048/baseline_experiments)
- Step 5: Compare expression patterns across tissues to identify the most relevant experimental system for your research.



## G4 – GTEx (Genotype-Tissue Expression)

**Official Website URL:** <https://gtexportal.org>

**Resource Type:** Database / Dataset Collection

**Main Biological Domain:** RNA/transcriptomics / Variants

**What It Is Used For:** GTEx is a comprehensive resource for studying tissue-specific gene expression and regulation in humans, providing RNA-seq data from 54 non-diseased tissue sites across nearly 1,000 human donors. It is used for understanding tissue-specific gene expression patterns, identifying expression quantitative trait loci (eQTLs), and studying the relationship between genetic variation and gene expression. GTEx is the definitive resource for human tissue-specific expression and eQTL data.

**What Data It Contains:** GTEx contains RNA-seq data from 54 human tissues (approximately 17,000 samples), whole genome sequencing data from donors, eQTL data (variants associated with gene expression levels), sQTL data (variants associated with splicing), and tissue-specific expression profiles. GTEx data is available at multiple levels: raw reads (in dbGaP for controlled access), processed expression matrices, and pre-computed eQTL results.

**Main question it helps answer:** In what human tissues is this gene expressed, and what genetic variants affect its expression?

**Typical user:** Researcher / Bioinformatician / Clinician

**Example scientific questions:**

- In what human tissues is this gene most highly expressed?
- What eQTLs are associated with this gene in brain tissue?
- Is this GWAS variant an eQTL for a nearby gene?

**Example use cases:**

- Checking tissue-specific expression of a candidate gene from a GWAS study
- Identifying eQTLs to interpret GWAS variants
- Comparing expression levels of a gene across all human tissues

**Input Data Accepted:** Gene names, Ensembl IDs, variant IDs (rsIDs), tissue names

**Output Data Provided:** Tissue-specific expression profiles, eQTL data, violin plots of expression distributions, raw data downloads

**Strengths:** Definitive resource for human tissue-specific expression; Comprehensive eQTL data for all 54 tissues; High-quality, uniformly processed data; Interactive visualization tools; Pre-computed eQTL results available for download

**Limitations:** Human-specific; not applicable to other organisms; Data is from non-diseased tissues; not representative of disease states; Raw data requires dbGaP access (controlled access due to human subjects); Limited to 54 tissue types; some tissues not represented

**Common beginner mistakes:** Assuming GTEx data represents disease states — it is from non-diseased donors; Not realizing that raw GTEx data requires dbGaP access; Confusing eQTLs (variants affecting expression) with causal disease variants



**When to Use It:** Use GTEx when you need human tissue-specific expression data, when you need eQTL data to interpret GWAS variants, or when you want to understand the regulatory context of a genetic variant.

**When NOT to Use It:** For disease-specific expression, use GEO or Expression Atlas. For non-human organisms, use Expression Atlas or GEO. For raw data reanalysis, note that raw GTEx data requires dbGaP access.

**Related databases / alternatives:** Expression Atlas (<https://www.ebi.ac.uk/gxa>): Broader species coverage, less human-specific depth; GEO (<https://www.ncbi.nlm.nih.gov/geo>): Broader coverage including disease conditions; ENCODE (<https://www.encodeproject.org>): Regulatory genomics data

**How It Connects to Other Resources:** GTEx links to Ensembl for gene annotations, to dbSNP for variant information, and to the GWAS Catalog for disease associations. GTEx eQTL data is used by many GWAS interpretation tools (e.g., SMR, Mendelian randomization tools).

**API / FTP / programmatic access:** GTEx Portal API (<https://gtexportal.org/api/v2>) provides programmatic access to expression and eQTL data. Bulk data downloads available at <https://gtexportal.org/home/downloads/adult-gtex>. Raw data available through dbGaP (controlled access).

**Evidence/curation level:** Curated — uniformly processed by the GTEx Analysis Working Group

**Data Update Status:** Major releases (GTEx v8, v9, etc.) with new donors and tissues; current version is GTEx Analysis V10 (2023)

**Licensing / access restrictions:** Processed data open access; raw data requires dbGaP access agreement

**Citation / Recommended Reference:** GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330. doi:10.1126/science.aaz1776

**Beginner-Friendly Explanation:** GTEx is a project that measured gene activity in 54 different human tissues from nearly 1,000 donors. It tells you exactly how active each gene is in each tissue — for example, whether a gene is mostly active in the brain, the liver, or everywhere equally. It also tells you whether certain genetic variants affect how much a gene is expressed. GTEx is the most comprehensive resource for understanding where and how much human genes are expressed.

**Advanced Technical Explanation:** GTEx uses a standardized RNA-seq processing pipeline (STAR alignment, RSEM quantification, TMM normalization) applied uniformly across all samples. eQTL analysis uses FastQTL with permutation-based FDR correction, identifying cis-eQTLs within 1 Mb of the gene TSS. The GTEx Portal API provides access to expression data (median TPM by tissue), eQTL data (effect size, p-value, FDR), and individual-level data (for registered users). GTEx data is widely used for colocalization analysis (e.g., with GWAS summary statistics) to identify genes mediating GWAS signals.

#### One practical workflow example:

- Step 1: Navigate to <https://gtexportal.org> and search for your gene of interest.
- Step 2: View the tissue expression violin plot to identify tissues with highest expression.
- Step 3: Click on a specific tissue to view the expression distribution and individual data points.
- Step 4: Navigate to the eQTL section to find variants associated with expression of this gene.
- Step 5: Download the pre-computed eQTL results for your gene from the GTEx downloads page for use in colocalization analysis.

## G5 – SRA (Sequence Read Archive) — Cross-reference Entry

---

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/sra>

**Entry type:** Cross-reference entry — full database card provided in Category H, H1.

SRA is mentioned in Category G because many transcriptomics studies deposited in GEO or related expression resources link to raw RNA-seq reads stored in SRA. However, SRA is primarily a raw sequencing data repository, not a processed gene expression database. For full coverage of SRA, including raw read retrieval, access methods, limitations, and reproducibility notes, see Category H: Raw Sequencing Data Repositories, H1 – SRA.

### Beginner Example for category G

---

A student wants to find published RNA-seq data comparing gene expression in Alzheimer's disease brain tissue versus normal brain tissue. They navigate to GEO (<https://www.ncbi.nlm.nih.gov/geo>) and search for "Alzheimer's disease RNA-seq Homo sapiens brain." They find several relevant Series (GSE) records and examine the metadata to find a dataset with a suitable experimental design (e.g., multiple cases and controls, appropriate brain region). They download the Series Matrix file, which contains the processed expression matrix, and load it into R using the GEOquery package for differential expression analysis.

To check whether their gene of interest is expressed in brain tissue in healthy individuals, the student also visits GTEx (<https://gtexportal.org>) and searches for the gene. The GTEx portal shows expression levels across 54 tissues, confirming that the gene is highly expressed in multiple brain regions. This combination of GEO (for disease-specific data) and GTEx (for normal tissue expression) is a standard approach for transcriptomics research.

### Advanced Research Example for category G

---

A bioinformatician is performing a meta-analysis of RNA-seq datasets for a specific disease, requiring uniform reprocessing of raw data from multiple studies. They first search GEO and ArrayExpress to identify all relevant datasets, recording GSE and E-MTAB accession numbers. For each dataset, they download the raw FASTQ files from SRA (using fasterq-dump) or ENA (using FTP), then process all datasets through a standardized pipeline (STAR alignment, featureCounts, DESeq2 normalization) to enable cross-study comparisons.

They also check Expression Atlas to see which of their datasets have already been uniformly processed, potentially saving reprocessing time. For the final analysis, they compare their uniformly processed results to the pre-computed differential expression results in Expression Atlas and to the tissue-specific expression profiles in GTEx, providing context for their findings. All accession numbers, software versions, and pipeline parameters are recorded in a reproducible workflow using Snakemake or Nextflow.

## Common Confusion Points

---

GEO and SRA are complementary, not redundant. GEO stores processed expression data (count matrices, normalized values) submitted by researchers; SRA stores the raw sequencing reads (FASTQ files) associated with the same experiments. For most reanalysis workflows, you need both: GEO for metadata and processed data, SRA for raw reads.

ArrayExpress has been integrated into BioStudies. The old ArrayExpress URL (<https://www.ebi.ac.uk/arrayexpress>) now redirects to BioStudies. ArrayExpress accession numbers (E-MTAB-, E-GEOD-) are still valid and the data is still accessible, but the interface has changed. Some older tutorials may reference the old interface.

Expression Atlas contains only a curated subset of GEO/ArrayExpress data. Not all GEO datasets are available in Expression Atlas — only those that have been selected and uniformly reprocessed by the Expression Atlas team. If you cannot find a dataset in Expression Atlas, check GEO or ArrayExpress directly.

GTEx data is from non-diseased tissues. GTEx provides expression data from healthy donors, which is valuable for understanding normal tissue-specific expression but is not representative of disease states. For disease-specific expression, use GEO or Expression Atlas.

Raw GTEx data requires dbGaP access. The processed GTEx data (expression matrices, eQTL results) is freely available, but the raw sequencing reads and individual-level genotype data require a dbGaP data access agreement due to human subjects protections. Many analyses can be performed with the processed data without needing raw access.

## Which One Should I Use?

---

For finding publicly available expression datasets for a specific biological condition, start with GEO — it has the broadest coverage. For European datasets or data deposition in Europe, use ArrayExpress/BioStudies. For uniformly processed expression data enabling cross-experiment comparisons, use Expression Atlas. For human tissue-specific expression and eQTL data, use GTEx. For downloading raw sequencing reads for independent reanalysis, use SRA (or ENA for faster downloads). In practice, most transcriptomics workflows use multiple resources: GEO for dataset discovery, SRA/ENA for raw data download, and Expression Atlas or GTEx for context and comparison.

## Category H: Raw Sequencing Data Repositories

### CATEGORY OVERVIEW

Raw sequencing data repositories are specialized archives designed to store, preserve, and provide public access to the primary output of high-throughput sequencing instruments — the unprocessed or minimally processed reads generated by technologies such as Illumina short-read sequencing, Pacific Biosciences long-read sequencing, Oxford Nanopore sequencing, and older platforms such as 454 pyrosequencing and SOLiD. These repositories serve as the foundational layer of the genomics data ecosystem, ensuring that the raw experimental evidence underlying published findings remains accessible for independent verification, reanalysis, and meta-analysis. Major funding agencies and journals now mandate deposition of raw sequencing data in recognized repositories as a condition of publication, making familiarity with these archives essential for any researcher working in genomics, transcriptomics, metagenomics, or epigenomics.

A researcher needs raw sequencing data repositories in several distinct scenarios. When reproducing or extending a published study, the raw reads allow one to apply updated alignment tools, reference genomes, or variant-calling pipelines that may yield improved results compared to the original analysis. When conducting meta-analyses or systematic reviews, pooling raw data from multiple studies enables harmonized processing under a single computational pipeline, eliminating batch effects introduced by inconsistent analytical choices. Researchers also deposit their own data to these repositories to comply with journal and funder mandates, to enable community reuse, and to establish priority for datasets that accompany publications. Training bioinformaticians frequently use publicly available raw datasets to benchmark new tools or to practice analysis workflows without generating new experimental data.

It is important to distinguish raw data repositories from processed data repositories. Raw data repositories (SRA, ENA, DRA) store sequencing reads in formats such as FASTQ, BAM, or CRAM — the direct output of sequencing instruments before any biological interpretation. Processed data repositories, such as GEO (Gene Expression Omnibus) or ArrayExpress, store the results of analyses performed on those reads: count matrices, normalized expression values, differential expression tables, or peak calls. Many studies deposit data in both: raw reads in SRA/ENA and processed results in GEO/ArrayExpress. When a researcher wants to re-analyze data from scratch using their own pipeline, they need the raw repository; when they want to use the authors' processed results directly, the processed repository is more appropriate. Understanding this distinction prevents the common mistake of downloading processed data and treating it as raw reads, or vice versa.

## H1 – SRA (Sequence Read Archive)

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/sra>

**Resource Type:** Repository

**Main Biological Domain:** RNA/transcriptomics / DNA sequences / Omics

**What It Is Used For:** The SRA is the primary NCBI repository for storing and distributing raw high-throughput sequencing data submitted alongside publications or as standalone datasets. Researchers use it to deposit sequencing data to comply with journal and funding mandates, and to retrieve publicly available raw reads for reanalysis, benchmarking, or meta-analysis. It is the largest single archive of sequencing data in the world by volume, covering virtually all sequencing technologies and biological domains.

**What Data It Contains:** The SRA contains raw sequencing reads from a vast range of experiment types including whole-genome sequencing, RNA-seq, ChIP-seq, ATAC-seq, 16S amplicon metagenomics, whole-metagenome shotgun sequencing, single-cell RNA-seq, and many others. Data are organized hierarchically: BioProject (study), BioSample (biological sample), SRA Experiment (library preparation), and SRA Run (sequencing run). Metadata includes organism, tissue, experimental conditions, sequencing platform, library strategy, and links to associated publications.

**Main question it helps answer:** What raw sequencing reads are publicly available for a given organism, experiment type, or study, and how can I access them for reanalysis?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What RNA-seq datasets are available for human liver tissue that I can use to validate my differential expression findings?
- Can I download the raw reads from a published GWAS study to re-call variants using a newer pipeline?
- Are there publicly available metagenomic datasets from gut microbiome studies in patients with inflammatory bowel disease?

**Example use cases:**

- Downloading raw FASTQ files from a published RNA-seq study to re-align against an updated reference genome and re-quantify gene expression.
- Retrieving ChIP-seq reads for a transcription factor of interest to perform a meta-analysis of binding sites across multiple cell lines.
- Accessing single-cell RNA-seq data from a developmental biology study to apply a new cell-type annotation algorithm.

**Input Data Accepted:** SRA accession numbers including SRR, SRX, SRS, SRP, and PRJNA formats, search queries based on organism name, experiment type, or keywords, BioProject and BioSample identifiers, and Entrez query syntax for programmatic access.

**Output Data Provided:** Outputs include raw sequencing reads in SRA format that can be converted and downloaded as FASTQ files using the SRA Toolkit, BAM or CRAM files when submitted in aligned formats,

experiment and run metadata in XML or JSON formats, run selector tables with metadata-based filtering options, and links to associated BioProject, BioSample, and publication records.

**Strengths:** The SRA is the largest global archive of raw sequencing data, containing petabytes of data generated from diverse sequencing platforms and biological applications. It is tightly integrated with the NCBI ecosystem, including PubMed, GEO, BioProject, BioSample, and dbGaP, allowing seamless cross-resource navigation. The SRA Toolkit provides command-line utilities such as fastq-dump, fasterq-dump, and prefetch for efficient data retrieval, while cloud-based access through AWS and Google Cloud helps overcome download limitations for large datasets. Additionally, comprehensive metadata standards and controlled vocabularies improve dataset discoverability.

**Limitations:** SRA data often requires conversion to FASTQ format using the SRA Toolkit, introducing an additional processing step compared with repositories offering direct FASTQ downloads. Metadata quality can vary substantially, with many submissions containing incomplete or inconsistent annotations. Controlled-access datasets, particularly human genomic data managed through dbGaP, require separate approval processes. Large datasets such as whole-genome sequencing cohorts may be difficult to download without cloud infrastructure, and the archive does not provide built-in quality control or validation of submitted sequencing data.

**Common Beginner Mistakes:** A common mistake is using fast-q-dump instead of faster-q-dump for large datasets, despite fasterq-dump being considerably faster and better suited for parallel processing. Another frequent error is forgetting to include the --split-files option when downloading paired-end sequencing data, which results in interleaved reads that may not be compatible with many downstream bioinformatics tools.

**Confusing SRA accession levels:** downloading at the SRP (study) level when individual SRR (run) accessions are needed; Not checking whether data is controlled access before attempting download, leading to failure or incomplete retrievals

**When to Use It:** Use SRA when you need the raw sequencing reads from a published study to perform your own analysis pipeline from scratch. It is the appropriate resource when you want to apply updated tools, reference genomes, or analytical methods to existing data, or when you need to pool data from multiple studies under a harmonized pipeline.

**When NOT Use It:** Do not use SRA if you only need processed results such as gene expression counts or normalized values — GEO or ArrayExpress are more appropriate for those. If you are looking for assembled genome sequences rather than raw reads, NCBI Assembly or GenBank/INSDC are better resources.

#### **Related databases / alternatives:**

- ENA (European Nucleotide Archive): INSDC partner; mirrors most SRA data; often provides direct FASTQ download without format conversion.
- DRA (DDBJ Sequence Read Archive): INSDC partner in Japan; mirrors SRA/ENA
- GEO (Gene Expression Omnibus): stores processed expression data; often links to SRA for raw reads
- ArrayExpress: EBI equivalent of GEO; links to ENA for raw reads
- dbGaP: NCBI repository for controlled-access human genetic data



**How It Connects to Other Resources:** SRA is deeply integrated with the broader NCBI infrastructure. Each SRA submission is linked to a BioProject record (study-level metadata) and one or more BioSample records (sample-level metadata), and many submissions are cross-referenced with GEO series records that contain the processed data. SRA records frequently link to PubMed publications, and human genetic data with privacy restrictions are housed in the related dbGaP repository. The INSDC partnership ensures that data submitted to SRA is mirrored at ENA and DRA within 24-48 hours.

#### API / FTP / programmatic access:

- SRA Toolkit: command-line suite; key tools are prefetch (download .sra files), fasterq-dump (convert to FASTQ), and sam-dump (convert to SAM)
- NCBI Entrez API (E-utilities): esearch and efetch can query and retrieve SRA metadata in XML format  
Endpoint: <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/>
- AWS Open Data: many SRA files available in s3://sra-pub-run-odp/ bucket without egress charges
- Google Cloud: SRA data available via gs://sra-pub-run-1/ bucket
- FTP: <ftp://ftp.ncbi.nlm.nih.gov/sra/>
- pysradb (Python): programmatic metadata retrieval and download management
- SRA Run Selector: web interface for batch accession retrieval with metadata filtering

**Evidence/curation level:** Community-submitted; minimal curation of data content; metadata reviewed for format compliance but biological accuracy not independently verified.

**Data Update Status:** Continuously updated; new submissions processed daily. The archive has grown exponentially and as of 2024 contains over 50 petabytes of sequence data.

**Licensing / access restrictions:** Most data is openly accessible. Controlled-access data (primarily human genetic data with privacy implications) is housed in dbGaP and requires application to a Data Access Committee. Open-access data has no restrictions on use or redistribution.

**Citation / Recommended Reference:** Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19-21. doi:10.1093/nar/gkq1019

**Beginner-Friendly Explanation:** Think of the SRA as a giant library where scientists store the raw data files produced by DNA sequencing machines. When researchers publish a paper that involved sequencing, they are usually required to deposit their raw sequencing files here so that other scientists can check their work or use the data for new studies. You can search for datasets by organism, experiment type, or keyword, and then download the files to run your own analysis. The files need a special tool (SRA Toolkit) to convert them into the standard FASTQ format that most analysis programs expect.

**Advanced Technical Explanation:** The SRA stores data in a compressed, proprietary .sra format that encodes base calls, quality scores, and metadata in a columnar storage scheme optimized for compression ratios. The fasterq-dump utility parallelizes decompression and FASTQ conversion using multiple threads and a temporary disk cache, substantially outperforming the legacy fastq-dump. Data submitted to SRA is organized under the INSDC metadata model: Study (SRP/ERP/DRP) -> Sample (SRS/ERS/DRS) -> Experiment (SRX/ERX/DRX) ->



Run (SRR/ERR/DRR). Cloud-native access via AWS or GCP eliminates the need for local .sra file storage by streaming reads directly into analysis pipelines using the SRA Toolkit's cloud-aware prefetch or via direct S3/GCS URIs.

### One practical workflow example:

#### Downloading and processing RNA-seq data from SRA:

- Step 1: Search SRA (<https://www.ncbi.nlm.nih.gov/sra>) for your study using keywords or a known accession (e.g., SRP123456). Use the Run Selector to identify individual SRR accessions and download the accession list as a text file.
- Step 2: Install SRA Toolkit (`conda install -c bioconda sra-tools`) and configure it with a local cache directory using `vdb-config`.
- Step 3: Download .sra files using: `prefetch --option-file SRR_Acc_List.txt`
- Step 4: Convert to FASTQ using: `fasterq-dump --split-files SRR_XXXXXXX` (use `--split-files` for paired-end data)
- Step 5: Compress output: `gzip SRR_XXXXXXX_1.fastq SRR_XXXXXXX_2.fastq`
- Step 6: Proceed with quality control (FastQC), trimming (Trimmomatic/fastp), and alignment (STAR/HISAT2) as appropriate for your experiment type.

## H2 – ENA (European Nucleotide Archive)

**Official Website URL:** <https://www.ebi.ac.uk/ena>

**Resource Type:** Repository

**Main Biological Domain:** RNA/transcriptomics / DNA sequences / Omics

**What It Is Used For:** The ENA is the European INSDC partner repository for nucleotide sequence data, maintained by the European Bioinformatics Institute (EMBL-EBI). It serves as the primary submission and access point for sequencing data from European researchers and institutions, and mirrors data from SRA and DRA through the INSDC partnership. Researchers use ENA to deposit data for European journal and funder compliance, and to retrieve raw sequencing data with the advantage of direct FASTQ download without requiring format conversion tools.

**What Data It Contains:** ENA contains raw sequencing reads (FASTQ, BAM, CRAM), assembled sequences, annotated sequences, and associated metadata spanning all sequencing technologies and biological domains. It is organized into three main sections: the Sequence Read Archive (raw reads), the Sequence Archive (assembled and annotated sequences), and the Trace Archive (legacy capillary sequencing chromatograms).

**Main question it helps answer:** Where can I find and directly download raw sequencing reads from published studies, particularly with convenient FASTQ access and rich metadata?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- Can I download FASTQ files directly from ENA for a published metagenomics study without installing additional conversion software?
- What sequencing datasets are available for a specific pathogen outbreak that I can use for phylogenetic analysis?
- How do I submit my sequencing data to comply with European journal requirements?

**Example use cases:**

- Directly downloading FASTQ files via FTP or Aspera for a large RNA-seq study, bypassing the SRA format conversion step.
- Using the ENA browser to explore metadata and select specific samples from a large multi-sample study before downloading only the relevant runs.
- Submitting sequencing data from a European research project to ENA to satisfy Wellcome Trust or EMBL funding mandates.

**Input data accepted:** ENA/SRA/DRA accession numbers (ERR, ERX, ERS, ERP, PRJEB formats); Search queries via the ENA browser or API; FASTQ, BAM, CRAM files for submission; Metadata in XML or JSON format for programmatic submission

**Output data provided:** Direct FASTQ files via FTP or Aspera (no conversion required); BAM/CRAM files where submitted in aligned format; Metadata in XML, JSON, or TSV format; FTP download links for individual runs or entire studies; Submission receipts and accession numbers for deposited data.

**Strengths:** Direct FASTQ download via FTP or Aspera without requiring SRA Toolkit format conversion — a major practical advantage over SRA; Excellent metadata browsing interface with faceted search and filtering; Strong API (ENA Portal API) for programmatic metadata retrieval and download URL generation; Aspera high-speed transfer protocol available for large downloads; Comprehensive coverage through INSDC mirroring

**Limitations:** Some very recently submitted SRA data may have a short lag before appearing in ENA due to mirroring delays; Controlled-access human data has more limited infrastructure compared to NCBI's dbGaP system; The submission interface can be complex for first-time submitters with large or complex datasets; FTP speeds can be variable depending on geographic location and network conditions

**Common beginner mistakes:** Not using Aspera (ascp) for large downloads, resulting in slow FTP transfers that may time out; Confusing ENA accession prefixes (ERR = run, ERX = experiment, ERS = sample, ERP = project) with SRA prefixes (SRR, SRX, SRS, SRP); Downloading only the metadata report without the actual FASTQ files; Not verifying MD5 checksums after download to confirm file integrity

**When to Use It:** Use ENA when you want direct FASTQ file access without installing SRA Toolkit, or when you are based in Europe and need to submit data to comply with European funding mandates. ENA is also preferable when you need programmatic access to download URLs for batch processing pipelines.

**When NOT to Use It:** ENA is not the best choice for accessing controlled-access human genetic data, where NCBI's dbGaP infrastructure is more developed. If you need processed expression data rather than raw reads, ArrayExpress or GEO are more appropriate.

**Related databases / alternatives:** SRA (Sequence Read Archive): NCBI INSDC partner; same data, different access interface; requires SRA Toolkit for FASTQ conversion; DRA (DDBJ Sequence Read Archive): Japanese INSDC partner; ArrayExpress: EBI repository for processed expression data; links to ENA; EVA (European Variation Archive): EBI repository for variant data

**How It Connects to Other Resources:** ENA is part of the EMBL-EBI resource ecosystem and links to ArrayExpress for processed expression data, the European Variation Archive (EVA) for variant data, and UniProt/Ensembl for gene and protein annotations. Through the INSDC partnership, all ENA records are synchronized with NCBI SRA and DDBJ DRA. ENA also links to the European Genome-phenome Archive (EGA) for controlled-access human data.

**API / FTP / programmatic access:** ENA Portal API: RESTful API for metadata and download URL retrieval; Endpoint: <https://www.ebi.ac.uk/ena/portal/api/> Example: [https://www.ebi.ac.uk/ena/portal/api/filereport?accession=PRJEB12345&result=read\\_run&fields=run\\_accession,fastq ftp&format=tsv](https://www.ebi.ac.uk/ena/portal/api/filereport?accession=PRJEB12345&result=read_run&fields=run_accession,fastq ftp&format=tsv) ; FTP: <ftp://ftp.sra.ebi.ac.uk/vol1/fastq/>; Aspera: era-fasp@fasp.sra.ebi.ac.uk (requires Aspera Connect client); enaBrowserTools (Python): command-line scripts for ENA data retrieval; ffq (Python): fetches FTP/Aspera URLs for SRA/ENA accessions

**Evidence/curation level:** Community-submitted; metadata reviewed for format compliance; biological content not independently curated.

**Data Update Status:** Continuously updated; INSDC synchronization occurs daily. New submissions from European researchers appear immediately; mirrored SRA/DRA data typically appears within 24-48 hours.

**Licensing / access restrictions:** Open access for the vast majority of data. Controlled-access human data is managed through the European Genome-phenome Archive (EGA) with separate data access procedures.

**Citation / Recommended Reference:** Amid C, et al. The European Nucleotide Archive in 2019. *Nucleic Acids Res.* 2020;48(D1):D70-D76. doi:10.1093/nar/gkz1063

**Beginner-Friendly Explanation:** The European Nucleotide Archive (ENA) is Europe's version of the SRA — a large public library of raw DNA sequencing files. One of its biggest advantages for beginners is that you can download files directly in FASTQ format without needing to install any special conversion software. You can search for datasets by organism, experiment type, or study accession number, and then download the files using a standard FTP client or web browser. Because ENA and SRA share data through an international agreement, most datasets available in one are also available in the other.

**Advanced Technical Explanation:** ENA stores raw reads in their original submitted format (FASTQ, BAM, or CRAM) and makes them available via FTP and Aspera without the intermediate .sra container format used by NCBI. The ENA Portal API provides a RESTful interface that returns TSV, JSON, or XML metadata including direct FTP and Aspera download URLs, enabling fully automated pipeline integration. The INSDC data exchange protocol synchronizes records between ENA, SRA, and DRA using accession cross-references at all metadata levels. ENA's submission system supports Webin-CLI for programmatic submission of large datasets with validation against controlled vocabularies and ontologies.

**One Practical Workflow Example:** Batch downloading FASTQ files from ENA using the Portal API:

**Step 1:** Identify the study accession (e.g., PRJEB12345 or SRP123456) from a publication or the ENA browser.

**Step 2:** Query the ENA Portal API to retrieve FASTQ download links and MD5 checksums:

```
"https://www.ebi.ac.uk/ena/portal/api/filereport?accession=PRJEB12345&result=read_run&fields=run_accession,fastq ftp,fastq_md5&format=tsv" > file_report.tsv
```

**Step 3:** Extract FASTQ FTP URLs from the TSV file and download using wget:

```
awk -F'\t' 'NR>1 {print $2}' file_report.tsv | tr ';' '\n' | sed 's|^|ftp:/' | wget -i -
```

Alternatively, download using curl:

```
awk -F'\t' 'NR>1 {print $2}' file_report.tsv | tr ';' '\n' | sed 's|^|ftp:/' | xargs -n1 curl -O
```

**Step 4:** Optionally, use Aspera for faster transfer of large sequencing datasets:

```
ascp -QT -l 300m -P33001 era-fasp@fasp.sra.ebi.ac.uk:/vol1/fastq/ERR123/ERR123456/ERR123456_1.fastq.gz
```

**Step 5:** Verify file integrity using MD5 checksums provided in the API response:

```
md5sum ERR123456_1.fastq.gz
```

Compare the generated MD5 value with the fastq\_md5 field in file\_report.tsv.

**Step 6:** Perform quality assessment using FastQC and proceed with downstream analysis: fastqc \*.fastq.gz

## H3 – DRA (DDBJ Sequence Read Archive)

**Official Website URL:** <https://www.ddbj.nig.ac.jp/dra>

**Resource Type:** Repository

**Main Biological Domain:** DNA sequences / RNA/transcriptomics / Omics

**What It Is Used For:** The DRA is the Japanese INSDC partner repository for raw sequencing data, maintained by the DNA Data Bank of Japan at the National Institute of Genetics. It serves as the primary submission point for sequencing data from Japanese researchers and institutions, and mirrors data from SRA and ENA through the INSDC partnership.

**What Data It Contains:** DRA contains raw sequencing reads from all major sequencing platforms, organized under the same INSDC metadata hierarchy as SRA and ENA (Study -> Sample -> Experiment -> Run). It includes data from a wide range of biological domains including genomics, transcriptomics, metagenomics, and epigenomics. DRA accessions use the DRR (run), DRX (experiment), DRS (sample), and DRP (project) prefix system. All data submitted to DRA is synchronized with SRA and ENA.

**Main question it helps answer:** Where can I find and access raw sequencing data submitted by Japanese research institutions, and how do I deposit data to comply with Japanese funding requirements?

**Typical user:** Bioinformatician / Researcher (particularly in Japan and Asia-Pacific region)

**Example Scientific Questions:** Are there raw sequencing datasets from Japanese cohort studies that I can use for population genetics analysis? – How do I submit my sequencing data to comply with AMED or JSPS data sharing requirements? – What metagenomics datasets from Japanese environmental samples are available in DRA?

**Example use cases:**

- Accessing raw sequencing data from a Japanese cancer genomics consortium study for reanalysis.
- Submitting sequencing data from a JSPS-funded project to DRA to satisfy data sharing requirements.
- Retrieving DRA accessions for datasets submitted directly to DDBJ before being mirrored to SRA/ENA.

**Input Data Accepted:** DRA accession numbers (DRR, DRX, DRS, and DRP formats), as well as cross-referenced SRA or ENA accession numbers. Data can also be identified through search queries using the DRA search interface. For data submission, the DRA accepts sequencing file formats such as FASTQ and BAM files.

**Output Data Provided:** Raw sequencing reads downloadable via FTP or Aspera, metadata in XML format, cross-references to SRA and ENA accession numbers, as well as submission receipts and assigned DRA accession numbers.

**Strengths:** DRA serves as the primary submission repository for Japanese research data, enabling timely access to datasets generated by Japanese institutions. Through full integration with the International Nucleotide Sequence Database Collaboration (INSDC), DRA datasets are also accessible via SRA and ENA platforms. The database supports Japanese-language interfaces and documentation, provides direct FTP access to raw sequencing files, and is connected to the broader DDBJ ecosystem, including DDBJ and the Japanese Genotype-phenotype Archive (JGA).

**Limitations:** DRA contains a smaller overall dataset volume compared with SRA and ENA, and its web interface and documentation are generally less polished. API functionality is also more limited than ENA's Portal API. Controlled-access human datasets are managed separately through JGA, often requiring a distinct and potentially complex application process. In addition, DRA is less commonly used by researchers outside Japan.

**Common Beginner Mistakes:** Beginners often fail to recognize that DRA data are mirrored and therefore accessible through both SRA and ENA as part of the INSDC collaboration. Another common mistake is attempting to access JGA-controlled human datasets directly through DRA without completing the required JGA authorization process. Users may also confuse DRA, which stores raw sequencing reads, with DDBJ, which primarily archives annotated nucleotide sequences.

**When to Use It:** Use DRA when you are a Japanese researcher submitting data to comply with domestic funding mandates, or when you specifically need to access datasets submitted directly to DDBJ/DRA. For most international users, accessing the same data via SRA or ENA is equally valid and may offer a more familiar interface.

**When NOT to Use It:** For most non-Japanese researchers, there is rarely a compelling reason to use DRA specifically over SRA or ENA, since all three mirror the same data. Do not use DRA for controlled-access human genetic data from Japanese studies — use JGA instead.

**Related databases / alternatives:** SRA (Sequence Read Archive): NCBI INSDC partner; mirrors all DRA data; ENA (European Nucleotide Archive): EMBL-EBI INSDC partner; mirrors all DRA data; JGA (Japanese Genotype-phenotype Archive): controlled-access human genetic data from Japan; DDBJ (DNA Data Bank of Japan): annotated nucleotide sequences (not raw reads)

**How It Connects to Other Resources:** DRA is part of the DDBJ ecosystem at NIG, which includes DDBJ (annotated sequences), JGA (controlled-access human data), and the DDBJ Analysis Pipeline. Through INSDC, all DRA records are synchronized with NCBI SRA and EMBL-EBI ENA. DRA records link to DDBJ BioProject and BioSample records, which are also synchronized with NCBI BioProject and BioSample.

**API / FTP / programmatic access:**

- FTP: [ftp://ftp.ddbj.nig.ac.jp/ddbj\\_database/dra/](ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/)
- Aspera: available for high-speed transfers
- DDBJ Search: <https://ddbj.nig.ac.jp/search> (metadata search interface)
- SRA Toolkit: can be used to download DRA data using DRR accessions

**Evidence/curation level:** Community-submitted; metadata reviewed for format compliance by DDBJ curators; biological content not independently verified.

**Data Update Status:** Continuously updated; INSDC synchronization with SRA and ENA occurs daily.

**Licensing / access restrictions:** Open access for the vast majority of data. Controlled-access human genetic data is managed through JGA with separate data access procedures.

**Citation / Recommended Reference:** Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* 2012;40(Database issue):D54-6. doi:10.1093/nar/gkr854



**Beginner-Friendly Explanation:** The DDBJ Sequence Read Archive (DRA) is Japan's contribution to the international system for sharing raw DNA sequencing data. It works as part of a three-way partnership with the American SRA and European ENA, meaning that data submitted to any one of these archives is automatically shared with the other two. If you are a researcher in Japan, DRA is where you would submit your sequencing data to meet funding requirements. For most users outside Japan, the same data can be accessed through SRA or ENA.

**Advanced Technical Explanation:** DRA implements the full INSDC metadata schema with DRR/DRX/DRS/DRP accession namespaces that are cross-referenced to SRA and ENA equivalents through the INSDC exchange protocol. Data files are stored in FASTQ or BAM format on DDBJ's high-performance storage infrastructure at NIG and are accessible via FTP and Aspera. The JGA handles controlled-access human data using a separate access control framework with approval from the NBDC Human Data Review Board. DRA's submission system (D-way) supports both web-based and command-line submission workflows with validation against INSDC metadata standards.

**One Practical Workflow Example:** Accessing DRA data for a Japanese genomics study:

- Step 1: Identify the DRA accession from a publication (e.g., DRP003456) or search the DRA browser at <https://ddbj.nig.ac.jp/search>
- Step 2: Navigate to the study page to view associated samples and runs; note individual DRR accession numbers.
- Step 3: Since DRA data is mirrored in ENA, use the ENA Portal API for convenient FASTQ URL retrieval:  

```
curl "https://www.ebi.ac.uk/ena/portal/api/filereport?accession=DRP003456&result=read_run&fields=run_accession,fastqftp&format=tsv"
```
- Step 4: Download FASTQ files via FTP from ENA or directly from DDBJ FTP: `wget ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/DRR003/DRR003456/DRR003456.fastq.bz2`
- Step 5: Decompress (bzip2 -d) and proceed with quality control and analysis.

## BEGINNER EXAMPLE (Category H):

A graduate student reads a paper about RNA-seq analysis of Alzheimer's disease brain tissue and wants to reproduce the analysis with a newer alignment tool. They find the SRA accession (SRP123456) in the paper's data availability statement. They go to <https://www.ncbi.nlm.nih.gov/sra>, search for the accession, use the Run Selector to download the accession list, then use `fasterq-dump --split-files` on each SRR accession to obtain paired FASTQ files. Alternatively, they use the ENA Portal API to get direct FTP download links, avoiding the need for SRA Toolkit entirely.

## ADVANCED EXAMPLE (Category H):

A bioinformatician is building a meta-analysis pipeline that needs to automatically download and process 500 RNA-seq samples from 15 different SRA studies. They write a Snakemake workflow that: (1) queries the ENA Portal API for all run accessions and FTP URLs associated with each study accession; (2) downloads FASTQ files in parallel using Aspera with a bandwidth cap; (3) runs FastQC and MultiQC for



quality assessment; (4) aligns with STAR to GRCh38; (5) quantifies with featureCounts; and (6) performs batch correction with ComBat-seq before differential expression analysis. The pipeline uses pysradb to retrieve and validate metadata, cross-referencing SRA and ENA accessions to ensure no duplicates are included.

## CONFUSION POINTS:

---

**SRA vs. GEO:** SRA contains raw reads; GEO contains processed data. Many studies deposit in both. The GEO record links to the SRA accession for raw reads. Do not confuse the GSE (GEO series) accession with the SRP (SRA study) accession.

**SRA vs. ENA vs. DRA:** These three archives mirror each other through INSDC.

The same dataset has different accession numbers in each (SRR/ERR/DRR). Choose based on convenience of access, not data content.

**Controlled vs. open access:** Most SRA data is open, but human genetic data with privacy implications is in dbGaP (NCBI) or EGA (EBI) and requires separate application.

**Raw reads vs. assembled sequences:** SRA/ENA/DRA store raw reads. Assembled genome sequences are in NCBI Assembly/GenBank or ENA's sequence archive.

## DECISION GUIDE:

---

**Use SRA if:** you are comfortable with SRA Toolkit, you need tight integration with NCBI resources (BioProject, GEO, PubMed), or you need cloud access via AWS.

**Use ENA if:** you want direct FASTQ download without format conversion, you need a powerful API for batch URL retrieval, or you are submitting data from a European institution.

**Use DRA if:** you are a Japanese researcher submitting data, or you need to access data submitted directly to DDBJ.

**In practice:** for downloading, ENA is often the most convenient; for submission, use the archive appropriate to your institution's location.

## Category I: Protein Sequence and Function Databases

### CATEGORY OVERVIEW

Protein sequence and function databases are curated repositories that store amino acid sequences alongside functional annotations describing what a protein does, where it is found, how it is regulated, and what structural or functional domains it contains. These databases are central to molecular biology and biochemistry research because they transform raw sequence data — the output of genome sequencing or proteomics experiments — into biologically meaningful information. The most important resource in this category is UniProt, which integrates two complementary sections: Swiss-Prot (manually reviewed, high- confidence annotations) and TrEMBL (computationally predicted annotations for the vast majority of known protein sequences). Complementary resources such as InterPro, PROSITE, and SMART provide specialized information about protein domains, families, and functional motifs.

Researchers need protein sequence and function databases at multiple stages of a typical molecular biology project. When a new gene is identified through sequencing or genetic mapping, the first step is usually to search these databases to determine whether the encoded protein has known homologs with characterized functions. When designing experiments to study a protein of interest, these databases provide information about known isoforms, post-translational modifications, subcellular localization, interaction partners, and disease associations. In computational biology, protein function databases serve as the gold standard for training and evaluating machine learning models for function prediction, and as the annotation source for gene ontology enrichment analyses. Proteomics researchers use these databases as search libraries for peptide identification by mass spectrometry.

The distinction between manually reviewed and computationally predicted annotations is critically important in this category. Swiss-Prot entries have been individually reviewed by expert curators who read the primary literature, evaluate experimental evidence, and assign annotations with explicit evidence codes. TrEMBL entries, by contrast, are annotated automatically using rule- based systems (UniRule and HAMAP) and sequence similarity to reviewed entries. The practical consequence is that Swiss-Prot annotations are highly reliable but cover only a small fraction of known proteins (~570,000 entries as of 2024), while TrEMBL covers hundreds of millions of sequences but with variable annotation quality. For critical functional claims, researchers should always verify that the annotation comes from a Swiss-Prot entry or is supported by experimental evidence codes (ECO:0000269 for experimental evidence).

## I1 – UniProt (Universal Protein Resource)

---

**Official Website URL:** <https://www.uniprot.org>

**Resource Type:** Knowledgebase

**Main Biological Domain:** Proteins

**What It Is Used For:** UniProt is the world's most comprehensive and widely used protein sequence and functional annotation database. Researchers use it to look up protein sequences, retrieve functional annotations (molecular function, biological process, subcellular localization), find information about post-translational modifications, identify disease associations, and access cross-references to hundreds of other biological databases. It serves as the primary reference database for protein sequences in proteomics, structural biology, and computational biology.

**What Data It Contains:** UniProt contains protein sequences from all organisms, with functional annotations including Gene Ontology terms, enzyme classification numbers, pathway memberships, subcellular localization, tissue expression, post-translational modifications, natural variants, active sites, binding sites, and disease associations. It integrates data from Swiss-Prot (manually reviewed) and TrEMBL (computationally annotated) sections, and provides cross-references to over 150 external databases including PDB, Ensembl, KEGG, Reactome, and PubMed.

**Main question it helps answer:** What is the function, structure, localization, and biological context of this protein sequence?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist / Clinician

**Example scientific questions:**

- What is the known function and subcellular localization of the human TP53 protein, and what disease variants have been reported?
- Which proteins in the human proteome are annotated as kinases and are localized to the nucleus?
- What are the known post-translational modifications of BRCA1, and which are supported by experimental evidence?

**Example use cases:**

- Retrieving the canonical sequence and all known isoforms of a human protein for use as a reference in mass spectrometry database searching.
- Downloading all Swiss-Prot entries for a specific organism to build a local annotation database for a bioinformatics pipeline.
- Using the UniProt ID mapping tool to convert between gene names, Ensembl IDs, RefSeq accessions, and UniProt accessions.

**Input data accepted:**

- UniProt accession numbers (e.g., P04637 for human TP53)
- Gene names, protein names, organism names
- Protein sequences (for BLAST search)
- Accession numbers from other databases (for ID mapping)

- Advanced query syntax for filtered searches

### Output data provided

- Full protein sequence in FASTA format
- Functional annotations with evidence codes
- Sequence features (domains, active sites, modifications) in flat file or JSON format
- Cross-references to external databases
- Downloadable datasets in UniProtKB flat file, FASTA, TSV, or JSON format
- ID mapping results

### Strengths:

- Most comprehensive protein sequence and function database available
- Swiss-Prot section provides expert-curated, literature-backed annotations of the highest quality
- Extensive cross-references to over 150 external databases
- Powerful query language for complex filtered searches
- Excellent REST API and programmatic access tools
- Regular releases with clear versioning and change tracking

### Limitations:

- The vast majority of UniProt entries are in TrEMBL with computationally predicted annotations of variable reliability
- Manual curation in Swiss-Prot cannot keep pace with the exponential growth of sequencing data
- Annotations for non-model organisms are often sparse or absent
- Some functional annotations are transferred by sequence similarity and may not reflect the actual function of the specific protein
- The web interface can be slow for very large result sets

### Common beginner mistakes:

- Not distinguishing between Swiss-Prot (reviewed) and TrEMBL (unreviewed) entries, leading to over-reliance on computationally predicted annotations
- Using gene names for searches without specifying the organism, resulting in hits from multiple species
- Not checking evidence codes for annotations — an annotation labeled "By similarity" is not experimentally verified
- Confusing UniProt accession numbers (e.g., P04637) with entry names (e.g., P53\_HUMAN)

**When to Use It:** Use UniProt whenever you need functional information about a protein, including its molecular function, biological process, subcellular localization, domain structure, post-translational modifications, or disease associations. It is the first resource to consult when characterizing a newly identified protein or when building annotation pipelines.

**When NOT to Use It:** UniProt is not the best resource for three-dimensional structural information (use PDB/AlphaFold), for variant frequency data in human populations (use gnomAD), or for pathway-level analysis

(use KEGG or Reactome). For raw protein sequences from genome projects without functional annotation, NCBI RefSeq or Ensembl may be more appropriate.

### Related databases / alternatives:

- NCBI Protein: Less curated but broader coverage of genome-predicted proteins
- RefSeq: NCBI's curated reference sequence database for proteins and genes
- InterPro: domain and family annotations; integrated with UniProt
- neXtProt: human protein-focused knowledgebase with additional detail

**How It Connects to Other Resources:** UniProt serves as a central hub in the protein biology data ecosystem, providing cross-references to structural databases (PDB, AlphaFold), pathway databases (KEGG, Reactome, BioCyc), domain databases (InterPro, Pfam, PROSITE), genome databases (Ensembl, RefSeq), disease databases (OMIM, ClinVar), and literature (PubMed). The UniProt ID mapping service allows conversion between virtually all major biological identifier systems.

### API / FTP / programmatic access:

- REST API: <https://rest.uniprot.org/> - Example: <https://rest.uniprot.org/uniprotkb/P04637.json>
- SPARQL endpoint: <https://sparql.uniprot.org/sparql>
- FTP: <https://ftp.uniprot.org/pub/databases/uniprot/>
- Python: requests library with REST API; also unipressed package
- Batch retrieval: POST to <https://rest.uniprot.org/idmapping/run>
- UniProt BLAST: <https://www.uniprot.org/blast>

**Evidence/curation level:** Mixed: Swiss-Prot section is manually reviewed by expert curators with literature-based evidence codes; TrEMBL section is computationally annotated using UniRule and HAMAP rule systems.

**Data Update Status:** Major releases approximately every 8 weeks; Swiss-Prot is updated continuously with new manual curation; TrEMBL is updated with each new genome/proteome submission to INSDC.

**Licensing / access restrictions:** Fully open access under Creative Commons Attribution 4.0 (CC BY 4.0). All data freely downloadable and reusable with attribution.

**Citation / Recommended Reference:** The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in Nucleic Acids Res. 2023;51(D1):D523-D531. doi:10.1093/nar/gkac1052

**Beginner-Friendly Explanation:** UniProt is like an encyclopedia for proteins. For almost any protein you can think of, UniProt will tell you what it does in the cell, where in the cell it is found, what other proteins it interacts with, and whether mutations in it cause disease. It has two sections: Swiss-Prot, where expert scientists have carefully reviewed and verified the information, and TrEMBL, where the information has been predicted by computer programs. When you look up a protein, always check whether you are reading a Swiss- Prot (reviewed) or TrEMBL (unreviewed) entry, as this tells you how reliable the information is.

**Advanced Technical Explanation:** UniProtKB entries are structured records containing sequence data, feature annotations with positional information (active sites, binding sites, disulfide bonds, PTM sites), controlled vocabulary terms from Gene Ontology and UniProt-specific taxonomies, and evidence attribution using the

Evidence and Conclusion Ontology (ECO). Swiss-Prot curation follows a defined standard operating procedure that includes literature review, experimental evidence evaluation, and propagation of annotations to closely related proteins. The UniRef clusters (UniRef100, UniRef90, UniRef50) provide non-redundant sequence sets at different identity thresholds for use in sequence analysis. The UniParc archive stores all unique sequences ever submitted to major sequence databases, providing a historical record independent of annotation status.

**One Practical Workflow Example:** Retrieving functional annotations for a protein of interest:

- Step 1: Go to <https://www.uniprot.org> and search for your protein using gene name + organism (e.g., "BRCA1 human" or "TP53 Homo sapiens").
- Step 2: Select the Swiss-Prot (reviewed) entry if available (indicated by a gold star icon). Note the UniProt accession (e.g., P38398).
- Step 3: Review the "Function" section for molecular function, biological process, and pathway annotations with evidence codes.
- Step 4: Check the "PTM/Processing" section for post-translational modifications and the "Disease & Variants" section for pathogenic variants.
- Step 5: Use the "Cross-references" section to navigate to PDB for structures, Ensembl for genomic context, or OMIM for disease info.
- Step 6: For programmatic access, retrieve the full entry as JSON: `curl https://rest.uniprot.org/uniprotkb/P38398.json`

## I2 – Swiss-Prot (manually reviewed section of UniProtKB)

**Official Website URL:** <https://www.uniprot.org/uniprotkb?query=reviewed:true>

**Resource Type:** Knowledgebase

**Main Biological Domain:** Proteins

**What It Is Used For:** Swiss-Prot is the manually reviewed, expert-curated section of UniProtKB, representing the gold standard for protein sequence and functional annotation. Researchers use Swiss-Prot when they need high-confidence functional information that has been verified against primary literature, or when they need a non-redundant, high-quality reference proteome for computational analyses such as mass spectrometry database searching or sequence alignment benchmarking.

**What Data It Contains:** Swiss-Prot contains approximately 570,000 protein entries (as of 2024) with comprehensive, manually verified annotations including molecular function, biological process, subcellular localization, active sites, binding sites, post-translational modifications, natural variants, disease associations, and literature references. Each annotation is accompanied by an evidence code indicating whether it is based on experimental data, inferred from sequence similarity, or predicted computationally.

**Main question it helps answer:** What is the experimentally verified or expert-curated function of this protein, and what evidence supports each annotation?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:**

- What experimental evidence supports the annotation of this enzyme's catalytic mechanism?
- Which Swiss-Prot entries for human proteins are annotated as involved in DNA repair?
- What are the known active site residues of this protease, and are they conserved in the homolog I am studying?

**Example use cases:**

- Using Swiss-Prot as the search database for proteomics mass spectrometry experiments to minimize false positives from low-quality annotations.
- Downloading all Swiss-Prot human entries to build a high-confidence training set for a machine learning protein function predictor.
- Verifying that a functional annotation found in TrEMBL is supported by experimental evidence in the corresponding Swiss-Prot entry.

**Input Data Accepted:** UniProt accession numbers, gene names, and protein names combined with the reviewed:true filter. Users can also submit protein sequences for BLAST searches restricted to Swiss-Prot entries, as well as perform advanced queries such as reviewed:true AND organism\_id:9606 to retrieve reviewed human protein records.

**Output Data Provided:** fully annotated protein entries with evidence codes, FASTA protein sequences, flat file format (UniProtKB format) containing all annotation fields, JSON and XML formats for programmatic access, and downloadable complete Swiss-Prot datasets.



### Strengths:

- Highest annotation quality of any protein database; each entry individually reviewed by expert curators
- Explicit evidence codes allow users to distinguish experimental from predicted annotations
- Non-redundant: one canonical entry per protein per species
- Stable accession numbers that persist across database releases

### Limitations:

- Covers only ~570,000 proteins — a tiny fraction of all known protein sequences
- Curation cannot keep pace with the exponential growth of sequencing data
- Coverage is heavily biased toward well-studied model organisms (human, mouse, yeast, E. coli)
- Some annotations, even in Swiss-Prot, are transferred by similarity and may not reflect the actual function of the specific protein
- New proteins from recently sequenced organisms may wait years for manual curation

### Common beginner mistakes:

- Assuming all UniProt entries are Swiss-Prot quality (most are TrEMBL)
- Not filtering for reviewed:true when a high-confidence dataset is needed
- Treating "By similarity" annotations as experimentally verified
- Not checking the evidence code for each specific annotation field

**When to Use It:** Use Swiss-Prot when annotation quality is paramount — for example, when building training datasets for machine learning, when verifying functional claims for a manuscript, or when you need a non-redundant high-quality reference proteome for computational analysis.

**When NOT to Use It:** Do not use Swiss-Prot alone when you need comprehensive coverage of a specific proteome, especially for non-model organisms where Swiss-Prot coverage is sparse. For broad coverage, use the full UniProtKB (Swiss-Prot + TrEMBL) or organism-specific proteome downloads.

### Related databases / alternatives:

- TrEMBL: computationally annotated complement to Swiss-Prot
- neXtProt: human-focused protein knowledgebase with additional detail
- NCBI RefSeq: curated reference sequences, less detailed functional annotation than Swiss-Prot

**How It Connects to Other Resources:** Swiss-Prot entries serve as the authoritative source for protein annotations propagated to many other databases. InterPro, Pfam, and PROSITE use Swiss-Prot annotations to validate domain and motif assignments. Gene Ontology annotations in Swiss-Prot are contributed to the GO Consortium. PDB structures are cross-referenced to Swiss-Prot entries, and KEGG/Reactome pathway entries link to Swiss-Prot accessions.

### API / FTP / programmatic access:

- REST API with reviewed filter: <https://rest.uniprot.org/uniprotkb/search?query=reviewed:true&format=fasta>;  
FTP download of complete Swiss-Prot: [https://ftp.uniprot.org/pub/databases/uniprot/current\\_release/knowledgebase/complete/uniprot\\_sprot.fasta.gz](https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz);  
Flat file: uniprot\_sprot.dat.gz (same FTP location)

**Evidence/curation level:** Manually reviewed by expert curators; annotations supported by literature evidence codes (ECO:0000269 for experimental, ECO:0000250 for by similarity, ECO:0000255 for sequence analysis).

**Data Update Status:** Updated with each UniProt release (approximately every 8 weeks)

**Licensing / access restrictions:** : Fully open access under Creative Commons Attribution 4.0 (CC BY 4.0).

**Citation / Recommended Reference:** Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28(1):45-8. doi:10.1093/nar/28.1.45 (For current citation, use the UniProt Consortium paper above)

**Beginner-Friendly Explanation:** Swiss-Prot is the carefully checked high-quality section of the UniProt protein database. Think of it as the difference between a Wikipedia article that has been reviewed by experts versus one that was written by a computer program. Every entry in Swiss-Prot has been read and verified by a scientist who checked the original research papers. This makes Swiss-Prot smaller than the full UniProt database, but much more reliable. When you need to be sure that the information about a protein is accurate, Swiss-Prot is the place to look.

**Advanced Technical Explanation:** Swiss-Prot curation follows a standardized workflow in which curators perform literature mining, evaluate experimental evidence for each annotation, assign ECO evidence codes, and propagate annotations to closely related proteins using HAMAP family rules. The canonical sequence concept in Swiss-Prot defines a single representative sequence per protein per species, with isoforms described as sequence variants. Merge/demerge operations maintain accession history so that deprecated accessions redirect to current entries. The annotation score (1-5 stars) provides a quick quality indicator based on the number and type of annotations present.

**One Practical Workflow Example:** Downloading Swiss-Prot for use as a proteomics search database:

Step 1: Go to [https://ftp.uniprot.org/pub/databases/uniprot/current\\_release/knowledgebase/complete/](https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/)

Step 2: Download uniprot\_sprot.fasta.gz and uniprot\_sprot\_varsplic.fasta.gz (isoform sequences) if needed.

Step 3: Decompress: gunzip uniprot\_sprot.fasta.gz

Step 4: For species-specific searches, filter by taxonomy: `grep -A1 "OS=Homo sapiens" uniprot_sprot.fasta | grep "^>" | cut -d"|" -f2 > human_accessions.txt` (Or use the REST API with organism\_id:9606 filter)

Step 5: Use the resulting FASTA file as the search database in your proteomics software.

Step 6: Add common contaminants (cRAP database) to the search database before running the search.

## I3 – TrEMBL (computationally annotated section of UniProtKB)

**Official Website URL:** <https://www.uniprot.org/uniprotkb?query=reviewed:false>

**Resource Type:** Knowledgebase

**Main Biological Domain:** Proteins

**What It Is Used For:** TrEMBL is the computationally annotated section of UniProtKB, containing protein sequences that have not yet been manually reviewed for Swiss-Prot. It provides broad coverage of protein sequences from genome sequencing projects, including sequences from non-model organisms, environmental metagenomes, and newly sequenced species. Researchers use TrEMBL when they need to find sequences for proteins from organisms not covered by Swiss-Prot, or when performing large-scale comparative genomics analyses requiring comprehensive sequence coverage.

**What Data It Contains:** TrEMBL contains hundreds of millions of protein sequences with automated annotations generated by the UniRule and HAMAP rule-based systems, which transfer annotations from Swiss-Prot entries to similar sequences based on sequence identity and domain content. Annotations include predicted function, subcellular localization, and domain assignments, but with lower confidence than Swiss-Prot. The database grows rapidly as new genome sequences are deposited in INSDC databases.

**Main question it helps answer:** What protein sequences are available for this organism or gene family, including those not yet manually curated?

**Typical user:** Bioinformatician / Researcher (comparative genomics, metagenomics)

**Example scientific questions:**

- What protein sequences are available for a newly sequenced bacterial species that has no Swiss-Prot entries?
- How many proteins in the human proteome are still in TrEMBL and have not been manually reviewed?
- What is the predicted function of this hypothetical protein from a metagenomic assembly?

**Example use cases:**

- Using TrEMBL as a comprehensive sequence database for BLAST searches when Swiss-Prot does not contain the organism of interest.
- Downloading all TrEMBL entries for a specific taxon to build a local proteome database for a non-model organism.
- Identifying which proteins in a newly sequenced proteome have TrEMBL entries with functional predictions.

**Input Data Accepted:** UniProt accession numbers, gene names and organism names combined with the reviewed:false filter, protein sequences for BLAST searches, and taxonomy IDs for organism-specific queries.

**Output Data Provided:** Protein sequences in FASTA format, automated functional annotations with evidence codes, cross-references to source genome databases, and downloadable complete TrEMBL datasets, which are very large and may exceed 100 GB in compressed form.

**Strengths:** Extremely broad coverage, containing hundreds of millions of protein sequences from non-model organisms, environmental metagenomes, and newly sequenced species. Automated annotations provide a useful starting point for functional characterization, and the database is regularly updated as new genome sequences are

deposited. Additionally, it uses the same interface and API as Swiss-Prot, ensuring consistent and convenient access.

**Limitations:** Lower annotation quality compared with Swiss-Prot, as many annotations are computational predictions or transferred by sequence similarity. The database also contains high redundancy, with many similar sequences from closely related organisms. Functional annotations for novel proteins may be incorrect or absent, and the very large file sizes can make bulk downloads challenging. In addition, entries labeled as “hypothetical protein” or “uncharacterized protein” often provide little functional information.

#### Common beginner mistakes:

- Treating TrEMBL annotations as experimentally verified facts
- Using TrEMBL as a proteomics search database without filtering, leading to inflated false discovery rates
- Not recognizing that most UniProt entries are TrEMBL, not Swiss-Prot
- Downloading the full TrEMBL database without realizing its enormous size

**When to Use It:** Use TrEMBL when you need broad sequence coverage for organisms not well represented in Swiss-Prot, or when performing large-scale comparative analyses where completeness is more important than annotation quality. Always treat TrEMBL annotations as hypotheses to be verified.

**When NOT to Use It:** Do not use TrEMBL as the sole source of functional information for a specific protein without verifying the annotation against primary literature. For high-confidence proteomics searches, Swiss-Prot is preferable.

#### Related databases / alternatives:

- Swiss-Prot: manually reviewed complement; use when quality > quantity
- NCBI nr (non-redundant): broad coverage protein database from NCBI
- RefSeq: NCBI curated reference sequences

**How It Connects to Other Resources:** TrEMBL entries are automatically cross-referenced to source genome records in INSDC databases, and annotations are propagated from Swiss-Prot using UniRule rules. As proteins accumulate experimental evidence, TrEMBL entries are promoted to Swiss-Prot through manual curation.

#### API / FTP / programmatic access:

REST API: <https://rest.uniprot.org/uniprotkb/search?query=reviewed:false;> FTP: [https://ftp.uniprot.org/pub/databases/uniprot/current\\_release/knowledgebase/complete/uniprot\\_trembl.fasta.gz](https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.fasta.gz) (very large file); Proteome-specific downloads recommended over full TrEMBL download

**Evidence/curation level:** Computationally predicted; annotations generated by UniRule and HAMAP automated rule systems; not manually reviewed.

**Data Update Status:** Updated with each UniProt release (approximately every 8 weeks); grows rapidly with new genome sequencing projects.

**Licensing / access restrictions:** Fully open access under Creative Commons Attribution 4.0 (CC BY 4.0).

**Citation / Recommended Reference:** Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28(1):45-8. doi:10.1093/nar/28.1.45

**Beginner-Friendly Explanation:** TrEMBL is the large, automatically annotated section of UniProt. While Swiss-Prot has been carefully checked by human experts, TrEMBL entries are annotated by computer programs that look at the protein's sequence and compare it to already-known proteins. This means TrEMBL covers far more proteins — hundreds of millions — but the information is less reliable. Think of TrEMBL as a first draft: useful for getting an initial idea of what a protein might do but always require verification before you rely on it for important conclusions.

**Advanced Technical Explanation:** TrEMBL annotations are generated through the UniRule system, which applies condition-action rules derived from Swiss-Prot annotations to sequences meeting defined criteria (sequence identity thresholds, domain content, taxonomic constraints). HAMAP (High-quality Automated and Manual Annotation of Proteins) provides family-specific annotation rules for bacterial and archaeal proteins. The annotation pipeline also incorporates InterPro domain predictions, signal peptide and transmembrane topology predictions, and subcellular localization predictions. Redundancy in TrEMBL is partially addressed by the UniRef clustering system (UniRef100/90/50).

**One Practical Workflow Example:** Finding TrEMBL sequences for a non-model organism:

- Step 1: Go to <https://www.uniprot.org> and search with: organism:"Drosophila virilis" AND reviewed:false
- Step 2: Filter results by protein name or GO term if needed.
- Step 3: Download results as FASTA using the Download button, selecting "Uncompressed" or "Compressed" FASTA format.
- Step 4: For programmatic access: `curl "https://rest.uniprot.org/uniprotkb/search?query=organism_id:7244+AND+reviewed:false&format=fasta&compressed=true" -o trembl_dv.fasta.gz`
- Step 5: Use the sequences for BLAST searches, phylogenetic analysis, or as a proteomics search database.
- Step 6: For any protein of interest, check whether a Swiss-Prot entry exists for a homolog to get higher-quality functional annotations.

## I4 – InterPro — Cross-reference Entry

**Official Website URL:** <https://www.ebi.ac.uk/interpro>

**Entry type:** Cross-reference entry — full database card provided in Category J, J2.

InterPro is mentioned in Category I because protein function interpretation often depends on identifying domains, motifs, families, and conserved signatures. However, InterPro belongs more directly to protein family, domain, and motif resources. For full coverage, see Category J: Protein Family, Domain, and Motif Databases, J2 – InterPro.

## I5 – PROSITE — Cross-reference Entry

**Official Website URL:** <https://prosite.expasy.org>

**Entry type:** Cross-reference entry — full database card provided in Category J, J4.

PROSITE is mentioned in Category I because protein functional interpretation may require motif and domain pattern recognition. However, PROSITE is primarily a protein motif, domain, and signature database. For full coverage, see Category J: Protein Family, Domain, and Motif Databases, J4 – PROSITE.

## I6 – SMART (Simple Modular Architecture Research Tool)

**Official Website URL:** <https://smart.embl.de>

**Resource Type:** Database / Tool

**Main Biological Domain:** Proteins

**What Is Used For:** SMART is a database and web tool for the identification and analysis of protein domains, with a particular focus on signaling domains, extracellular domains, and domains involved in cell communication. Researchers use SMART to analyze the domain architecture of signaling proteins, compare domain combinations across species, and identify proteins containing specific domain types. SMART provides both a curated domain database and an interactive visualization tool for domain architecture analysis.

**What Data It Contains:** SMART contains manually curated HMM-based models for approximately 1,300 domain families, with depth in signaling domains (kinase domains, SH2, SH3, PH, PDZ, WD40, etc.) and extracellular domains. Each domain entry includes a description, literature references, phylogenetic distribution, and links to known structures. SMART also provides pre-computed domain annotations for all proteins in major model organism proteomes.

**Main question it helps answer:** What signaling domains and modular architecture does this protein contain, and how does this compare to related proteins across species?

**Typical user:** Researcher / Bioinformatician (particularly in cell signaling and developmental biology)

**Example scientific questions:**

- What domain architecture does this receptor tyrosine kinase have, and how does it compare to other family members?
- Which proteins in the human proteome contain both a kinase domain and a RING finger domain?
- What is the evolutionary history of the SH2 domain across eukaryotes?

**Example use cases:** Analyzing the domain architecture of a newly identified signaling protein to predict its function and interaction partners; Comparing domain combinations in a protein family across multiple species to understand evolutionary diversification; Identifying all proteins in a proteome that contains a specific domain combination relevant to a signaling pathway.

**Input Data Accepted:** Protein sequences in FASTA format, UniProt accession numbers, and SMART domain accession numbers (SM format).

**Output Data Provided:** Outputs include domain architecture diagrams displaying domain positions and E-values, links to SMART domain entries with descriptions and supporting literature, comparisons of domain architecture across species, and downloadable domain annotation results.

**Strengths:** SMART provides particularly strong coverage of signaling and extracellular domains, along with interactive domain architecture visualization and phylogenetic distribution information for each domain. It integrates structural information through links to PDB entries and is especially useful for comparative analyses of domain combinations across proteins and species.



**Limitations:** SMART has smaller domain coverage compared with broader resources such as Pfam or InterPro, and its update frequency has decreased in recent years. The web interface may be slow when handling large-scale analyses, and the database is generally less suitable for annotating non-signaling domains compared with Pfam.

**Common Beginner Mistakes:** A frequent mistake is using SMART as the sole domain annotation resource without cross-referencing results with Pfam or InterPro to achieve more comprehensive coverage. Another common error is failing to distinguish between SMART's "normal" and "genomic" modes, where genomic mode includes both SMART and Pfam domains while normal mode reports only SMART domains.

**When to Use It:** Use SMART when you are specifically interested in signaling domains, extracellular domains, or the modular architecture of cell communication proteins. It is particularly useful for comparative analysis of domain combinations in signaling protein families.

**When NOT to Use It:** For comprehensive domain annotation of a full proteome, use InterProScan rather than SMART alone. SMART is not the best choice for metabolic enzymes or structural proteins where its domain coverage is limited.

**Related databases / alternatives:** Pfam: broader domain coverage; HMM-based; InterPro: integrates SMART with other databases; PROSITE: pattern-based domain detection

**How It Connects to Other Resources:** SMART is a member database of InterPro and its domain models are included in InterProScan. SMART links to PDB structures, UniProt entries, and literature. The SMART database is maintained at EMBL Heidelberg.

**API / FTP / programmatic access:** Web interface at <https://smart.embl.de> for individual sequence analysis; Batch analysis available through the web interface; SMART is included in InterProScan for programmatic large-scale analysis; FTP access limited; use InterPro FTP for bulk data

**Evidence/curation level:** Manually curated domain models; HMM profiles derived from curated multiple sequence alignments.

**Data Update Status:** Updates have become less frequent; the database is maintained but not as actively expanded as in earlier periods. Check the website for current release information.

**Licensing / access restrictions:** Freely accessible for academic use.

**Citation / Recommended Reference:** Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 2015;43(Database issue):D257-60. doi:10.1093/nar/gku949

**Beginner-Friendly Explanation:** SMART is a specialized tool for analyzing the "building blocks" (domains) of proteins involved in cell signaling — the molecular communication systems that tell cells when to grow, divide, or respond to their environment. If you have a protein that you think might be involved in signaling, SMART can show you which functional modules it contains and compare its architecture to similar proteins in other organisms. It is particularly good at recognizing the types of domains found in receptors, kinases, and other signaling proteins.

**Advanced Technical Explanation:** SMART uses profile HMMs built from manually curated multiple sequence alignments of domain families, with particular emphasis on domains involved in signal transduction, cell-cell communication, and extracellular recognition. The database operates in two modes: "normal" mode (SMART domains only) and "genomic" mode (SMART + Pfam domains), the latter providing more comprehensive coverage.





Domain boundaries are defined by the HMM match states and overlapping domain predictions are resolved by E-value ranking. The phylogenetic distribution tool uses pre-computed domain annotations across 13 model organism proteomes to display the evolutionary conservation of each domain.

**One Practical Workflow Example:** Analyzing the domain architecture of a signaling protein:

Step 1: Go to <https://smart.embl.de> and paste your protein sequence or enter a UniProt accession in the sequence input box.

Step 2: Select "Genomic" mode to include both SMART and Pfam domains for comprehensive coverage.

Step 3: Submit the analysis and review the domain architecture diagram, which shows all predicted domains with their positions and E-values.

Step 4: Click on each domain to access the SMART entry with description, literature, and phylogenetic distribution.

Step 5: Use the "Compare architecture" feature to find other proteins with similar domain combinations.

Step 6: Cross-reference with InterPro and UniProt for additional functional context and experimental evidence.

## BEGINNER EXAMPLE (Category I)

---

A student identifies a novel gene in a GWAS study and wants to understand what the encoded protein does. They search UniProt for the gene name (e.g., "PCSK9 human"), find the Swiss-Prot entry (Q8NBP7), and read the Function section to learn it is a serine protease involved in LDL receptor degradation. They check the evidence codes to confirm which annotations are experimentally verified, then use the cross-references to navigate to OMIM for disease associations and Reactome for pathway context.

## ADVANCED EXAMPLE (Category I)

---

A bioinformatician is annotating a newly sequenced fungal genome. They run InterProScan on all predicted proteins to assign domain annotations and GO terms. They then use the PROSITE patterns to identify proteins with specific catalytic motifs, and cross-reference with Swiss-Prot to find well-characterized homologs. For proteins with no InterPro hits, they perform BLAST searches against TrEMBL to find the closest characterized relatives. The resulting annotation table is used for GO enrichment analysis of differentially expressed genes.

## CONFUSION POINTS

---

UniProt vs. Swiss-Prot vs. TrEMBL: UniProt is the umbrella database containing both Swiss-Prot (reviewed) and TrEMBL (unreviewed). When people say "UniProt entry," they may mean either section.

InterPro vs. Pfam vs. PROSITE: These are all domain databases, but InterPro integrates the others. Pfam uses HMMs, PROSITE uses patterns/profiles, and InterPro combines both plus many more.

Annotation by similarity: Many UniProt annotations are transferred from characterized proteins by sequence similarity. These are labeled "By similarity" and are not experimentally verified for the specific protein.

## DECISION GUIDE

---

For high-quality functional annotation of a specific protein: Swiss-Prot

For broad sequence coverage including non-model organisms: TrEMBL

For comprehensive domain annotation of a proteome: InterPro/InterProScan

**For specific functional motif detection:** PROSITE

**For signaling domain analysis:** SMART

**For all-in-one protein information:** UniProt (start here, then drill down)

## Category J: Protein Family, Domain, and Motif Databases

### CATEGORY OVERVIEW

Protein family, domain, and motif databases provide systematic classifications of proteins based on shared sequence features that reflect evolutionary relationships and functional conservation. A protein domain is an independently folding structural and functional unit that can occur in different combinations in different proteins; a protein family groups proteins that share a common evolutionary ancestor and typically a common function; and a motif is a short, conserved sequence pattern associated with a specific function or modification. These databases are essential for inferring the function of uncharacterized proteins by identifying their structural and functional building blocks, and for understanding the evolutionary history of protein families across the tree of life.

The computational methods used by these databases vary significantly and complement each other. Pfam and TIGRFAMs use profile hidden Markov models (HMMs) built from curated multiple sequence alignments, which are sensitive enough to detect remote homologs with low sequence identity. PROSITE uses regular expression patterns and position-specific scoring matrices, which are more interpretable but less sensitive for divergent sequences. PRINTS uses fingerprints — sets of conserved motifs that together characterize a family — which can be more specific than single-domain models. SUPERFAMILY uses structural classifications from SCOP to assign proteins to superfamilies based on predicted structural similarity. CDD (Conserved Domain Database) at NCBI integrates multiple domain databases and provides a convenient interface for NCBI users. Each approach has strengths and weaknesses, which is why InterPro integrates all of them.

A key development in this category is the migration of Pfam into InterPro. Since 2022, Pfam has been fully integrated into the InterPro infrastructure at EMBL-EBI, and the standalone Pfam website ([pfam.xfam.org](http://pfam.xfam.org)) has been retired. Pfam data is now accessible through the InterPro website and API, and Pfam accessions (PF numbers) remain valid and searchable within InterPro. This integration means that researchers who previously used Pfam directly should now use InterPro as their primary interface, while Pfam HMM models remain available for download and use with HMMER software. Understanding this transition is important to avoid confusion when following older tutorials or methods papers that reference the standalone Pfam website.

## J1 – Pfam (Protein Families Database) NOTE: Pfam has been integrated into InterPro.

**Official Website URL:** <https://www.ebi.ac.uk/interpro> (Pfam data accessible here) Legacy URL (redirects): <https://pfam.xfam.org> (retired)

**Resource Type:** Database (now integrated into InterPro)

**Main Biological Domain:** Proteins

**What It Is Used For:** Pfam is a database of protein families represented by profile hidden Markov models (HMMs) built from curated multiple sequence alignments. It is used to identify protein domains in query sequences, classify proteins into families, and annotate proteomes at scale. Pfam HMMs are widely used in genome annotation pipelines and are the most commonly cited domain database in the literature. Since integration into InterPro, Pfam data is accessed through the InterPro interface, but Pfam HMM files remain available for download and use with HMMER.

**What Data It Contains:** Pfam contains over 19,000 protein family entries (as of 2024), each represented by a seed alignment (manually curated representative sequences), a full alignment (all detected family members), and a profile HMM. Each entry is classified as a Family, Domain, Repeat, Coiled-coil, Disordered, or Motif. Pfam entries include descriptions, literature references, GO term mappings, and clan assignments (grouping related families into superfamilies).

**Main question it helps answer:** What protein family or domain does this sequence belong to, based on sensitive profile HMM matching?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- What Pfam domains are present in this novel protein sequence?
- How many proteins in the human proteome contain the Pfam kinase domain (PF00069)?
- What is the clan membership of this domain, and what other families are evolutionarily related to it?

**Example use cases:**

- Running HMMER with Pfam HMMs to annotate all proteins in a newly sequenced genome.
- Using Pfam domain annotations to group proteins into functional categories for enrichment analysis.
- Downloading the Pfam HMM library to build a custom domain annotation pipeline.

**Input Data Accepted:** Protein sequences in FASTA format analyzed through InterPro or HMMER, Pfam accession numbers (PF format), and UniProt accession numbers.

**Output Data Provided:** Outputs include domain annotations with positions, E-values, and bit scores; Pfam entry descriptions, multiple sequence alignments, and HMM models; Gene Ontology (GO) term mappings; and clan assignments that group related domain families for evolutionary interpretation. Pfam HMM libraries can also be downloaded for use in custom annotation pipelines.

**Strengths:** Pfam is one of the most widely used protein domain databases and is supported by extensive literature. Its profile Hidden Markov Models (HMMs) are highly sensitive for detecting remote homologs and provide broad

coverage, with more than 19,000 protein families representing the majority of known protein sequences. Pfam HMM libraries are downloadable and compatible with HMMER, making them a standard resource for custom genome annotation workflows. In addition, the clan system enables evolutionary analysis by grouping related protein families.

**Limitations:** The standalone Pfam website was retired in 2022, requiring users to access Pfam data through the InterPro interface. Some Pfam entries define broad domain boundaries that may not precisely match biological domain limits, and HMM-based detection can produce false positives when analyzing highly divergent sequences. Coverage gaps also remain for novel protein families originating from poorly studied organisms.

**Common Beginner Mistakes:** Beginners often attempt to access the retired pfam.xfam.org website instead of using InterPro, neglect to download the Pfam HMM library when constructing custom HMMER-based annotation pipelines, or confuse Pfam accession numbers (PF identifiers) with InterPro accession numbers (IPR identifiers). Another common oversight is failing to use downloadable Pfam HMM profiles when performing large-scale domain annotation.

**When to Use It:** Use Pfam (via InterPro) when you need HMM-based domain annotation with broad coverage and sensitivity for remote homologs. Pfam HMMs are the standard choice for genome annotation pipelines using HMMER.

**When NOT to Use It:** For specific functional motif detection, PROSITE may be more appropriate. For structural family classification, SUPERFAMILY or CATH-Gene3D are better choices.

**Related databases / alternatives:** Related resources include InterPro, which now hosts and integrates Pfam data with other annotation databases; TIGRFAMs/NCBIfam, an HMM-based protein family resource maintained by NCBI; and CDD (Conserved Domain Database), NCBI's database for conserved domain identification and annotation.

**How It Connects to Other Resources:** Pfam is a core member database of InterPro. Pfam annotations are incorporated into UniProt entries, Ensembl gene annotations, and genome annotation pipelines. The Pfam HMM library is used by HMMER, which is integrated into many bioinformatics tools and pipelines.

**API / FTP / programmatic access:**

**InterPro API (for Pfam data):** <https://www.ebi.ac.uk/interpro/api/>; **Pfam HMM library download:**

[https://ftp.ebi.ac.uk/pub/databases/Pfam/current\\_release/Pfam-A.hmm.gz](https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz); **HMMER (command-line):** `hmmsearch --domtblout results.txt Pfam-A.hmm query.faa`

InterProScan also incorporates Pfam scanning as part of its integrated protein annotation workflow.

**Evidence/curation level:** Pfam uses manually curated seed alignments, while HMM profiles are computationally generated from these curated alignments. Gene Ontology (GO) term mappings are manually assigned to support reliable functional interpretation.

**Data Update Status:** Updated with each InterPro release (approximately every 2 months).

**Licensing / access restrictions:** Pfam data available under Creative Commons Zero (CC0) license; freely downloadable and reusable.

**Citation / Recommended Reference:** Mistry J, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412-D419. doi:10.1093/nar/gkaa913

**Beginner-Friendly Explanation:** Pfam is a database of protein "building blocks" called domains. Each entry in Pfam represents a type of domain found in many different proteins, and the database uses mathematical models (called HMMs) to recognize these domains even in proteins that are only distantly related. If you have a protein sequence and want to know what functional parts it contains, Pfam can identify these parts and tell you what they do. Note that the old Pfam website has been retired, and you now access Pfam data through the InterPro website.

**Advanced Technical Explanation:** Pfam uses profile HMMs built with HMMER3 from manually curated seed alignments. Each HMM encodes position-specific amino acid frequencies and gap penalties derived from the seed alignment, enabling sensitive detection of family members with as little as 20-25% sequence identity. The Pfam-A database contains high-quality, manually curated entries, while the discontinued Pfam-B contained automatically generated families. Clan assignments group related Pfam families based on structural and sequence evidence, providing a higher-level classification. The gathering threshold (GA) for each HMM is set to minimize false positives while maintaining sensitivity, and is used as the default cutoff in HMMER searches.

**One Practical Workflow Example:** Annotating a proteome with Pfam domains using HMMER:

Step 1: Download the Pfam HMM library: `wget https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/ Pfam-A.hmm.gz && gunzip Pfam-A.hmm.gz`

Step 2: Press the HMM database for faster searching: `hmmcompress Pfam-A.hmm`

Step 3: Run `hmmsearch` against your protein sequences: `hmmsearch --domtblout pfam_results.txt --cut_ga Pfam-A.hmm proteins.faa > /dev/null`

Step 4: Parse the domain table output (`pfam_results.txt`) to extract domain assignments, E-values, and positions for each protein.

Step 5: Filter results using the gathering threshold (already applied with `--cut_ga`) and remove overlapping domain hits if needed.

Step 6: Map Pfam accessions to GO terms using the `Pfam-A.clans.tsv` file from the InterPro FTP for functional enrichment analysis.

## J2 – InterPro

InterPro functions both as a protein function database and as the primary protein family and domain integration resource. Within the context of protein family and domain analysis (Category J), InterPro is generally considered the recommended starting point because it integrates signatures from multiple major databases, including Pfam, PROSITE, PRINTS, SUPERFAMILY, SMART, TIGRFAMs, PIRSF, HAMAP, CDD, CATH-Gene3D, and several additional member resources into a unified classification framework. This integration provides comprehensive coverage by combining different domain detection and annotation approaches, while InterProScan serves as the standard command-line tool for large-scale proteome annotation.

## J3 – CDD (Conserved Domain Database)

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/cdd>

**Resource Type:** Database

**Main Biological Domain:** Proteins

**What It Is Used For:** The CDD is NCBI's database of conserved protein domains, providing annotations for protein sequences submitted to or searched against NCBI resources. It is used to identify conserved domains in protein sequences, understand the functional architecture of proteins, and access domain annotations integrated into NCBI's protein and structure databases. CDD is particularly convenient for users already working within the NCBI ecosystem, as domain annotations are automatically displayed for protein sequences in NCBI databases.

**What Data It Contains:** CDD contains domain models from multiple sources including NCBI-curated domains, Pfam, SMART, COG (Clusters of Orthologous Groups), KOG, PRK (Protein Clusters), and TIGRFAMs. Each domain model is represented as a position-specific scoring matrix (PSSM) or profile HMM. CDD also includes 3D structure-based domain models derived from MMDB (Molecular Modeling Database) that link domain annotations to structural data.

**Main question it helps answer:** What conserved domains does this protein contain, and how do they relate to known protein structures and functional annotations in NCBI databases?

**Typical user:** Bioinformatician / Researcher (particularly NCBI ecosystem users)

**Example scientific questions:**

- What conserved domains are present in this protein sequence retrieved from NCBI?
- Does this bacterial protein contain a domain that is structurally similar to a known eukaryotic domain?
- What is the functional annotation of this conserved domain based on structural and sequence evidence?

**Example use cases:**

- Running RPS-BLAST (Reverse PSI-BLAST) against CDD to annotate domains in a set of protein sequences.
- Using the CDD web interface to explore the domain architecture of a protein retrieved from NCBI Protein.
- Accessing structure-based domain annotations to understand the 3D context of a conserved domain.

**Input Data Accepted:** Protein sequences in FASTA format, NCBI protein accession numbers, and CDD accession identifiers including cd, pfam, smart, and COG formats.

**Output Data Provided:** Outputs include domain annotations with positions, E-values, and bit scores, domain descriptions with functional annotations, links to NCBI structural resources such as MMDB and PDB, and RPS-BLAST results available in standard BLAST output formats.

**Strengths:** CDD is tightly integrated with NCBI protein and structural databases, allowing seamless navigation between sequence, domain, and structural information. It includes structure-based domain models that connect protein sequences with three-dimensional structures and uses RPS-BLAST, which is computationally efficient for large-scale domain searches. CDD also incorporates COG and KOG classifications, supporting functional and orthologous group assignments for prokaryotic and eukaryotic proteins. These features make CDD particularly convenient for users already working within the NCBI ecosystem.



**Limitations:** Compared with InterPro, CDD provides less comprehensive coverage of eukaryotic protein domains. Position-Specific Scoring Matrix (PSSM)-based models may be less sensitive than profile Hidden Markov Models (HMMs) for detecting remote homologs, and the web interface is generally less polished than InterPro. Some domain models may also be updated less frequently than their Pfam counterparts.

**Common Beginner Mistakes:** Beginners often fail to recognize that CDD integrates multiple source databases, including Pfam and SMART, meaning results may overlap with those from other annotation tools. Another common mistake is relying on the web interface for large-scale analyses instead of using the command-line RPS-BLAST tool, which is more efficient for bulk processing.

**When to Use It:** Use CDD when you are working within the NCBI ecosystem and want domain annotations integrated with NCBI protein and structure data. It is particularly useful for prokaryotic proteins where COG/KOG annotations provide functional context.

**When NOT to Use It:** For comprehensive eukaryotic domain annotation, InterPro/InterProScan provides broader coverage. For the most sensitive remote homolog detection, Pfam HMMs with HMMER are preferable.

**Related databases / alternatives:** Related resources include InterPro, which provides broader integration of protein domain databases; Pfam, an HMM-based domain database now accessed through InterPro; and COG/KOG ortholog group databases, which are integrated into CDD.

**How It Connects to Other Resources:** CDD is integrated into NCBI's protein database, structure database (MMDB), and BLAST services. Domain annotations from CDD appear automatically in NCBI protein records. CDD links to PDB structures for structure-based domain models.

**API / FTP / programmatic access:** CDD can be queried programmatically using RPS-BLAST, for example: `rpblast -query proteins.faa -db CDD -out results.txt -outfmt 6`

The CDD database files are available through the NCBI FTP server: [CDD FTP download](#). Additionally, NCBI E-utilities can be used to retrieve CDD annotations associated with protein accession numbers.

**Evidence/curation level:** Mixed: NCBI-curated domains are manually reviewed; imported Pfam/SMART domains follow their respective curation standards.

**Data Update Status:** Updated periodically; check NCBI CDD release notes for current version.

**Licensing / access restrictions:** Freely available; data downloadable from NCBI FTP.

**Citation / Recommended Reference:** Lu S, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 2020;48(D1):D265-D268. doi:10.1093/nar/gkz991

**Beginner-Friendly Explanation:** The Conserved Domain Database (CDD) is NCBI's tool for identifying the functional building blocks (domains) in protein sequences. When you look up a protein in NCBI's databases, CDD annotations are often shown automatically, telling you what known domains the protein contains. It draws on information from several domain databases at once, including Pfam and SMART, and also connects domain information to 3D protein structures. It is particularly useful if you are already using NCBI tools for your research.

**Advanced Technical Explanation:** CDD uses RPS-BLAST (Reverse PSI-BLAST) to search protein sequences against a database of PSSMs derived from multiple sequence alignments of domain families. The CDD database integrates models from Pfam, SMART, COG, KOG, PRK, TIGRFAMs, and NCBI-curated domains, with each

model stored as a PSSM in the RPS-BLAST database format. Structure-based domain models are derived from MMDB structural alignments and provide 3D context for domain annotations. The SPARCLE (Subfamily Protein Architecture Labeling Engine) system classifies proteins by their domain architecture and assigns functional labels to architecture classes.

#### One Practical Workflow Example: Running batch domain annotation with CDD:

- Step 1: Download the CDD database from NCBI FTP: `wget https://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/little_endian/Cdd_LE.tar.gz && tar -xzf Cdd_LE.tar.gz`
- Step 2: Run RPS-BLAST against your protein sequences: `rpsblast -query proteins.faa -db Cdd -out cdd_results.txt -outfmt "6 qseqid sseqid pident length eval evalue bitscore stitle" -evalue 0.01`
- Step 3: Parse the output to extract domain assignments for each protein.
- Step 4: Map CDD accessions to functional descriptions using the `cddid.tbl` file from the CDD FTP.
- Step 5: For web-based analysis of individual proteins, use the CD-Search tool at <https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>
- Step 6: Cross-reference results with InterPro for additional domain coverage and GO term annotations.

## J4 – PROSITE (in Category J context)

[See full PROSITE card in Category I (Card I5). Key Category J-specific notes:] PROSITE is particularly valuable in the domain/motif context for:

Detecting short functional motifs (phosphorylation sites, glycosylation sites, nuclear localization signals) that are too short for HMM-based detection  
Identifying active site residues and binding site patterns  
Providing interpretable regular expression patterns that can be manually inspected and modified.

## J5 – PRINTS (Protein Fingerprint Database)

**Official Website URL:** <https://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS>

**Resource Type:** Database (legacy; limited updates)

**Main Biological Domain:** Proteins

**What It Is Used For:** PRINTS is a database of protein family fingerprints, where each fingerprint consists of a set of conserved motifs (ungapped multiple sequence alignments) that together characterize a protein family. Unlike single-domain databases, PRINTS uses multiple motifs simultaneously to classify proteins, which can improve specificity. It is now primarily of historical interest and is included in InterPro, but the standalone database has not been actively updated since approximately 2012.

**What Data It Contains:** PRINTS contains approximately 2,000 fingerprint entries, each consisting of 2-20 conserved motifs derived from multiple sequence alignments of protein family members. Each entry includes a description, literature references, and links to Swiss-Prot entries. The database covers a range of protein families with particular depth in receptor families and enzyme classes.

**Main question it helps answer:** Does this protein belong to a specific family based on the presence of multiple conserved sequence motifs characteristic of that family?

**Typical user:** Bioinformatician (primarily through InterPro; standalone use is rare)

**Example scientific questions:**

- Does this sequence match the fingerprint for a specific GPCR subfamily?
- What protein families are characterized by this combination of conserved motifs?

**Example use cases:**

- Accessing PRINTS annotations through InterPro for proteins where PRINTS provides unique family assignments not covered by other databases.
- Historical reference for understanding the fingerprint approach to protein family classification.

**Input data accepted:** Protein sequences (via FingerPRINTScan tool); PRINTS accession numbers (PR format)

**Output data provided:** Fingerprint matches with scores for each motif; Family descriptions and literature references

**Strengths:** Multi-motif approach can be more specific than single-domain models; Integrated into InterPro for continued accessibility; Useful for GPCR and receptor family classification

**Limitations:** Not actively maintained since approximately 2012; limited updates; Smaller coverage than Pfam or InterPro; The standalone FingerPRINTScan tool may have compatibility issues with modern systems; Superseded by more comprehensive and actively maintained databases

**Common beginner mistakes:** Attempting to use the standalone PRINTS database for current research without recognizing its limited update status; Not recognizing that PRINTS data is accessible through InterPro

**When to Use It:** PRINTS is best accessed through InterPro rather than as a standalone resource. It may provide unique family assignments for some protein families, particularly GPCRs, that complement other databases.

**When NOT to Use It:** Do not use PRINTS as a primary domain annotation tool for current research. Use InterPro (which includes PRINTS) for comprehensive annotation.

**Related databases / alternatives:** **InterPro:** integrates PRINTS with actively maintained databases; **Pfam:** actively maintained HMM-based alternative; **GPCRDB:** specialized database for GPCR classification

**How It Connects to Other Resources:** PRINTS is a member database of InterPro. PRINTS entries are cross-referenced to Swiss-Prot and are accessible through the InterPro interface.

**API / FTP / programmatic access:** Accessible through InterPro API using PRINTS accession numbers; Standalone FingerPRINTScan tool available from the PRINTS website; InterProScan includes PRINTS scanning

**Evidence/curation level:** Manually curated fingerprints; limited updates since ~2012.

**Data Update Status:** Limited updates; last major update approximately 2012. Maintained within InterPro but not independently expanded.

**Licensing / access restrictions:** Freely accessible.

**Citation / Recommended Reference:** Attwood TK, et al. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. 2003;31(1):400-2. doi:10.1093/nar/gkg030

**Beginner-Friendly Explanation:** PRINTS is an older protein family database that uses a unique approach: instead of looking for a single domain, it looks for a set of several short sequence patterns that together identify a protein family. Think of it like identifying a bird species by looking for multiple features simultaneously (beak shape, wing pattern, tail length) rather than just one. While this approach can be very specific, the database has not been actively updated since around 2012, so it is now mainly useful as part of the InterPro integrated database rather than as a standalone tool.

**Advanced Technical Explanation:** PRINTS fingerprints are sets of ungapped motifs derived from iterative database scanning using the FingerPRINTScan algorithm, which scores sequences by the number and quality of motif matches. The multi-motif approach provides higher specificity than single-motif methods because false positives matching one motif are unlikely to match all motifs in the fingerprint. However, the lack of updates means that many recently characterized protein families are not represented, and the database is now primarily valuable for the subset of families it covers well.

**One Practical Workflow Example:** Accessing PRINTS data through InterPro:

Step 1: Go to <https://www.ebi.ac.uk/interpro> and search for a protein family of interest (e.g., "GPCR rhodopsin").

Step 2: In the results, filter by "Database: PRINTS" to see PRINTS-specific entries.

Step 3: Click on a PRINTS entry (PR accession) to view the fingerprint description and member proteins.

Step 4: Use InterProScan to scan your sequences against all member databases including PRINTS simultaneously.

Step 5: In InterProScan output, look for entries with "PRINTS" in the database column.

## J6: SUPERFAMILY

**Official Website URL:** <https://supfam.org>

**Resource Type:** Database

**Main Biological Domain:** Proteins

**What It Is Used For:** SUPERFAMILY is a database of structural and functional annotations for all proteins and genomes, based on the SCOP (Structural Classification of Proteins) superfamily classification. It uses profile HMMs derived from structural alignments to assign protein sequences to SCOP superfamilies, providing a structure-based classification that can detect remote homologs beyond the reach of sequence-based methods. Researchers use SUPERFAMILY to assign proteins to structural superfamilies, understand the structural evolution of protein families, and annotate proteomes with structure-based functional predictions.

**What Data It Contains:** SUPERFAMILY contains HMM-based models for all SCOP superfamilies, with pre-computed assignments for proteins in hundreds of completely sequenced genomes. Each superfamily entry includes a description, representative structures, and links to SCOP and PDB. The database provides genome-level statistics on superfamily distributions, enabling comparative genomics analyses of structural repertoires.

**Main question it helps answer:** What structural superfamily does this protein belong to, and what does this imply about its evolutionary origin and potential function?

**Typical user:** Bioinformatician / Researcher (structural bioinformatics, evolutionary genomics)

### Example scientific questions:

- What SCOP superfamily does this protein with no detectable sequence homologs belong to?
- How does the structural domain repertoire of this newly sequenced organism compare to related species?
- What is the evolutionary distribution of this structural superfamily across the tree of life?

### Example use cases:

- Assigning a protein with low sequence identity to known proteins to a structural superfamily using SUPERFAMILY HMMs.
- Comparing the structural domain repertoires of multiple genomes to identify lineage-specific expansions or losses.
- Using SUPERFAMILY assignments to infer function for proteins with no sequence-based annotations.

### Input data accepted:

- Protein sequences in FASTA format
- SCOP superfamily identifiers
- Genome identifiers for pre-computed assignments

### Output data provided

- SCOP superfamily assignments with E-values
- Links to representative PDB structures
- Genome-level superfamily distribution statistics
- Downloadable assignment tables for sequenced genomes

**Strengths:**

- Structure-based classification detects remote homologs beyond sequence methods
- Pre-computed assignments for hundreds of genomes
- Genome-level comparative analysis tools
- Links to PDB structures for structural context

**Limitations:**

- Limited to proteins with structural homologs in SCOP; novel folds are not classified
- Update frequency depends on SCOP updates, which have been irregular
- Less comprehensive than Pfam for sequence-based domain annotation
- Web interface is less polished than InterPro

**Common beginner mistakes:**

- Expecting SUPERFAMILY to classify all proteins (it can only classify those with structural homologs in SCOP)
- Confusing SCOP superfamily classification with Pfam domain classification

**When to Use It:** Use SUPERFAMILY when you need structure-based protein classification, particularly for proteins with low sequence identity to characterized proteins, or when performing comparative genomics analyses of structural domain repertoires.

**When NOT to Use It:** For routine domain annotation of well-characterized protein families, Pfam/InterPro is more comprehensive and easier to use. SUPERFAMILY is most valuable for structural bioinformatics applications.

**Related databases / alternatives:**

- SCOP/SCOPE: structural classification database
- CATH: alternative structural classification (integrated into InterPro as CATH-Gene3D)
- InterPro: integrates SUPERFAMILY with other databases

**How It Connects to Other Resources:** SUPERFAMILY is a member database of InterPro and its assignments are included in InterProScan output. SUPERFAMILY links to SCOP and PDB for structural data, and to UniProt for sequence data.

**API / FTP / programmatic access:**

- Web interface at <https://supfam.org> for individual sequence analysis
- FTP: <https://supfam.org/SUPERFAMILY/downloads.html>
- InterProScan includes SUPERFAMILY scanning
- Downloadable HMM library for use with HMMER

**Evidence/curation level:** Computationally derived from SCOP structural classifications; SCOP classifications are manually curated.

**Data Update Status:** Updates depend on SCOP releases; check the SUPERFAMILY website for current version information.

**Licensing / access restrictions:** Freely accessible for academic use.

**Citation / Recommended Reference:** Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol.* 2001;313(4):903-19. doi:10.1006/jmbi.2001.5080

**Beginner-Friendly Explanation:** SUPERFAMILY classifies proteins based on their three-dimensional structure rather than just their sequence. This is useful because proteins that have evolved from a common ancestor can look very different in sequence but still have similar shapes. SUPERFAMILY uses mathematical models built from known protein structures to recognize these structural similarities, even when sequence similarity is too low for other methods to detect. It is particularly useful for proteins that cannot be classified by sequence-based methods alone.

**Advanced Technical Explanation:** SUPERFAMILY uses profile HMMs built from structural alignments of SCOP superfamily members, enabling detection of structural homologs at sequence identities below the "twilight zone" (~20-30%). The HMMs are constructed using SAM (Sequence Alignment and Modeling) software and searched against query sequences using a two-pass procedure: first a fast database scan, then a more sensitive alignment of candidate hits. Assignments are made at the SCOP superfamily level (grouping families with common evolutionary origin) rather than the family level (grouping proteins with clear sequence similarity). Pre-computed assignments for sequenced genomes are stored in a relational database and accessible through the web interface and FTP.

**One Practical Workflow Example: Assigning structural superfamilies to a set of protein sequences:**

Step 1: Go to <https://supfam.org> and use the "Assign" tool to submit protein sequences in FASTA format.

Step 2: Alternatively, run InterProScan with the SUPERFAMILY database included (it is included by default).

Step 3: Review the SUPERFAMILY assignments in the InterProScan output (look for entries with "SUPERFAMILY" in the database column).

Step 4: For each assigned superfamily, click the SCOP link to view the structural classification and representative structures.

Step 5: Use the SUPERFAMILY genome browser to compare superfamily distributions across related organisms.

Step 6: For proteins with no SUPERFAMILY assignment, use Pfam/InterPro for sequence-based domain annotation.



## BEGINNER EXAMPLE (Category J)

---

A researcher identifies a novel bacterial protein with no obvious sequence homologs. They run InterProScan on the sequence and find a Pfam domain (PF00069, protein kinase domain) and a SUPERFAMILY assignment to the protein kinase superfamily. They then check PROSITE for the kinase activation loop pattern (PS00107) and find a match, confirming the kinase assignment. They use the InterPro entry for the kinase domain to find GO terms and pathway associations.

## ADVANCED EXAMPLE (Category J)

---

A bioinformatician is performing a comparative genomics study of domain evolution across 50 fungal genomes. They run InterProScan on all predicted proteomes, extract Pfam domain assignments, and build a domain presence/absence matrix. They use SUPERFAMILY assignments to identify proteins with structural homologs in other kingdoms despite low sequence identity. They then perform phylogenetic profiling to identify domain families that are lineage-specific or show evidence of horizontal gene transfer.

## CONFUSION POINTS

---

**Pfam website retirement:** The standalone pfam.xfam.org website was retired in 2022. Pfam data is now at InterPro. Pfam accessions (PF numbers) still work when searched in InterPro.

**InterPro vs. member databases:** InterPro integrates Pfam, PROSITE, PRINTS, SMART, SUPERFAMILY, and others. An InterPro entry (IPR number) may correspond to multiple member database entries. Use InterPro for integrated results; use member databases directly for specific approaches.

**Domain vs. family vs. superfamily:** In InterPro/Pfam, a "domain" is a structural/functional unit that can occur in multiple protein contexts; a "family" groups proteins with common function and ancestry; a "superfamily" (in SCOP/SUPERFAMILY) groups families with common structural fold.

## DECISION GUIDE (Category J):

---

**For comprehensive domain annotation:** InterPro/InterProScan (integrates all)

**For HMM-based domain detection with HMMER:** Pfam HMM library

**For NCBI-integrated domain annotation:** CDD For specific functional motifs and active sites: PROSITE

**For structure-based classification:** SUPERFAMILY

**For signaling domain analysis:** SMART For legacy GPCR/receptor family classification: PRINTS (via InterPro)

## Category K: Protein Structure Databases

### CATEGORY OVERVIEW

Protein structure databases store three-dimensional coordinate data for biological macromolecules, primarily proteins and nucleic acids, determined by experimental methods such as X-ray crystallography, cryo-electron microscopy (cryo-EM), and nuclear magnetic resonance (NMR) spectroscopy. The three-dimensional structure of a protein is intimately linked to its function: the precise arrangement of atoms in the active site determines catalytic specificity, the shape of binding interfaces governs protein-protein and protein-ligand interactions, and structural changes underlie allosteric regulation and conformational signaling. Access to structural data is therefore essential for rational drug design, understanding disease mechanisms at the molecular level, and interpreting the functional consequences of mutations.

The Protein Data Bank (PDB) is the single global archive for experimentally determined macromolecular structures, maintained jointly by three partner organizations: RCSB PDB (USA), PDBe (Europe), and PDBj (Japan). All three partners hold identical copies of the PDB archive and provide different interfaces and tools for accessing and analyzing the data. This means that a researcher can use any of the three portals to access the same structural data, choosing based on the tools and interface they prefer. The PDB archive has grown from a handful of structures in the 1970s to over 220,000 structures as of 2024, driven by advances in crystallography, cryo-EM, and computational structure determination.

A transformative development in this category was the release of the AlphaFold Protein Structure Database in 2021, which provides predicted three-dimensional structures for virtually all proteins in UniProt — over 200 million structures — generated by the AlphaFold2 deep learning system. This has fundamentally changed the landscape of structural biology by making structural predictions available for proteins that have never been experimentally characterized. However, it is critical to understand that AlphaFold predictions are computational models, not experimental structures, and their accuracy varies significantly depending on the protein and region. The per-residue confidence score (pLDDT) provides a useful guide to prediction reliability, but predicted structures should always be interpreted with appropriate caution and validated against experimental data where possible.

## K1: RCSB PDB (Protein Data Bank)

**Database Name:** RCSB Protein Data Bank (RCSB PDB)

**Official Website URL:** <https://www.rcsb.org>

**Resource Type:** Repository / Database

**Main Biological Domain:** Structures

**What It Is Used For:** The RCSB PDB is the US partner of the worldwide Protein Data Bank, providing access to the global archive of experimentally determined three-dimensional structures of biological macromolecules. Researchers use it to retrieve structural coordinates for proteins and nucleic acids, visualize molecular structures, analyze structural features, and access associated experimental data and validation reports. It is the primary resource for structural biologists, computational chemists, and drug discovery researchers.

**What Data It Contains:** The RCSB PDB contains over 220,000 structures (as of 2024) determined by X-ray crystallography, cryo-EM, NMR, and other methods. Each entry includes atomic coordinates in PDB or mmCIF format, experimental data (structure factors for crystallography, restraints for NMR), validation reports, sequence information, ligand descriptions, and biological assembly information. The database also provides pre-computed structural analyses including secondary structure assignments, domain annotations, and evolutionary conservation data.

**Main question it helps answer:** What is the three-dimensional structure of this protein, and what does it reveal about its function, binding sites, and interactions?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist (structural biologist, biochemist, drug discovery scientist)

**Example scientific questions:**

- What is the structure of the SARS-CoV-2 spike protein in complex with the ACE2 receptor?
- What ligands have been co-crystallized with this enzyme, and what are their binding modes?
- How does the structure of this mutant protein differ from the wild-type?

**Example use cases:** Downloading the PDB file for a protein of interest to perform molecular docking simulations for drug discovery; Using the RCSB PDB search interface to find all structures of a protein family and analyze structural conservation; Retrieving the biological assembly coordinates for a protein complex to understand its quaternary structure.

**Input Data Accepted:** PDB accession codes consisting of four-character alphanumeric identifiers (e.g., 1TUP), search queries based on protein name, organism, experimental method, resolution, or ligand, sequence queries using BLAST or FASTA searches against PDB sequences, and structure-based queries for shape similarity searches.

**Output Data Provided:** Outputs include atomic coordinates in PDB (.pdb) or mmCIF (.cif) formats, experimental data files such as structure factors and NMR restraints, validation reports containing MolProbity scores, clashscores, and Ramachandran plots, biological assembly files, sequence and structural annotations, and interactive three-dimensional visualization through the Mol\* viewer.

**Strengths:** The Protein Data Bank (PDB) serves as the authoritative global archive for experimentally determined protein structures. It provides comprehensive validation reports for assessing structure quality, powerful search interfaces with extensive filtering options, and advanced visualization tools such as Mol\*. The resource also offers rich APIs for programmatic access and integrates structural information with sequence databases, drug resources, and scientific literature.

**Limitations:** PDB coverage is biased toward proteins that are experimentally tractable, particularly crystallizable and stable proteins, resulting in underrepresentation of membrane proteins, intrinsically disordered proteins, and very large complexes. Structure quality varies substantially among entries, particularly older structures that may contain modeling or refinement errors. Biological assembly information can sometimes be ambiguous, ligand annotations may vary in quality, and not all deposited structures include associated experimental datasets.

**Common Beginner Mistakes:** A common mistake is using asymmetric unit coordinates instead of biological assembly coordinates, since the asymmetric unit may represent only part of the biologically functional complex. Users also frequently neglect to evaluate resolution and validation metrics before using a structure, confuse legacy PDB (.pdb) files with modern mmCIF (.cif) formats preferred for large structures, or fail to account for missing residues corresponding to disordered regions not resolved experimentally.

**When to Use It:** Use RCSB PDB when you need experimentally determined three-dimensional structural data for a protein or nucleic acid. It is the essential resource for structural biology, molecular docking, structure-based drug design, and any analysis requiring atomic-level structural information.

**When NOT to Use It:** If no experimental structure exists for your protein of interest, use AlphaFold or SWISS-MODEL for predicted structures. For sequence-based functional annotation without structural analysis, UniProt is more appropriate.

**Related databases / alternatives:** Related resources include PDBe, the European partner of the Protein Data Bank that provides the same structural data through a different interface and specialized analysis tools; PDBj, the Japanese PDB partner offering alternative interfaces and services for accessing the same core dataset; AlphaFold DB, which provides predicted protein structures for proteins lacking experimentally determined models; the SWISS-MODEL Repository, which hosts homology-based structural models for proteins without available experimental structures; and EMDB (Electron Microscopy Data Bank), which archives raw cryo-electron microscopy density maps associated with structural studies.

**How It Connects to Other Resources:** RCSB PDB links to UniProt for sequence annotations, PubMed for associated publications, ChEMBL and DrugBank for ligand information, and Ensembl for genomic context. The PDB archive is synchronized with PDBe and PDBj through the wwPDB partnership. RCSB PDB also provides links to AlphaFold predictions for proteins with PDB structures, enabling comparison of experimental and predicted structures.

**API / FTP / programmatic access:** REST API: <https://data.rcsb.org/rest/v1/> - Example: <https://data.rcsb.org/rest/v1/core/entry/1TUP>; GraphQL API: <https://data.rcsb.org/graphql>; FTP: <https://files.rcsb.org/pub/pdb/>; PDB file download: <https://files.rcsb.org/download/1TUP.pdb>; mmCIF download: <https://files.rcsb.org/download/1TUP.cif>; Python: pypdb, rcsbsearchapi packages; Biopython: Bio.PDB module for structure parsing and analysis

**Evidence/curation level:** Experimentally determined structures; each entry includes a validation report assessing structure quality. Structures are not independently re-determined but are validated against experimental data.

**Data update status:** New structures added weekly; the archive is updated every Wednesday with newly released entries.

**Licensing / access restrictions:** Fully open access; all PDB data is freely available with no restrictions on use or redistribution (CC0 license for data).

**Citation / recommended reference:** Berman HM, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28(1):235-42. doi:10.1093/nar/28.1.235 (For RCSB PDB specifically: Burley SK, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules. *Nucleic Acids Res.* 2021;49(D1):D437-D451. doi:10.1093/nar/gkaa1038)

**Beginner-Friendly Explanation:** The RCSB Protein Data Bank is the world's archive of protein shapes. Just as a blueprint shows the exact layout of a building, a PDB structure shows the exact positions of every atom in a protein. Scientists use X-ray crystallography, cryo-electron microscopy, and other techniques to determine these structures, and then deposit them in the PDB for everyone to use. You can search for the structure of almost any well-studied protein, view it in 3D in your web browser, and download the coordinate file to use in your own analysis. Over 220,000 structures are available as of 2024.

**Advanced Technical Explanation:** PDB entries are stored in mmCIF format (the current standard) and legacy PDB format, with atomic coordinates, anisotropic displacement parameters, occupancy values, and alternate conformations for each atom. Biological assembly information specifies the symmetry operations needed to generate the functional oligomeric state from the asymmetric unit. Validation reports use MolProbity metrics (Ramachandran outliers, rotamer outliers, clashscore) and real-space correlation coefficients to assess model quality against experimental data. The RCSB PDB search API supports complex queries combining sequence similarity, structure similarity, chemical component properties, and metadata filters using a JSON query language.

**One practical workflow example:** Retrieving and analyzing a protein structure for molecular docking:

Step 1: Search RCSB PDB (<https://www.rcsb.org>) for your target protein using the protein name or UniProt accession. Filter by resolution (e.g., < 2.5 Angstrom) and method (X-ray crystallography).

Step 2: Select the highest-quality structure (lowest resolution value, best validation scores). Check the validation report for Ramachandran outliers and clashscore.

Step 3: Download the biological assembly file in PDB or mmCIF format.

Step 4: Prepare the structure for docking: remove water molecules and non-essential ligands, add hydrogen atoms, and assign charges using tools like AutoDockTools or Schrodinger Protein Preparation Wizard.

Step 5: Identify the binding site using the ligand position in the co-crystal structure or using cavity detection tools (fpocket, SiteMap).

Step 6: Perform docking with AutoDock Vina, Glide, or similar tools, using the prepared structure and defined binding site.

## K2 – PDBe (Protein Data Bank in Europe)

**Official Website URL:** <https://www.ebi.ac.uk/pdbe>

**Resource Type:** Repository / Database

**Main Biological Domain:** Structures

**What It Is Used For:** PDBe is the European partner of the worldwide Protein Data Bank, maintained by EMBL-EBI. It provides access to the same PDB archive as RCSB PDB and PDBj, but with a distinct set of tools and interfaces developed at EBI. PDBe is particularly known for its PDBe-KB (Knowledge Base) resource, which aggregates structural and functional annotations from multiple sources for each UniProt protein, and for its strong API that is widely used in programmatic structural bioinformatics workflows.

**What Data It Contains:** PDBe holds the complete PDB archive (identical to RCSB PDB and PDBj) plus additional value-added resources: PDBe-KB aggregates annotations from FunPDBe partner resources (binding sites, conservation, disease variants, etc.); PDBe-REDO provides re-refined and re-built versions of PDB structures with improved geometry; and EMDB (Electron Microscopy Data Bank) stores raw cryo-EM density maps associated with PDB entries.

**Main question it helps answer:** What structural data and integrated structural annotations are available for this protein, and how can I access them programmatically?

**Typical user:** Bioinformatician / Researcher (structural bioinformatics)

**Example scientific questions:**

- What is the aggregated structural and functional annotation for this UniProt protein across all available PDB structures?
- What cryo-EM maps are available for this large protein complex?
- How can I programmatically retrieve all PDB structures for a given UniProt accession?

**Example use cases:**

- Using the PDBe API to retrieve all PDB structures for a protein of interest and their associated metadata.
- Accessing PDBe-KB to view aggregated annotations (binding sites, disease variants, conservation) mapped onto the protein structure.
- Downloading re-refined structures from PDBe-REDO for improved geometry in computational analyses.

**Input Data Accepted:** PDBe accepts PDB accession codes, UniProt accession numbers, search queries through the PDBe interface, and sequence-based queries.

**Output Data Provided:** Outputs include PDB and mmCIF coordinate files identical to those in RCSB PDB, aggregated annotations through PDBe-KB, re-refined structures from PDBe-REDO, cryo-EM maps from EMDB, and API responses in JSON format.

**Strengths:** PDBe provides several value-added services beyond the core PDB archive. PDBe-KB offers integrated structural annotations aggregated from multiple partner resources, while PDBe-REDO supplies re-refined versions of PDB structures with improved geometry and model quality. The platform has a robust REST API widely used in bioinformatics workflows and integrates closely with EMDB for cryo-electron microscopy data. As part of the EMBL-EBI ecosystem, PDBe links directly to resources such as UniProt, InterPro, and related biological databases.



**Limitations:** PDBe contains the same experimentally determined structures as RCSB PDB and PDBj and does not host unique structural entries. PDBe-KB annotations depend on the update cycles of contributing partner databases, and some users may find the interface less intuitive than RCSB PDB for routine browsing and visualization.

**Common Beginner Mistakes:** Beginners often fail to recognize that PDBe, RCSB PDB, and PDBj share the same underlying structural archive and differ primarily in interfaces and tools. Another common oversight is not using PDBe-KB when integrated structural annotations and functional context are needed.

**When to Use It:** PDBe is particularly useful when programmatic access through a well-documented API is required, when aggregated structural annotations from PDBe-KB are needed, or when working with EMDB cryo-EM datasets and related structural resources.

**When NOT to Use It:** For straightforward structure browsing and visualization, RCSB PDB may provide a more intuitive user experience. In practice, choosing between PDBe, RCSB PDB, and PDBj is often based on personal preference and workflow compatibility rather than data availability.

**Related databases / alternatives:** Related resources include RCSB PDB and PDBj, which provide the same structural data through alternative interfaces and analysis tools; EMDB, which archives cryo-EM density maps and is closely integrated with PDBe; and PDBe-REDO, which provides automatically re-refined structural models.

**How It Connects to Other Resources:** PDBe is part of the EMBL-EBI ecosystem and links to UniProt, InterPro, Ensembl, and ChEMBL. PDBe-KB aggregates annotations from FunPDBe partner resources including P2rank (binding sites), ConSurf (conservation), and others. EMDB is co-managed by PDBe and RCSB.

#### API / FTP / programmatic access:

- **PDBe REST API:** <https://www.ebi.ac.uk/pdbe/api/>
- **Example:** <https://www.ebi.ac.uk/pdbe/api/pdb/entry/summary/1TUP>
- **PDBe-KB API:** <https://www.ebi.ac.uk/pdbe/graph-api/>
- **FTP:** <https://ftp.ebi.ac.uk/pub/databases/pdb/>
- **EMDB API:** <https://www.ebi.ac.uk/emdb/api/>

**Evidence/curation level:** Experimentally determined structures (same as PDB); PDBe-KB annotations from partner resources with varying evidence levels.

**Data Update Status:** Updated weekly in synchronization with the wwPDB release cycle.

**Licensing / access restrictions:** Fully open access; CC0 license for PDB data.

**Citation / Recommended Reference:** Armstrong DR, et al. PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* 2020;48(D1):D335-D343. doi:10.1093/nar/gkz990

**Beginner-Friendly Explanation:** PDBe is the European version of the Protein Data Bank — it contains exactly the same protein structures as the American RCSB PDB, but with different tools and features. One of PDBe's special features is PDBe-KB, which collects information about a protein from many different sources and shows it all together on the protein structure, making it easy to see where important sites are located. PDBe also has a powerful programming interface (API) that makes it easy to automatically retrieve structural data for large numbers of proteins.



**Advanced Technical Explanation:** PDBe maintains the European mirror of the wwPDB archive and provides additional value-added services. The PDBe REST API follows a consistent URL scheme for retrieving entry summaries, molecules, ligands, binding sites, secondary structure, and other structural features in JSON format. PDBe-KB uses a graph-based data model to aggregate annotations from FunPDBe partner resources, mapping them to UniProt residue positions for cross-structure comparison. PDBe-REDO applies automated re-refinement and model rebuilding to PDB structures using the latest refinement software and restraint libraries, often improving geometry and fit to experimental data.

**One Practical Workflow Example: Using PDBe API to retrieve all structures for a UniProt protein:**

Step 1: Query the PDBe API for all PDB entries mapped to a UniProt accession: `curl "https://www.ebi.ac.uk/pdbe/api/mappings/best_structures/P04637" | python -m json.tool`

Step 2: Parse the JSON response to extract PDB IDs, chain IDs, resolution, and coverage information.

Step 3: For each PDB entry, retrieve the structure summary: `curl "https://www.ebi.ac.uk/pdbe/api/pdb/entry/summary/1TUP"`

Step 4: Download the mmCIF files for selected structures: `wget https://files.ebi.ac.uk/pdbe/entry-files/download/1tup.cif`

Step 5: Access PDBe-KB for aggregated annotations: `https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P04637`

Step 6: Use the annotations to guide structural analysis and interpretation.

## K3 – PDBj (Protein Data Bank Japan)

**Official Website URL:** <https://pdj.org>

**Resource Type:** Repository / Database

**Main Biological Domain:** Structures

**What It Is Used For:** PDBj is the Japanese partner of the worldwide Protein Data Bank, maintained by the Institute for Protein Research at Osaka University. It provides access to the same PDB archive as RCSB PDB and PDBe, with tools and interfaces developed specifically by the Japanese structural biology community. PDBj offers unique tools including eF-site (electrostatic surface analysis), Promode Elastic (normal mode analysis), and Mine2 (structural similarity search), as well as Japanese-language support.

**What Data It Contains:** PDBj holds the complete PDB archive (identical to RCSB PDB and PDBe) plus value-added resources including structural analysis tools, a structural similarity search engine (Mine2), and links to Japanese structural biology resources. PDBj also maintains the BMRB (Biological Magnetic Resonance Data Bank) mirror for NMR data.

**Main question it helps answer:** What structural data is available for this protein, and what specialized structural analyses (electrostatics, normal modes, structural similarity) can be performed?

**Typical user:** Researcher / Bioinformatician (particularly in Japan and Asia-Pacific)

**Example scientific questions:** What is the electrostatic surface potential of this protein, and how does it relate to its binding properties? | What PDB structures are most similar to my query structure? | What NMR data are available for this protein?

**Example use cases:** Using PDBj's Mine2 tool to find structurally similar proteins to a query structure; Accessing eF-site for electrostatic surface analysis of a protein; Retrieving NMR restraint data from the BMRB mirror.

**Input Data Accepted:** PDBj accepts PDB accession codes, search queries through the PDBj interface, and uploaded structure files for structural similarity searches.

**Output Data Provided:** Outputs include PDB and mmCIF coordinate files identical to those available from other wwPDB partners, electrostatic surface analyses through eF-site, normal mode analysis results using Promode Elastic, and structural similarity search results generated by the Mine2 platform.

**Strengths:** PDBj offers several specialized structural analysis tools not typically available through other PDB partners, including electrostatic surface analysis, normal mode analysis, and structural similarity searching. It provides Japanese-language support and documentation, hosts a mirror of the Biological Magnetic Resonance Data Bank (BMRB) for NMR-related data, and participates fully in the worldwide Protein Data Bank (wwPDB) partnership, ensuring synchronized and consistent structural datasets.

**Limitations:** PDBj contains the same experimentally determined structures available through RCSB PDB and PDBe and therefore does not provide unique structural entries. Some of its specialized tools may have more limited English-language documentation, and the platform is generally less familiar to researchers outside Japan.

**Common Beginner Mistakes:** A common misunderstanding is failing to recognize that PDBj, RCSB PDB, and PDBe host the same structural archive. Another frequent oversight is ignoring PDBj's specialized tools, which may

provide useful analyses such as electrostatic mapping or normal mode modeling unavailable through other interfaces.

**When to Use It:** Use PDBj when you need its unique structural analysis tools (electrostatics, normal modes, structural similarity), or when you need NMR data from the BMRB mirror.

**When NOT to Use It:** For general structure browsing and download, RCSB PDB or PDBe are equally suitable and may be more familiar. The choice is largely based on which tools are needed.

**Related databases / alternatives:** Related resources include RCSB PDB and PDBe, which host the same wwPDB structural archive through different interfaces and tools, and BMRB, the Biological Magnetic Resonance Data Bank, which is mirrored through PDBj for NMR data access.

**How It Connects to Other Resources:** PDBj is part of the wwPDB partnership and synchronizes data with RCSB PDB and PDBe weekly. PDBj links to UniProt, PubMed, and Japanese structural biology resources.

**API / FTP / programmatic access:** PDBj offers REST-based programmatic access through the PDBj API [PDBj REST API](#) and provides structural data downloads through its FTP archive [PDBj FTP archive](#). Structural similarity analysis through Mine2 is available at [PDBj Mine2 structural similarity search](#).

**Evidence/curation level:** Experimentally determined structures (same as PDB).

**Data Update Status:** Updated weekly in synchronization with the wwPDB release cycle.

**Licensing / access restrictions:** Fully open access; CC0 license for PDB data.

**Citation / Recommended Reference:** Kinjo AR, et al. Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures.

**Beginner-Friendly Explanation:** PDBj is Japan's version of the Protein Data Bank, containing the same protein structures as the American and European versions. What makes PDBj special are its unique analysis tools: you can calculate the electrical charge distribution on a protein's surface, analyze how a protein might move and flex, and search for proteins with similar three-dimensional shapes. For most users, the choice between PDBj, RCSB PDB, and PDBe comes down to which tools you need for your specific analysis.

**Advanced Technical Explanation:** PDBj maintains the Japanese mirror of the wwPDB archive with weekly synchronization. The Mine2 structural similarity search uses a graph-based algorithm to compare secondary structure element arrangements, enabling fast all-against-all structural comparison. The eF-site tool calculates electrostatic potentials using the Poisson-Boltzmann equation and maps them onto the molecular surface. Promode Elastic implements elastic network model normal mode analysis for predicting protein conformational dynamics without molecular dynamics simulation.

**One Practical Workflow Example: Using PDBj for structural similarity search:**

Step 1: Go to <https://pdbj.org/mine2> and enter a PDB accession code or upload a structure file.

Step 2: Select the search parameters (chain, similarity threshold).

Step 3: Review the list of structurally similar proteins with similarity scores and structural alignment visualizations.

Step 4: Download the alignment results for further analysis.

Step 5: Use the eF-site tool (<https://pdbj.org/eF-site>) to analyze the electrostatic surface of your protein of interest.

## K4 – AlphaFold Protein Structure Database

**Database Name:** AlphaFold Protein Structure Database

**Official Website URL:** <https://alphafold.ebi.ac.uk>

**Resource Type:** Database

**Main Biological Domain:** Structures

**What It Is Used For:** The AlphaFold Protein Structure Database provides predicted three-dimensional structures for virtually all proteins in UniProt, generated by the AlphaFold2 deep learning system developed by DeepMind and EMBL-EBI. It is used to obtain structural models for proteins that have no experimentally determined structure, to guide experimental structure determination, to predict the structural consequences of mutations, and to perform structure-based functional annotation at proteome scale. The database has transformed structural biology by making structural predictions available for the vast majority of known proteins.

**What Data It Contains:** The AlphaFold DB contains predicted structures for over 200 million proteins from UniProt, including complete proteomes for hundreds of organisms. Each entry provides a PDB-format coordinate file with per-residue confidence scores (pLDDT, ranging from 0-100), a predicted aligned error (PAE) matrix indicating confidence in relative domain positions, and links to the corresponding UniProt entry. Structures are available for individual proteins and as bulk downloads for complete proteomes.

**Main question it helps answer:** What is the predicted three-dimensional structure of this protein, and how confident is the prediction for each region?

**Typical user:** Researcher / Bioinformatician / Wet-lab scientist

**Example scientific questions:**

- What does the predicted structure of this uncharacterized protein suggest about its function?
- Which regions of this protein are predicted with high confidence, and which are likely disordered?
- How does the predicted structure of this protein compare to its closest structural homolog in the PDB?

**Example use cases:**

- Using an AlphaFold structure as a starting model for molecular replacement in X-ray crystallography.
- Predicting the structural impact of a disease-associated mutation by comparing wild-type and mutant AlphaFold models.
- Downloading all AlphaFold structures for a proteome to perform large-scale structure-based functional annotation.

**Input data accepted:** UniProt accession numbers; Protein names or gene names (via search); Organism names for proteome-level access

**Output Data Provided:** Predicted structure in PDB format with pLDDT scores in the B-factor column; Predicted aligned error (PAE) matrix in JSON format; Confidence score visualization; Links to UniProt entry and related PDB structures

**Strengths:**

- Unprecedented coverage: predicted structures for >200 million proteins



- High accuracy for single-domain proteins and well-folded regions
- pLDDT confidence scores provide per-residue reliability estimates
- PAE matrix enables assessment of inter-domain orientation confidence
- Freely available for all proteins in UniProt
- Bulk download available for complete proteomes

**Limitations:** Predictions are computational models, not experimental structures; must be interpreted with appropriate caution; Accuracy is lower for intrinsically disordered regions, novel folds, and proteins requiring cofactors or binding partners for folding; Does not predict protein complexes (AlphaFold-Multimer is separate); Low pLDDT regions (< 50) are likely disordered and should not be interpreted as structured; Cannot predict conformational changes or multiple functional states; Predictions may not reflect the biologically relevant conformation

**Common beginner mistakes:** Treating AlphaFold predictions as equivalent to experimental structures; Not checking pLDDT scores before interpreting structural features; Interpreting low-confidence (low pLDDT) regions as structured; Using AlphaFold structures for drug docking without experimental validation of the binding site; Not recognizing that the B-factor column contains pLDDT scores, not actual B-factors

**When to Use It:** Use AlphaFold DB when no experimental structure exists for your protein of interest and you need a structural model for hypothesis generation, as a starting model for experimental structure determination, or for large-scale structural annotation. It is particularly valuable for proteins from non-model organisms with no structural data.

**When NOT to Use It:** Do not use AlphaFold structures as the sole basis for drug design without experimental validation. For proteins with existing high-quality experimental structures, use the PDB. Do not interpret low-pLDDT regions as structured.

**Related databases / alternatives:** RCSB PDB: experimental structures; use when available; SWISS-MODEL Repository: homology models; useful when a close template exists; ESMFold: alternative AI structure prediction (Meta AI); RoseTTAFold: alternative AI structure prediction

**How it connects to other resources:** AlphaFold DB is maintained by EMBL-EBI and links to UniProt entries for each predicted structure. RCSB PDB and PDBe now display AlphaFold predictions alongside experimental structures for proteins in UniProt. AlphaFold structures are used by InterPro and other databases for structure-based functional annotation.

**API / FTP / programmatic access:** REST API: <https://alphafold.ebi.ac.uk/api/>; Example: <https://alphafold.ebi.ac.uk/api/prediction/P04637/>; FTP bulk download: <https://ftp.ebi.ac.uk/pub/databases/alphafold/>; Python: requests library with REST API; Bulk proteome downloads available as tar archives on FTP

**Evidence/curation level:** Computationally predicted by AlphaFold2; not experimentally determined. Confidence scores (pLDDT) provide per-residue reliability estimates.

**Data Update Status:** Major releases periodically; AlphaFold DB v4 (2022) covers >200 million UniProt proteins. Updates follow UniProt releases.



**Licensing / access restrictions:** Freely available under Creative Commons Attribution 4.0 (CC BY 4.0). AlphaFold2 code available under Apache 2.0 license on GitHub.

**Citation / Recommended Reference:** Varadi M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022;50(D1):D439-D444. doi:10.1093/nar/gkab1061 (AlphaFold2 method: Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583-589. doi:10.1038/s41586-021-03819-2)

**Beginner-Friendly Explanation:** The AlphaFold Protein Structure Database contains predicted 3D shapes for almost every known protein — over 200 million of them. These shapes were predicted by a computer program called AlphaFold2, which uses artificial intelligence to figure out how a protein folds based on its sequence. This is revolutionary because most proteins have never had their structure determined experimentally. Each predicted structure comes with a confidence score for every part of the protein: blue regions are predicted with high confidence, while orange and red regions are less certain and may actually be flexible or disordered in the real protein.

**Advanced Technical Explanation:** AlphaFold2 uses a transformer-based neural network architecture (Evoformer) that processes multiple sequence alignments (MSAs) and pairwise residue distance information to predict 3D coordinates. The pLDDT (predicted local distance difference test) score is a per-residue confidence metric calibrated against the IDDT-Ca metric used in CASP assessments. The PAE (predicted aligned error) matrix provides confidence estimates for the relative position of every pair of residues, enabling assessment of inter-domain orientation reliability. Regions with pLDDT < 50 are typically intrinsically disordered; regions with pLDDT 50-70 should be interpreted cautiously; regions with pLDDT > 90 are generally reliable for structural analysis.

#### **One Practical Workflow Example: Using AlphaFold structures for functional annotation:**

Step 1: Search AlphaFold DB (<https://alphafold.ebi.ac.uk>) using the UniProt accession or protein name.

Step 2: View the structure in the browser, noting the pLDDT color coding (blue = high confidence, red = low confidence).

Step 3: Download the PDB file and PAE JSON file.

Step 4: Open the PDB file in PyMOL or ChimeraX; color by B-factor (which contains pLDDT scores) to visualize confidence.

Step 5: Use DALI or Foldseek to search for structurally similar proteins in the PDB, which may suggest function.

Step 6: For drug discovery applications, validate the predicted binding site against experimental data before proceeding with docking.



## K5 – SWISS-MODEL Repository

**Official Website URL:** <https://swissmodel.expasy.org/repository>

**Resource Type:** Database / Tool

**Main Biological Domain:** Structures

**What It Is Used For:** The SWISS-MODEL Repository provides pre-computed homology models for protein sequences in UniProtKB, generated by the SWISS-MODEL automated homology modeling server. Researchers use it to obtain structural models for proteins that lack experimental structures but have sufficient sequence similarity to a template with a known structure. SWISS-MODEL is particularly useful when a close structural template exists (>30% sequence identity), as homology models in this range are generally reliable for structural analysis.

**What Data It Contains:** The SWISS-MODEL Repository contains homology models for proteins in UniProt, generated using the best available templates from the PDB. Each model entry includes the model coordinates, the template used, sequence identity to the template, model quality scores (QMEAN, MolProbity), and coverage information.

**Main question it helps answer:** What is the best homology model available for this protein based on known structural templates, and how reliable is it?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- Is there a homology model available for this protein that I can use for docking, given that no experimental structure exists?
- What template was used to build the model, and what is the sequence identity to the template?
- How does the SWISS-MODEL homology model compare to the AlphaFold prediction for this protein?

**Example use cases:**

- Retrieving a pre-computed homology model for a protein of interest to use as a starting point for molecular dynamics simulation.
- Comparing SWISS-MODEL and AlphaFold predictions for a protein to assess consistency.
- Using the SWISS-MODEL server to build a custom homology model with a specific template.

**Input Data Accepted:** SWISS-MODEL accepts UniProt accession numbers for accessing repository models, protein sequences in FASTA format for custom modeling, and PDB template identifiers for template-guided homology modeling.

**Output Data Provided:** Outputs include homology model coordinates in PDB format, model quality scores such as QMEAN and MolProbity assessments, template information including PDB ID, chain, sequence identity, and sequence coverage, as well as structural alignments comparing the generated model with the selected template.

**Strengths:** SWISS-MODEL provides pre-computed structural models for many UniProt proteins and offers transparent template selection and model quality evaluation. The QMEAN scoring system provides a useful and widely accepted estimate of model reliability, and the repository is updated regularly as new experimental structures



become available in the PDB. The SWISS-MODEL server also supports custom modeling workflows, allowing users to specify preferred structural templates for targeted homology modeling.

**Limitations:** The quality of SWISS-MODEL predictions depends strongly on template availability and sequence similarity. Models built from templates with less than approximately 30% sequence identity are generally considered unreliable. Homology modeling cannot predict proteins with novel folds lacking suitable templates, and structural coverage is limited to regions supported by available templates. For proteins without close structural homologs, AlphaFold predictions now often provide superior performance.

**Common Beginner Mistakes:** A frequent mistake is using homology models generated from low-identity templates without recognizing their limited accuracy. Users also often neglect to examine QMEAN quality scores before relying on a model or confuse SWISS-MODEL homology models with AlphaFold predictions, despite their fundamentally different methodologies and assumptions.

**When to Use It:** Use SWISS-MODEL when a close structural template exists (>30% sequence identity) and you want a homology model with transparent template selection and quality assessment. It is also useful when you want to model a specific region using a specific template.

**When NOT to Use It:** For proteins without close templates, AlphaFold provides better predictions. For proteins with experimental structures, use the PDB directly.

**Related databases / alternatives:** Related resources include AlphaFold DB for AI-based structure prediction, particularly useful for proteins without close homologous templates; Modeller for command-line homology modeling; I-TASSER as an alternative structure prediction platform; and RoseTTAFold as another AI-based protein structure prediction approach.

**How It Connects to Other Resources:** SWISS-MODEL is part of the ExPASy bioinformatics portal and links to UniProt, PDB, and PROSITE. Models are built using PDB templates and quality-assessed using MolProbity and QMEAN. The SWISS-MODEL server integrates with UniProt to provide pre-computed models for UniProt entries.

**API / FTP / programmatic access:** SWISS-MODEL documentation and API information are available through [SWISS-MODEL API documentation](#). Repository entries can be queried directly using UniProt accessions via [SWISS-MODEL repository search](#), and programmatic model generation is supported through authenticated API-based workflows.

**Evidence/curation level:** Computationally generated homology models; quality assessed by QMEAN and MolProbity; not manually curated.

**Data Update Status:** Repository updated regularly as new PDB templates become available.

**Licensing / access restrictions:** Freely accessible for academic use.

**Citation / Recommended Reference:** Waterhouse A, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296-W303. doi:10.1093/nar/gky427

**Beginner-Friendly Explanation:** SWISS-MODEL is a tool that builds predicted protein structures by comparing your protein to proteins with known structures. If your protein is similar enough to a protein whose structure has been determined experimentally, SWISS-MODEL can use that known structure as a template to build a model of your protein. The repository contains pre-built models for many proteins, so you can often find a model without

having to run the modeling yourself. The quality of the model depends on how similar your protein is to the template — the more similar, the more reliable the model.

**Advanced Technical Explanation:** SWISS-MODEL uses a pipeline of template identification (BLAST, HHblits), template selection (based on sequence identity, coverage, and template quality), target-template alignment, and model building using ProMod3. The QMEAN (Qualitative Model Energy ANalysis) score estimates absolute model quality by comparing the model to a set of high-resolution experimental structures, with scores near 0 indicating good quality and scores below -4 indicating poor quality. The repository provides models for all UniProt sequences where a template with >30% sequence identity and >50% coverage can be identified in the PDB.

**One Practical Workflow Example: Retrieving and evaluating a SWISS-MODEL homology model:**

Step 1: Go to <https://swissmodel.expasy.org/repository> and enter the UniProt accession of your protein.

Step 2: Review the available models, noting the template PDB ID, sequence identity, coverage, and QMEAN score.

Step 3: Select the model with the best combination of high sequence identity, good coverage, and QMEAN score near 0.

Step 4: Download the model in PDB format.

Step 5: Compare the SWISS-MODEL model with the AlphaFold prediction (if available) using structural alignment in PyMOL or ChimeraX.

Step 6: Use the model for downstream analysis (docking, MD simulation), keeping in mind the limitations based on template identity.

## BEGINNER EXAMPLE (Category K)

---

A biochemist wants to understand the binding site of a drug target protein. They search RCSB PDB for the protein name and find three structures: one apo structure and two co-crystal structures with inhibitors. They download the highest-resolution co-crystal structure, open it in PyMOL, and visualize the inhibitor binding mode. They also check AlphaFold DB for the full-length protein structure, as the PDB structures only cover the catalytic domain.

## ADVANCED EXAMPLE (Category K)

---

A computational chemist is performing virtual screening against a novel drug target with no experimental structure. They retrieve the AlphaFold prediction, check pLDDT scores to identify the well-folded binding site region (pLDDT > 80), and compare it with the SWISS-MODEL homology model built on a 45%-identity template. They use the PAE matrix to assess inter-domain orientation confidence. After validating the predicted binding site against mutagenesis data from the literature, they perform molecular dynamics simulation to generate an ensemble of conformations for ensemble docking.

## CONFUSION POINTS (Category K)

---

AlphaFold vs. PDB: AlphaFold structures are predictions; PDB structures are experimental. Always prefer experimental structures when available. AlphaFold pLDDT scores are NOT B-factors, even though they are stored in the B-factor column of the PDB file. RCSB PDB vs. PDBe vs. PDBj: All three contain identical structural data. The choice is based on tools and interface preference, not data content. Asymmetric unit vs. biological assembly: The asymmetric unit in a crystal structure may not represent the functional oligomeric state. Always check the biological assembly annotation. Resolution: Lower resolution numbers mean higher quality. Beginners sometimes confuse this.

## DECISION GUIDE (Category K):

---

**Experimental structure available:** Use RCSB PDB (or PDBe/PDBj) No experimental structure, close template exists (>30% identity): Consider SWISS-MODEL for homology model No experimental structure, no close template: Use AlphaFold DB

**Need programmatic access to structural data:** PDBe API is well-documented

**Need structural similarity search:** PDBj Mine2 or DALI server

**Need cryo-EM maps:** EMDB (via PDBe)

## Category L: Variant and Mutation Databases

### CATEGORY OVERVIEW

Variant and mutation databases catalog genetic variants — differences in DNA sequence between individuals or between a sample and a reference genome — and provide information about their frequency, functional consequences, and clinical significance. These databases are essential for interpreting the results of genome sequencing in both research and clinical contexts. The spectrum of variants covered ranges from common single nucleotide polymorphisms (SNPs) with population frequencies above 1%, through rare variants associated with Mendelian diseases, to somatic mutations found specifically in cancer cells. Each type of variant requires different databases and analytical approaches.

The distinction between germline and somatic variants is fundamental in this category. Germline variants are inherited and present in every cell of an individual; they are the subject of population genetics, GWAS studies, and inherited disease research. Somatic variants arise during the lifetime of an organism, typically in specific tissues, and are not inherited; they are the primary focus of cancer genomics. Databases such as dbSNP and gnomAD focus on germline variants and their population frequencies, while COSMIC focuses exclusively on somatic mutations in cancer. ClinVar and ClinGen bridge the germline-clinical divide by curating evidence for the clinical significance of germline variants in disease. Understanding which type of variant you are studying is essential for choosing the appropriate database.

Variant interpretation is a rapidly evolving field with standardized frameworks for classifying variants as pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, or benign. The ACMG/AMP guidelines (Richards et al., 2015) provide the most widely used framework for germline variant classification, and ClinVar and ClinGen implement these guidelines. For somatic variants, the AMP/ASCO/CAP guidelines provide a tiered classification system. Researchers and clinicians using variant databases must understand these classification systems and recognize that variant classifications can change as new evidence accumulates — a variant classified as VUS today may be reclassified as pathogenic or benign as more data becomes available.

## dbSNP (Database of Single Nucleotide Polymorphisms)

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/snp>

**Resource Type:** Database

**Main Biological Domain:** Variants

**What It Is Used For:** dbSNP is NCBI's primary repository for short genetic variants, including single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), microsatellites, and other short variants. It is used to look up known variants by genomic position or rsID, to determine whether a variant has been previously observed, to retrieve population frequency data, and to obtain variant annotations for use in downstream analyses. dbSNP assigns stable reference SNP identifiers (rsIDs) that are widely used across the genomics literature and in variant annotation pipelines.

**What Data It Contains:** dbSNP contains hundreds of millions of variant records from diverse sources including genome sequencing projects, GWAS studies, and clinical sequencing. Each variant record includes the genomic position (on multiple reference assemblies), allele information, population frequency data from projects such as 1000 Genomes, gnomAD, and TOPMed, functional annotations (coding, intronic, UTR, etc.), and links to ClinVar for clinical significance data. The database covers variants from humans and many other organisms.

**Main question it helps answer:** Has this genetic variant been previously observed, what is its population frequency, and does it have a stable rsID identifier?

**Typical user:** Bioinformatician / Researcher / Clinician

**Example scientific questions:**

- What is the population frequency of this SNP in different ethnic groups?
- Does this variant have an rsID, and has it been reported in any GWAS studies?
- What is the functional annotation of this variant (synonymous, missense, splice site)?

**Example use cases:**

- Annotating variants from a WGS study with rsIDs and population frequencies using dbSNP as a reference.
- Filtering common variants (MAF > 1%) from a rare disease study using dbSNP frequency data.
- Looking up a specific rsID to retrieve all available information about a variant of interest.

**Input Data Accepted:** dbSNP accepts rsID identifiers (e.g., rs1234567), genomic coordinates including chromosome position and reference/alternate alleles, gene names or genomic regions, and batch queries performed through NCBI E-utilities.

**Output Data Provided:** Outputs include variant records containing rsIDs, genomic positions, and allele information; population frequency data derived from multiple studies and projects; functional annotations describing variant type and genomic context; links to related NCBI resources such as ClinVar and PubMed; and downloadable variant sets in VCF format.

**Strengths:** dbSNP is the largest repository of short genetic variants and serves as the primary source for rsID assignment. It provides population frequency information from multiple large-scale datasets and supplies stable rsIDs that function as universal variant identifiers throughout genetics and genomics research. The database is



tightly integrated with the broader NCBI ecosystem, including ClinVar, Gene, and PubMed, and supports variant information for many organisms beyond humans.

**Limitations:** dbSNP contains numerous legacy submissions with variable quality, and not all variants have consistent population frequency information available. Clinical significance annotations are limited, making ClinVar more appropriate for clinical interpretation. The web interface may perform slowly for large-scale queries, and some rsIDs have been merged or deprecated over time as database curation progresses.

**Common Beginner Mistakes:** Beginners frequently use dbSNP as the primary resource for clinical interpretation, despite ClinVar being more suitable for evaluating pathogenicity and clinical relevance. Another common mistake is failing to specify the genome assembly version when searching by genomic coordinates, since positions differ between assemblies such as GRCh37 and GRCh38. Users may also incorrectly assume that all dbSNP entries are experimentally validated and high quality, although many represent preliminary or lower-confidence submissions.

**When to Use It:** Use dbSNP when you need to look up rsIDs for variants, retrieve population frequency data, or annotate variants from a sequencing study with known identifiers. It is the standard reference for variant annotation pipelines.

**When NOT to Use It:** For clinical variant interpretation, use ClinVar. For high-quality population frequency data, gnomAD provides more detailed and better-curated frequency information. For somatic cancer mutations, use COSMIC.

**Related databases / alternatives:** Related resources include ClinVar for clinical significance interpretation, gnomAD for population allele frequencies, the European Variation Archive (EVA) maintained by EMBL-EBI, and the 1000 Genomes Project for population-scale genetic variation data.

**How It Connects to Other Resources:** dbSNP is integrated with NCBI's ClinVar (clinical significance), Gene (gene context), and PubMed (literature). rsIDs from dbSNP are used as universal identifiers in GWAS Catalog, gnomAD, Ensembl, and virtually all variant annotation tools. The EVA at EBI mirrors much of the dbSNP content.

**API / FTP / programmatic access:** Variant retrieval can be performed programmatically through NCBI E-utilities using esearch and efetch. The dbSNP REST API is available at [dbSNP REST API](#), and bulk data downloads are accessible through the NCBI FTP archive [dbSNP FTP archive](#). VCF datasets are available by chromosome and genome assembly version. Many annotation tools, including ANNOVAR and VEP, incorporate dbSNP as a reference source for variant annotation.

**Evidence/curation level:** Community-submitted; rsID assignment is automated; some curation for merging duplicate entries; quality varies widely by submission source.

**Data Update Status:** Continuously updated; major releases (builds) periodically. Build 156 (2023) contains >1 billion variant records.

**Licensing / access restrictions:** Freely available; no restrictions on use or redistribution.

**Citation / Recommended Reference:** Sherry ST, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308-11. doi:10.1093/nar/29.1.308

**Beginner-Friendly Explanation:** dbSNP is NCBI's catalog of genetic variants — the places in the genome where people differ from each other. Each variant gets a unique identifier called an rsID (like rs1234567) that scientists

use to refer to that specific variant in papers and databases. When you sequence someone's genome and find a variant, you can look it up in dbSNP to see if it has been seen before, how common it is in different populations, and whether it falls in a gene. It is one of the first places to check when interpreting sequencing results.

**Advanced Technical Explanation:** dbSNP assigns rsIDs through a clustering algorithm that merges variant submissions at the same genomic position with the same alleles. Each rsID record contains a RefSNP cluster with all submitted variant records (ssIDs) that have been merged into it. Population frequency data is aggregated from multiple sources including 1000 Genomes, gnomAD, TOPMed, and others, with allele frequencies reported per population. The dbSNP VCF files are organized by chromosome and genome assembly version (GRCh37/GRCh38) and are used as reference files in variant annotation pipelines such as ANNOVAR, VEP, and SnpEff.

**One Practical Workflow Example:** Annotating variants from a WGS study with dbSNP rsIDs:

- Step 1: Obtain your variant calls in VCF format from your variant calling pipeline (GATK HaplotypeCaller, DeepVariant, etc.).
- Step 2: Download the dbSNP VCF for the appropriate genome assembly: `wget https://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b156_GRCh38p14/VCF/GCF_000001405.40.gz`
- Step 3: Annotate your VCF with rsIDs using bcftools annotate: `bcftools annotate -a GCF_000001405.40.gz -c ID your_variants.vcf.gz -o annotated.vcf.gz`
- Step 4: Filter common variants ( $MAF > 0.01$ ) using the annotated frequency data.
- Step 5: For remaining rare variants, check ClinVar and gnomAD for clinical significance and population frequency.
- Step 6: Use the rsIDs to cross-reference with GWAS Catalog and literature.



## ClinVar

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/clinvar>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Variants / Clinical genomics

**What It Is Used For:** ClinVar is NCBI's archive of relationships between human genetic variants and phenotypes, with supporting evidence. It is the primary resource for clinical variant interpretation, aggregating submissions from clinical laboratories, research groups, and expert panels about the clinical significance of variants. Clinicians and researchers use ClinVar to determine whether a variant identified in a patient has been previously classified as pathogenic, benign, or of uncertain significance, and to access the evidence supporting each classification.

**What Data It Contains:** ClinVar contains over 2 million variant-condition pairs (as of 2024), with clinical significance classifications (pathogenic, likely pathogenic, VUS, likely benign, benign), submitter information, evidence summaries, and links to associated conditions (via MedGen/OMIM). Each submission includes the variant description, the condition, the classification, the review status (number of stars indicating evidence quality), and the submitting organization. ClinVar also aggregates submissions from multiple sources to provide a consensus classification.

**Main question it helps answer:** Has this variant been classified as pathogenic or benign for a specific condition, and what is the evidence supporting that classification?

**Typical user:** Clinician / Researcher / Bioinformatician (clinical genomics)

**Example scientific questions:**

- Has this BRCA1 variant been classified as pathogenic for hereditary breast and ovarian cancer?
- What is the review status of this variant's classification in ClinVar?
- Are there conflicting interpretations of this variant's clinical significance across different laboratories?

**Example use cases:**

- Interpreting a variant identified in a clinical exome sequencing report by checking its ClinVar classification and evidence.
- Filtering variants from a research cohort to identify those with established pathogenic classifications.
- Submitting variant interpretations from a clinical laboratory to contribute to the community knowledge base.

**Input Data Accepted:** ClinVar accepts variant identifiers including rsIDs, ClinVar accession numbers, and HGVS nomenclature; genomic coordinates; gene names; and disease or condition names, including MedGen and OMIM identifiers.

**Output Data Provided:** Outputs include clinical significance classifications accompanied by review status ratings ranging from 0 to 4 stars, submitter information and submission dates, evidence summaries with supporting literature references, disease and condition associations linked to MedGen, OMIM, and Human Phenotype Ontology (HPO) terms, downloadable VCF files for variant collections, and interpretation history documenting changes in clinical assessment over time.

**Strengths:** ClinVar is the primary public resource for clinical interpretation of genetic variants and aggregates submissions from hundreds of diagnostic laboratories, research groups, and expert panels. Its review status system provides a useful indicator of evidence quality and consensus, with expert panel and practice guideline

classifications representing the highest confidence levels. ClinVar integrates tightly with the NCBI ecosystem, linking to dbSNP, Gene, MedGen, and PubMed, and supports HGVS nomenclature for precise and standardized variant description.

**Limitations:** Many ClinVar variants have conflicting interpretations submitted by different laboratories or organizations, and a substantial proportion remain classified as variants of uncertain significance (VUS). Interpretation quality can vary substantially among submitters, and not all pathogenic variants are represented within ClinVar. Variant classifications may also become outdated as additional evidence accumulates, meaning historical interpretations may not always reflect current knowledge.

**Common Beginner Mistakes:** A common mistake is treating a VUS designation as evidence supporting either pathogenicity or benignity, despite its uncertain meaning. Users also often fail to examine the review status before relying on a classification, assume that a variant absent from ClinVar is benign, or overlook conflicting submissions that may significantly affect interpretation.

**When to Use It:** Use ClinVar whenever you need to interpret the clinical significance of a germline variant identified in a patient or research cohort. It is the standard reference for clinical genomics variant interpretation.

**When NOT to Use It:** ClinVar is not appropriate for somatic cancer variants (use COSMIC or OncoKB). For population frequency data, gnomAD is more appropriate. For variants in non-human organisms, ClinVar does not apply.

**Related databases / alternatives:** Related resources include LOVD, which provides locus-specific variant databases with gene-focused detail; ClinGen for expert curation of gene-disease and variant relationships; HGMD, a comprehensive mutation database requiring subscription for full access; and gnomAD, which complements ClinVar by providing large-scale population allele frequency data.

**How It Connects to Other Resources:** ClinVar links to dbSNP (rsIDs), Gene (gene context), MedGen (condition information), OMIM (disease associations), and PubMed (literature). ClinGen expert panels submit their variant classifications to ClinVar, providing the highest-quality (4-star) classifications. ClinVar data is used by variant annotation tools (VEP, ANNOVAR) and clinical reporting systems.

**API / FTP / programmatic access:** ClinVar records can be queried programmatically using NCBI E-utilities (esearch and efetch). Bulk downloads are available through the ClinVar FTP archive [ClinVar FTP archive](#), including VCF datasets such as clinvar.vcf.gz containing clinically annotated variants. ClinVar data can also be retrieved through the NCBI E-utilities API [NCBI E-utilities](#) (ClinVar) using db=clinvar. Python access is commonly implemented using requests with E-utilities or dedicated packages such as clinvar-this.

**Evidence/curation level:** Community-submitted with varying quality; review status (0-4 stars) indicates curation level. Expert panel submissions (4 stars) are the highest quality.

**Data Update Status:** Updated weekly; new submissions processed continuously.

**Licensing / access restrictions:** Freely available; no restrictions on use or redistribution.

**Citation / Recommended Reference:** Landrum MJ, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153

**Beginner-Friendly Explanation:** ClinVar is a database where clinical laboratories and researchers share their interpretations of genetic variants — whether a particular change in DNA is likely to cause disease or not. When a patient has genetic testing, the laboratory looks up variants in ClinVar to see if other labs have already figured out

whether that variant is harmful. Each entry has a "star rating" that tells you how much evidence supports the classification: four stars means an expert panel has carefully reviewed all the evidence, while one star means only one lab has submitted an interpretation.

**Advanced Technical Explanation:** ClinVar implements the ACMG/AMP five-tier variant classification system (pathogenic, likely pathogenic, VUS, likely benign, benign) with evidence codes from the Richards et al. 2015 guidelines. The review status system uses a star rating: 0 stars (no assertion criteria), 1 star (criteria provided, single submitter), 2 stars (criteria provided, multiple submitters, no conflicts), 3 stars (reviewed by expert panel), 4 stars (practice guideline). Variants are described using HGVS nomenclature at the DNA, RNA, and protein levels. The ClinVar VCF file includes CLNSIG (clinical significance), CLNREVSTAT (review status), and CLNDISEASE (associated condition) fields for use in variant annotation pipelines.

**One Practical Workflow Example:** Interpreting variants from a clinical exome sequencing study:

Step 1: Annotate your VCF with ClinVar classifications using VEP or ANNOVAR, which include ClinVar as a standard annotation source.

Step 2: Filter for variants with ClinVar classifications of "Pathogenic" or "Likely pathogenic" with review status  $\geq 1$  star.

Step 3: For each candidate variant, go to ClinVar directly and review the full record: check the review status, submitter information, and evidence summary.

Step 4: Check for conflicting interpretations from different submitters.

Step 5: For VUS variants in genes of interest, check ClinGen for gene-disease validity and any expert panel classifications.

Step 6: Document the ClinVar accession number and classification date in your variant interpretation report.

### L3: gnomAD (Genome Aggregation Database)

**Official Website URL:** <https://gnomad.broadinstitute.org>

**Resource Type:** Database

**Main Biological Domain:** Variants

**What It Is Used For:** gnomAD is the largest publicly available database of human genetic variation frequencies, aggregating exome and genome sequencing data from tens of thousands of individuals across diverse populations. It is used to determine the population frequency of genetic variants, to assess whether a variant is too common to be pathogenic for a rare disease, to evaluate constraint metrics (pLI, LOEUF) indicating whether genes tolerate loss-of-function variants, and to provide a reference for variant filtering in rare disease and population genetics studies.

**What Data It Contains:** gnomAD v4 (2023) contains variant frequencies from over 730,000 exomes and 76,000 genomes from individuals of diverse ancestries. For each variant, gnomAD provides allele frequency, allele count, allele number, and population-stratified frequencies across multiple ancestry groups (African, Admixed American, Ashkenazi Jewish, East Asian, Finnish, Middle Eastern, Non-Finnish European, South Asian, and others). Gene-level constraint metrics (pLI, LOEUF, Z-scores for missense and synonymous variants) are also provided.

**Main question it helps answer:** How common is this variant in the general population, and is it too frequent to be causative of a rare Mendelian disease?

**Typical user:** Bioinformatician / Clinician / Researcher (rare disease, population genetics)

**Example scientific questions:**

- What is the allele frequency of this variant in the non-Finnish European population?
- Is this gene intolerant to loss-of-function variants (high pLI score)?
- What is the maximum allele frequency of this variant across all gnomAD populations?

**Example use cases:**

- Filtering variants from a rare disease exome study by removing those with gnomAD allele frequency > 0.1% in any population.
- Using pLI scores to prioritize genes with loss-of-function variants in a developmental disorder cohort.
- Assessing the population frequency of a ClinVar VUS to inform its pathogenicity classification.

**Input Data Accepted:** gnomAD accepts genomic coordinates including chromosome position and reference/alternate alleles, rsID identifiers, gene names, and variant descriptions using HGVS nomenclature.

**Output Data Provided:** Outputs include allele frequency, allele count, and allele number across different populations, homozygote counts, variant quality metrics including filter status, gene constraint metrics such as pLI, LOEUF, and Z-scores, sequencing coverage information, and downloadable VCF files containing large-scale variant datasets.

**Strengths:** gnomAD is the largest publicly available human population variant frequency database and is widely regarded as the standard reference for allele frequency interpretation. It provides population-stratified frequencies across diverse ancestries, includes high-quality variant calls generated through rigorous quality control procedures, and covers both exome and genome sequencing datasets. Gene constraint metrics such as pLI and LOEUF are

extensively used in disease gene prioritization and variant interpretation. The resource is freely available and accessible without usage restrictions.

**Limitations:** Although gnomAD includes highly diverse datasets, ancestry representation remains somewhat biased toward populations of European descent. The database excludes individuals with severe pediatric disease to reduce ascertainment bias, although some pathogenic variants may still be present. Structural variants and large insertions/deletions are less comprehensively represented, mitochondrial variants require dedicated analyses, and certain rare variants may have artificially elevated frequencies because of sequencing or technical artifacts.

**Common Beginner Mistakes:** Beginners frequently use overall allele frequency instead of the maximum allele frequency observed across populations, which can lead to incorrect filtering decisions. Another common error is failing to recognize that gnomAD excludes severe pediatric disease cohorts, assuming that absence from gnomAD implies pathogenicity, or neglecting to examine filter status and quality labels such as PASS versus filtered variants before interpretation.

**When to Use It:** Use gnomAD whenever you need population frequency data for human variants, particularly for rare disease variant filtering and pathogenicity assessment. It is the standard reference for population frequency in clinical variant interpretation.

**When NOT to Use It:** gnomAD is not appropriate for somatic cancer variants. For non-human organisms, species-specific population databases are needed. gnomAD does not provide clinical significance information — use ClinVar for that.

**Related databases / alternatives:** Related resources include dbSNP, which provides a broader catalog of genetic variants but less detailed population frequency information; the 1000 Genomes Project, which offers a smaller but well-characterized population dataset; TOPMed, which includes large-scale sequencing data from diverse populations; and the UK Biobank, a large phenotype-linked sequencing resource with restricted access.

**How It Connects to Other Resources:** gnomAD frequencies are incorporated into ClinVar variant records, VEP annotations, and clinical reporting systems. gnomAD links to Ensembl for gene annotations and to ClinVar for clinical significance. The gnomAD constraint metrics (pLI, LOEUF) are widely used in gene prioritization tools such as Exomiser.

**API / FTP / programmatic access:** gnomAD provides a GraphQL API for programmatic access through [gnomAD GraphQL API](#). Bulk datasets are available through Google Cloud storage at <gs://gcp-public-data--gnomad/release/>, and downloadable release files can be accessed through [gnomAD downloads](#). Python access is supported through the gnomad package or by making direct API queries using requests.

**Evidence/curation level:** Computationally processed population sequencing data with rigorous QC; not manually curated for individual variants.

**Data Update Status:** Major releases periodically; gnomAD v4 (2023) is the current major release.

**Licensing / access restrictions:** Freely available under Creative Commons Attribution 4.0 (CC BY 4.0).

**Citation / Recommended Reference:** Chen S, et al. A genomic mutational constraint map using variation in 1,000 human exomes. *Nature*. 2024;625:92-100. doi:10.1038/s41586-023-06045-0 (gnomAD v4: Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434-443. doi:10.1038/s41586-020-2308-7)

**Beginner-Friendly Explanation:** gnomAD is a database that tells you how common a genetic variant is in the general population. It was built by combining sequencing data from hundreds of thousands of people who do not have severe genetic diseases. If you find a variant in a patient and want to know whether it could cause a rare disease, one of the first things to check is gnomAD: if the variant is common (found in many healthy people), it is unlikely to cause a rare disease. gnomAD also tells you how tolerant each gene is to having its function disrupted, which helps prioritize which variants are most likely to be harmful.

**Advanced Technical Explanation:** gnomAD aggregates exome and genome sequencing data processed through a standardized pipeline (BWA-MEM alignment, GATK variant calling, VQSR filtering) with additional sample-level and variant-level QC. Population structure is inferred using principal component analysis, and ancestry labels are assigned using a random forest classifier trained on reference populations. The pLI (probability of loss-of-function intolerance) score models the expected vs. observed number of loss-of-function variants per gene; pLI > 0.9 indicates strong intolerance. The LOEUF (loss-of-function observed/expected upper bound fraction) is a more continuous metric that has largely replaced pLI in recent analyses.

#### One practical workflow example:

##### Filtering rare disease variants using gnomAD:

Step 1: Annotate your VCF with gnomAD allele frequencies using VEP: `vep --input_file variants.vcf --output_file annotated.vcf --plugin gnomADc,gnomad.genomes.vcf.gz`

Step 2: Filter variants with gnomAD allele frequency > 0.001 (0.1%) in any population (use gnomAD\_AF\_popmax field).

Step 3: For remaining rare variants, check gnomAD homozygote counts: variants with many homozygotes in gnomAD are unlikely to cause recessive disease.

Step 4: For loss-of-function variants, check the gene's pLI or LOEUF score from the gnomAD gene constraint table.

Step 5: Cross-reference remaining candidates with ClinVar for clinical significance.

Step 6: For variants absent from gnomAD, check the coverage at that position to confirm the absence is not due to poor sequencing.

## L4: COSMIC (Catalogue of Somatic Mutations in Cancer) — Cross-reference Entry

**Database Name:** COSMIC (Catalogue of Somatic Mutations in Cancer)

**Entry type:** Cross-reference entry — full database card provided in Category W, W4.

COSMIC is mentioned in Category L because it is a major source for somatic cancer variant information. However, its primary domain is cancer genomics rather than general variant interpretation. For full coverage, including cancer-specific use cases, licensing restrictions, strengths, limitations, and workflow examples, see Category W: Cancer Genomics Databases, W4 – COSMIC.



## L5: Ensembl VEP (Variant Effect Predictor)

**Database Name:** Ensembl Variant Effect Predictor (VEP)

**Official Website URL:** <https://www.ensembl.org/vep>

**Resource Type:** Tool / Database

**Main Biological Domain:** Variants

**What It Is Used For:** Ensembl VEP is a tool for annotating and predicting the functional consequences of genetic variants. It is used to determine the effect of variants on genes, transcripts, and proteins, and to annotate variants with information from multiple databases including dbSNP, ClinVar, gnomAD, COSMIC, and others. VEP is one of the most widely used variant annotation tools in genomics and is available as a web tool, command-line tool, and REST API.

**What Data It Contains:** VEP itself is a tool rather than a database, but it integrates data from Ensembl gene annotations, dbSNP, ClinVar, gnomAD, COSMIC, SIFT, PolyPhen-2, CADD, and many other sources. It provides consequence terms (missense, synonymous, stop gained, splice site, etc.) using the Sequence Ontology vocabulary, along with pathogenicity predictions, population frequencies, and clinical significance annotations.

**Main question it helps answer:** What are the functional consequences of these variants on genes and proteins, and what do multiple annotation databases say about their significance?

**Typical user:** Bioinformatician / Researcher / Clinician

**Example scientific questions:**

- What is the predicted functional consequence of this variant on the encoded protein?
- Which of my variants are predicted to be damaging by SIFT and PolyPhen-2?
- What is the combined annotation from ClinVar, gnomAD, and COSMIC for this variant?

**Example use cases:**

- Running VEP on a VCF file from a WGS study to annotate all variants with functional consequences and database cross-references.
- Using VEP to filter variants by consequence type (e.g., keep only missense, stop gained, and splice site variants).
- Integrating VEP annotations into a clinical variant interpretation pipeline.

**Input Data Accepted:** The Ensembl Variant Effect Predictor (VEP) accepts VCF files containing variant calls, genomic coordinates in multiple formats, rsID identifiers, and variants described using HGVS nomenclature.

**Output Data Provided:** VEP produces functional consequence annotations using standardized Sequence Ontology terminology, pathogenicity predictions such as SIFT and PolyPhen-2 scores, population frequencies from databases including gnomAD and the 1000 Genomes Project, ClinVar clinical significance annotations, COSMIC somatic mutation information, and optional scores such as CADD through plugins. Results can be exported in VCF, TSV, or JSON formats.

**Strengths:** VEP integrates annotations from numerous biological databases within a single workflow, making it one of the most comprehensive variant annotation tools available. It uses standardized Sequence Ontology



consequence terms, supports extensive customization through plugins, and is accessible through web, command-line, Docker, and REST API interfaces. Regular updates synchronized with Ensembl releases ensure compatibility with current genome annotations and reference datasets.

**Limitations:** Annotation quality depends directly on the quality and update status of the underlying databases. Large VCF datasets may be processed slowly without parallelization or optimized computing resources, and some plugins require separate installation and database downloads. Results may differ across VEP versions due to updates in genome assemblies, transcript models, and annotation resources.

**Common Beginner Mistakes:** A frequent error fails to specify the correct genome assembly version, particularly when distinguishing between GRCh37 and GRCh38. Users also often neglect to enable comprehensive annotations using the `--everything` option or overlook transcript selection, leading to confusion between canonical transcript annotations and annotations across all transcripts.

**When to Use It:** Use VEP as the primary variant annotation tool for any genomics study requiring functional consequence annotation and database cross-referencing. It is the standard tool for variant annotation in both research and clinical genomics pipelines.

**When NOT Use It:** VEP is a tool, not a database; it does not store variant data itself. For looking up specific variants, use the underlying databases (ClinVar, gnomAD, COSMIC) directly.

**Related databases / alternatives:** Alternative variant annotation tools include ANNOVAR, SnpEff, and Nirvana, each with distinct strengths and workflow preferences.

**How It Connects to Other Resources:** VEP integrates data from Ensembl, dbSNP, ClinVar, gnomAD, COSMIC, SIFT, PolyPhen-2, and many other resources through its plugin system. VEP output is used as input for downstream variant prioritization tools such as Exomiser and CADD.

**API / FTP / programmatic access:** VEP provides a REST API for programmatic annotation through [Ensembl VEP REST API](#). Command-line annotation can be performed using:

```
vep --input_file variants.vcf --output_file annotated.vcf --everything --assembly GRCh38
```

Containerized deployment is available via Docker ([ensemblorg/ensembl-vep](#)), and installation through Conda can be performed using: `conda install -c bioconda ensembl-vep`

**Evidence/curation level:** Tool that integrates data from multiple sources with varying curation levels.

**Data update status:** Updated with each Ensembl release (approximately every 3 months).

**Licensing / access restrictions:** Freely available under Apache 2.0 license.

**Citation / recommended reference:** McLaren W, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17(1):122. doi:10.1186/s13059-016-0974-4

**Beginner-friendly explanation:** Ensembl VEP is a tool that takes a list of genetic variants and tells you what effect each one might have. For each variant, it tells you whether it falls in a gene, whether it changes the protein sequence, whether it has been seen in healthy people (and how often), and whether it has been linked to disease. Think of it as a one-stop annotation service that looks up your variants in many different databases simultaneously and summarizes all the information in one place.

**Advanced Technical Explanation:** VEP maps variants to Ensembl transcript coordinates and assigns consequence terms using the Sequence Ontology hierarchy, with the most severe consequence reported first. The `--pick` flag selects one consequence per variant based on a configurable priority order. VEP's plugin system allows integration of additional annotation sources (CADD, SpliceAI, AlphaMissense, etc.) through standardized plugin scripts. The REST API supports batch queries of up to 200 variants per request and returns JSON-formatted annotations.

**One practical workflow example: Annotating a VCF file with VEP:**

Step 1: Install VEP: `conda install -c bioconda ensembl-vep`

Step 2: Download the VEP cache for your assembly: `vep_install -a cf -s homo_sapiens -y GRCh38`

Step 3: Run VEP with standard annotations: `vep --input_file variants.vcf --output_file annotated.vcf --everything --assembly GRCh38 --cache --offline --format vcf --vcf --fork 4`

Step 4: Add gnomAD and ClinVar annotations via plugins if not included in the cache.

Step 5: Parse the annotated VCF to filter by consequence type and pathogenicity predictions.

Step 6: Prioritize variants using the combined annotations for downstream analysis.

## L6: LOVD (Leiden Open Variation Database)

**Official Website URL:** <https://www.lovd.nl>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Variants / Clinical genomics

**What It Is Used For:** LOVD is a flexible, freely available tool for gene-centered collection and display of DNA variants. It is used to create and maintain locus-specific databases (LSDBs) for individual genes, providing detailed variant information including clinical phenotypes, functional data, and literature references. LOVD is particularly valuable for rare disease genes where detailed, gene-specific variant curation is needed beyond what ClinVar provides.

**What Data It Contains:** LOVD hosts thousands of gene-specific databases contributed by research groups and clinical laboratories worldwide. Each database contains variants in a specific gene with associated phenotype information, variant classifications, functional data, and submitter information. LOVD3 (the current version) supports sharing of variant data across installations and integration with ClinVar.

**Main question it helps answer:** What variants have been reported in this specific gene, with what phenotypes and functional evidence?

**Typical user:** Clinician / Researcher (rare disease genetics)

**Example scientific questions:**

- What variants in BRCA2 have been reported in LOVD, and what phenotypes are associated with them?
- Is there a LOVD database for this rare disease gene that I can submit my patient's variant to?

**Example use cases:**

- Checking a gene-specific LOVD database for detailed variant information not available in ClinVar.
- Setting up a LOVD database for a newly characterized disease gene.
- Submitting patient variants to a gene-specific LOVD database to contribute to the community knowledge base.

**Input data accepted:** Gene names; Variant descriptions in HGVS notation; Patient phenotype information

**Output data provided:** Gene-specific variant lists with phenotype and classification data; Variant detail pages with functional evidence and literature; Export in various formats

**Strengths:** Gene-specific depth of annotation not available in general databases; Supports detailed phenotype information; Community-driven curation by gene experts; Free to install and use for creating new databases

**Limitations:** Coverage is uneven; not all genes have active LOVD databases; Quality varies by database curator; Less comprehensive than ClinVar for overall variant coverage; Some databases are not actively maintained

**Common beginner mistakes:** Not checking whether a LOVD database exists for the gene of interest; Overlooking LOVD when ClinVar has limited information for a specific gene

**When to Use It:** Use LOVD when you need gene-specific variant information with detailed phenotype data, particularly for rare disease genes with active LOVD databases maintained by expert curators.

**When NOT to Use It:** For broad variant interpretation across many genes, ClinVar is more comprehensive. LOVD is most valuable for specific genes with active community databases.

**Related databases / alternatives:** ClinVar: broader coverage; less gene-specific depth | HGMD: comprehensive but subscription required for full access | Gene-specific databases (e.g., BRCA Exchange for BRCA1/2)

**How It Connects to Other Resources:** LOVD integrates with ClinVar for variant sharing and links to OMIM, PubMed, and Ensembl. LOVD3 supports the GA4GH variant representation standards.

**API / FTP / programmatic access:** LOVD API: available at individual LOVD installations; Data export in TSV and other formats; LOVD3 supports REST API for programmatic access

**Evidence/curation level:** Community-submitted with varying quality; some databases are expertly curated, others are less actively maintained.

**Data Update Status:** Varies by individual database; some are actively updated, others are static.

**Licensing / access restrictions:** Freely accessible; LOVD software is open source.

**Citation / Recommended Reference:** Fokkema IF, et al. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011;32(5):557-63. doi:10.1002/humu.21438

**Beginner-Friendly Explanation:** LOVD is a system for creating gene-specific databases of genetic variants. While ClinVar collects variants from all genes in one place, LOVD allows experts who specialize in a particular gene to create a detailed database just for that gene, with more information about each variant than a general database can provide. If you are studying a rare disease gene, there may be a LOVD database maintained by the leading experts in that gene, which could have more detailed and up-to-date information than ClinVar.

**Advanced Technical Explanation:** LOVD3 implements a distributed database model where individual installations can share data through a central LOVD sharing infrastructure. Variants are described using HGVS nomenclature at the DNA, RNA, and protein levels, with validation against reference sequences. The database schema supports customizable phenotype fields, allowing gene-specific databases to capture disease-relevant clinical information. LOVD3 supports submission to ClinVar through a standardized export format.

#### **One practical workflow example: Finding gene-specific variant information in LOVD:**

Step 1: Go to <https://www.lovd.nl> and search for your gene of interest.

Step 2: If a LOVD database exists for the gene, navigate to the variant list and search for your specific variant.

Step 3: Review the variant detail page for phenotype information, functional data, and literature references.

Step 4: Compare the LOVD classification with ClinVar for consistency.

Step 5: If your variant is not in LOVD, consider submitting it to contribute to the community database.

## **L7: ClinGen (Clinical Genome Resource) — Cross-reference Entry**

**Official Website URL:** <https://clinicalgenome.org>

**Entry type:** Cross-reference entry — full database card provided in Category AF, AF3.

ClinGen is mentioned in Category L because variant interpretation depends on expert-reviewed clinical validity and actionability frameworks. For full coverage of ClinGen, including gene-disease validity, dosage sensitivity, expert panels, clinical actionability, and evidence frameworks, see Category AF: GWAS, Rare Disease, and Clinical Variant Interpretation Resources, AF3 – ClinGen.

## BEGINNER EXAMPLE (Category L):

---

A clinical genetics trainee receives an exome sequencing report for a patient with suspected hereditary breast cancer. The report lists a BRCA1 variant (c.5266dupC, p.Gln1756Profs\*74). They look it up in ClinVar and find it classified as Pathogenic with 4-star review status from the ClinGen BRCA1/2 expert panel. They check gnomAD and find it is absent from the general population. They confirm the classification in LOVD's BRCA1 database and document the evidence for the clinical report.

## ADVANCED EXAMPLE (Category L):

---

A cancer genomics researcher is analyzing somatic mutations from 200 colorectal cancer tumor-normal pairs. They run VEP to annotate all somatic variants, then filter for variants in COSMIC Cancer Gene Census Tier 1 genes. They use gnomAD to remove any variants present in the germline population (potential germline contamination). They analyze the mutational signatures using the COSMIC signature framework and identify SBS1, SBS5, and SBS44 (mismatch repair deficiency) as dominant signatures. They cross-reference driver mutations with OncoKB for therapeutic implications.

## CONFUSION POINTS (Category L):

---

Germline vs. somatic: dbSNP, ClinVar, gnomAD = germline variants; COSMIC = somatic cancer mutations. Never use gnomAD frequencies to filter somatic variants from tumor sequencing.

ClinVar vs. ClinGen: ClinVar is a submission database; ClinGen is a curation resource. ClinGen submits to ClinVar. ClinGen 4-star entries in ClinVar are the most reliable.

VUS: A variant of uncertain significance (VUS) in ClinVar is NOT evidence of pathogenicity. It means there is insufficient evidence to classify it either way.

VEP is a tool, not a database: VEP annotates variants using other databases; it does not store variant data itself.

## DECISION GUIDE (Category L):

---

For variant rsIDs and population frequency overview: dbSNP; For high-quality population frequencies and gene constraint: gnomAD; For clinical significance of germline variants: ClinVar; For highest-quality expert panel classifications: ClinGen; For somatic cancer mutations: COSMIC; For gene-specific variant databases: LOVD; For functional consequence annotation: Ensembl VEP; For comprehensive variant annotation pipeline: VEP + gnomAD + ClinVar + COSMIC

## Category M: Disease and Clinical Genomics Databases

### CATEGORY OVERVIEW

Disease and clinical genomics databases catalog the relationships between genes, genetic variants, and human diseases, providing the knowledge infrastructure for translating genomic discoveries into clinical understanding. These databases serve as the bridge between basic research findings — the identification of genes and variants associated with disease — and clinical application, including genetic counseling, diagnostic testing, and therapeutic decision-making. They range from comprehensive catalogs of Mendelian disease genes (OMIM) to resources focused on rare diseases (Orphanet), evidence-based gene-disease curation (ClinGen), gene-disease association mining (DisGeNET), and integrated disease information portals (MalaCards).

The distinction between different types of gene-disease evidence is critical in this category. Some databases (OMIM) catalog gene-disease relationships based on published literature without applying a formal evidence-grading framework, while others (ClinGen) apply rigorous semi-quantitative frameworks to classify the strength of evidence. DisGeNET aggregates gene-disease associations from multiple sources including curated databases, text mining, and animal models, providing a broad but heterogeneous evidence base. MalaCards integrates information from dozens of sources to provide a comprehensive disease card for each condition. Understanding the evidence basis for each database's claims is essential for appropriate use in research and clinical contexts.

A key challenge in this category is disease nomenclature and ontology. Different databases use different disease classification systems: OMIM uses its own MIM numbers, Orphanet uses ORPHA codes, ICD-10/11 is used in clinical settings, and the Human Phenotype Ontology (HPO) provides a standardized vocabulary for clinical features. Many databases provide cross-references between these systems, but mapping between them is not always straightforward. Researchers working across multiple disease databases should be aware of these nomenclature differences and use ontology mapping tools (e.g., the Monarch Initiative's disease ontology mappings) when integrating data from multiple sources.

## M1: OMIM (Online Mendelian Inheritance in Man)

---

**Official Website URL:** <https://www.omim.org>

**Resource Type:** Knowledgebase

**Main Biological Domain:** Diseases / Clinical genomics

**What It Is Used For:** OMIM is the authoritative catalog of human genes and genetic disorders, providing comprehensive, regularly updated information about the molecular basis of Mendelian diseases. Researchers and clinicians use OMIM to look up the genetic basis of inherited diseases, identify genes associated with specific phenotypes, understand the molecular mechanisms of genetic disorders, and access curated literature summaries for gene-disease relationships. OMIM is the standard reference for Mendelian disease genetics.

**What Data It Contains:** OMIM contains over 26,000 entries (as of 2024) covering genes (gene entries, asterisk prefix), phenotypes (phenotype entries, hash prefix), and gene-phenotype relationships. Each entry includes a detailed text summary of the clinical features, molecular genetics, inheritance pattern, and history of the disorder, with extensive literature references. OMIM assigns unique MIM numbers to each gene and phenotype, which are widely used as identifiers in the genetics literature.

**Main question it helps answer:** What is the genetic basis of this inherited disease, and what genes are associated with this phenotype?

**Typical user:** Clinician / Researcher / Genetic counselor

**Example scientific questions:**

- What genes are known to cause autosomal recessive intellectual disability?
- What is the molecular mechanism of Marfan syndrome, and which gene is responsible?
- What phenotypes are associated with mutations in the FBN1 gene?

**Example use cases:**

- Looking up the genetic basis of a suspected Mendelian disorder in a patient with a characteristic phenotype.
- Identifying all genes associated with a specific disease for inclusion in a clinical sequencing panel.
- Reviewing the clinical features and molecular genetics of a disease before designing a research study.

**Input Data Accepted:** OMIM accepts MIM identifiers for genes or phenotypes, gene names, disease names, phenotype keywords, and search queries based on clinical features.

**Output Data Provided:** Outputs include detailed expert-written entries describing clinical features, molecular genetics, inheritance patterns, and historical context; gene-phenotype relationship tables; MIM identifiers for cross-referencing; and links to related resources such as ClinVar, Ensembl, and PubMed.

**Strengths:** OMIM is widely regarded as the most comprehensive and authoritative catalog of Mendelian disease genetics. Its entries provide detailed, expert-curated summaries supported by extensive literature references and are updated regularly by specialists. MIM identifiers function as universal reference numbers throughout medical genetics, and OMIM's phenotypic series framework enables comparative analysis of related genetic disorders.

**Limitations:** Access to some OMIM features, including API services, requires registration or institutional access. Unlike ClinGen, OMIM does not apply a formal evidence-grading framework to gene-disease relationships. Coverage is focused primarily on Mendelian single-gene disorders and therefore does not comprehensively



represent complex multifactorial diseases. Its text-heavy format can also make computational analysis challenging, and some entries may lag behind the latest published evidence.

**Common Beginner Mistakes:** A common mistake is confusing gene entries and phenotype entries within OMIM. Gene entries are typically indicated by an asterisk (e.g., \*604370), whereas phenotype entries commonly use a hash symbol (e.g., #154700). Users may also overlook the fact that OMIM does not formally grade evidence strength and may incorrectly rely on OMIM alone for variant interpretation without consulting ClinVar or ClinGen.

**When to Use It:** Use OMIM when you need comprehensive information about the genetic basis of a Mendelian disorder, including clinical features, molecular mechanisms, inheritance patterns, and literature history. It is the first resource to consult for Mendelian disease genetics.

**When NOT to Use It:** OMIM is not appropriate for complex multifactorial diseases, somatic cancer mutations, or population-level variant frequency data. For evidence-graded gene-disease relationships, ClinGen is more rigorous.

**Related databases / alternatives:** Related resources include Orphanet for rare disease information with broader disease coverage, ClinGen for evidence-based gene-disease validity assessment, DisGeNET for broader gene-disease associations including complex disorders, and DECIPHER for developmental disorder genomics and phenotype interpretation.

**How It Connects to Other Resources:** OMIM links to ClinVar (variant data), Ensembl (gene information), PubMed (literature), HPO (phenotype ontology), and Orphanet (rare disease information). MIM numbers are used as disease identifiers in ClinVar, ClinGen, and many other databases. OMIM is integrated into clinical genomics tools such as Exomiser for phenotype-driven variant prioritization.

**API / FTP / programmatic access:** OMIM provides API access through [OMIM API](#), which requires an API key and is freely available for academic use. Downloadable datasets are available through [OMIM downloads](#) with registration. Important downloadable files include genemap2.txt for gene–phenotype mappings and mim2gene.txt for mapping MIM identifiers to gene IDs.

**Evidence/curation level:** Manually curated by expert editors; literature-based; no formal evidence- grading framework applied.

**Data Update Status:** Continuously updated; new entries and updates added regularly.

**Licensing / access restrictions:** Free registration required for full text access. API access requires a free API key. Commercial use requires a license.

**Citation / Recommended Reference:** Amberger JS, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789-98. doi:10.1093/nar/gku1205

**Beginner-Friendly Explanation:** OMIM is like an encyclopedia of inherited genetic diseases. For each known genetic disorder, OMIM has a detailed entry written by medical genetics experts that describes what the disease looks like, which gene causes it, how it is inherited, and what is known about the molecular mechanism. It also has entries for each gene, describing all the diseases that mutations in that gene can cause. OMIM is the standard reference that doctors and researchers consult when they want to understand the genetic basis of an inherited condition.

**Advanced Technical Explanation:** OMIM uses a structured entry format with standardized section headers (Description, Clinical Features, Inheritance, Molecular Genetics, etc.) and a controlled vocabulary for inheritance

patterns (autosomal dominant, autosomal recessive, X-linked, etc.). The gene-phenotype relationship table uses confidence codes (3 = molecular basis known, 2 = phenotype mapped to chromosomal region, 1 = phenotype placed on map by statistical methods) to indicate the strength of evidence. MIM numbers follow a prefix convention: asterisk (\*) for genes, hash (#) for phenotypes with known molecular basis, percent (%) for phenotypes with unknown molecular basis, and plus (+) for genes with known phenotype. The OMIM API provides programmatic access to entry text, gene-phenotype maps, and clinical synopsis data in JSON format.

**One Practical Workflow Example: Using OMIM to identify genes for a clinical sequencing panel:**

Step 1: Go to <https://www.omim.org> and search for the disease of interest (e.g., "Ehlers-Danlos syndrome").

Step 2: Review the phenotype entries (hash prefix) to identify all recognized subtypes and their associated genes.

Step 3: For each gene, check the confidence code in the gene-phenotype table (code 3 = molecular basis known).

Step 4: Download the `genemap2.txt` file from OMIM downloads for programmatic access to all gene-phenotype relationships.

Step 5: Cross-reference with ClinGen gene validity classifications to assess evidence strength for each gene-disease pair.

Step 6: Use the resulting gene list to design or evaluate a clinical sequencing panel.

## M2: Orphanet

**Official Website URL:** <https://www.orpha.net>

**Resource Type:** Knowledgebase / Portal

**Main Biological Domain:** Diseases / Clinical genomics

**What It Is Used For:** Orphanet is the reference portal for information on rare diseases and orphan drugs, maintained by a European consortium. It provides comprehensive information about rare diseases including clinical descriptions, diagnostic criteria, management guidelines, genetic information, and links to expert centers, patient registries, and clinical trials. Orphanet is the primary resource for rare disease information in Europe and is widely used globally for rare disease research and clinical management.

**What Data It Contains:** Orphanet contains information on over 10,000 rare diseases, each with a unique ORPHA code. For each disease, Orphanet provides a clinical description, epidemiology (prevalence, incidence), genetic information (causative genes, inheritance), diagnostic criteria, differential diagnosis, management guidelines, and links to expert centers, patient organizations, and clinical trials. Orphanet also maintains a classification of rare diseases and a nomenclature system.

**Main question it helps answer:** What is known about this rare disease, including its clinical features, genetic basis, prevalence, and available resources for patients and clinicians?

**Typical user:** Clinician / Researcher / Patient / Genetic counselor

**Example scientific questions:**

- What is the prevalence of this rare disease in Europe?
- What genes are associated with this rare syndrome, and what is the inheritance pattern?
- Are there expert centers or patient registries for this rare disease?

**Example use cases:**

- Looking up clinical information about a rare disease for patient counseling or clinical management.
- Identifying expert centers and patient registries for a rare disease research project.
- Using ORPHA codes as disease identifiers in rare disease research databases and registries.

**Input Data Accepted:** Orphanet accepts disease names, ORPHA codes, gene names, and clinical features that can be used for disease search and differential diagnosis.

**Output Data Provided:** Outputs include disease information pages describing clinical presentation, genetics, and management; ORPHA codes used for disease identification; links to expert centers, patient organizations, and clinical trials; and downloadable disease classifications and nomenclature datasets.

**Strengths:** Orphanet is one of the most comprehensive resources dedicated to rare diseases and is widely used as a reference source in rare disease medicine and research. ORPHA codes function as standardized identifiers for rare disorders and facilitate interoperability between databases. The platform supports multiple European languages and provides practical resources such as expert clinical centers and patient organizations. Information is regularly curated and updated by rare disease specialists.

**Limitations:** Orphanet has a primarily European focus, and some diseases may be represented more comprehensively in region-specific resources. The level of clinical detail varies between diseases, and Orphanet is not intended for variant interpretation or pathogenicity assessment. Some disease entries may also lag behind the most recent scientific publications.

**Common Beginner Mistakes:** A common misunderstanding is confusing Orphanet with variant interpretation resources such as ClinVar. Orphanet focuses on disease-level information rather than variant pathogenicity. Another frequent oversight is failing to use ORPHA codes during database integration and cross-referencing, despite their importance as standardized rare disease identifiers.

**When to Use It:** Use Orphanet when you need comprehensive information about a rare disease, including clinical features, genetics, prevalence, and practical resources. It is the standard reference for rare disease information in Europe.

**When NOT to Use It:** For variant interpretation, use ClinVar. For common diseases, OMIM or DisGeNET may be more appropriate. Orphanet focuses on rare diseases (prevalence < 1 in 2,000).

**Related databases / alternatives:** Related resources include OMIM for Mendelian disease genetics, NORD for United States-focused rare disease information, and GARD (Genetic and Rare Diseases Information Center) for NIH-supported rare disease information and patient-oriented resources.

**How It Connects to Other Resources:** Orphanet links to OMIM (MIM numbers), ClinVar, Ensembl, and clinical trial databases. ORPHA codes are used as disease identifiers in the Human Phenotype Ontology (HPO), DECIPHER, and many rare disease research databases. Orphanet data is integrated into the Monarch Initiative's disease ontology.

**API / FTP / programmatic access:** Orphanet provides programmatic access through the Orphacode API [Orphanet API](#). Downloadable datasets are available through [Orphadata downloads](#) and are freely accessible for academic use. Orphadata provides XML datasets covering disease classifications, gene–disease associations, epidemiological information, and nomenclature resources.

**Evidence/curation level:** Manually curated by rare disease experts; literature-based; regularly reviewed and updated.

**Data Update Status:** Continuously updated; major releases periodically.

**Licensing / access restrictions:** Freely accessible; data downloads available for academic use through Orphadata.

**Citation / Recommended Reference:** Orphanet: an online rare disease and orphan drug data base. Copyright, INSERM 1997. Available at <http://www.orpha.net>. Accessed [date].

**Beginner-Friendly Explanation:** Orphanet is a comprehensive information resource for rare diseases — those affecting fewer than 1 in 2,000 people. For each rare disease, Orphanet provides a detailed description of the symptoms, how common it is, which genes are involved, how it is inherited, and how it is managed. It also provides links to specialist centers, patient support groups, and clinical trials. Orphanet assigns each disease a unique code (ORPHA code) that is used as a standard identifier in rare disease research across Europe and internationally.

**Advanced Technical Explanation:** Orphanet uses a hierarchical disease classification system with multiple levels of granularity, from broad disease groups to specific subtypes. ORPHA codes are stable identifiers that persist

across database updates, enabling longitudinal tracking of disease entities. The Orphanet nomenclature system provides standardized disease names in multiple languages with cross-references to ICD-10, ICD-11, OMIM, UMLS, MeSH, and other classification systems. The Orphadata platform provides programmatic access to Orphanet data in XML format, including gene-disease associations with evidence levels and inheritance patterns.

### One Practical Workflow Example: Using Orphanet for rare disease research:

Step 1: Go to <https://www.orpha.net> and search for the rare disease by name or clinical features.

Step 2: Note the ORPHA code for the disease (e.g., ORPHA:558 for Marfan syndrome).

Step 3: Review the disease page for clinical description, genetics, and management information.

Step 4: Use the ORPHA code to cross-reference with other databases (HPO, DECIPHER, patient registries).

Step 5: Download gene-disease association data from Orphadata for computational analysis: `wget https://www.orphadata.com/data/xml/en_product6.xml`

Step 6: Parse the XML to extract gene-disease relationships for your analysis.

## M3: ClinGen — Cross-reference Entry

---

**Official Website URL:** <https://clinicalgenome.org>

### L7 – ClinGen (Clinical Genome Resource) — Cross-reference Entry

**Entry type:** Cross-reference entry — full database card provided in Category AF, AF3.

ClinGen is mentioned in Category L because variant interpretation depends on expert-reviewed clinical validity and actionability frameworks. For full coverage of ClinGen, including gene-disease validity, dosage sensitivity, expert panels, clinical actionability, and evidence frameworks, see Category AF: GWAS, Rare Disease, and Clinical Variant Interpretation Resources, AF3 – ClinGen.

## M4: DisGeNET

**Official Website URL:** <https://www.disgenet.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Diseases / Clinical genomics

**What It Is Used For:** DisGeNET is a comprehensive platform integrating information on gene-disease associations from multiple curated databases, GWAS catalogs, animal models, and text mining of the biomedical literature. It is used to explore the genetic basis of human diseases, identify genes associated with a disease of interest, find diseases associated with a gene of interest, and perform network and enrichment analyses of gene-disease relationships. DisGeNET provides a broader view of gene-disease associations than OMIM, including complex multifactorial diseases.

**What Data It Contains:** DisGeNET contains over 1.5 million gene-disease associations (as of 2024) covering over 30,000 diseases and traits, sourced from curated databases (UniProt, ClinVar, OMIM, Orphanet, CTD), GWAS catalogs (GWAS Catalog, PharmGKB), animal models (MGD, RGD), and text mining. Each association includes a score (GDA score) reflecting the strength and number of evidence sources, evidence type, and source information.

**Main question it helps answer:** What genes are associated with this disease, and what diseases are associated with this gene, based on integrated evidence from multiple sources?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What genes have the strongest evidence for association with type 2 diabetes?
- What diseases are associated with the APOE gene?
- What is the overlap in genetic architecture between Alzheimer's disease and Parkinson's disease?

**Example use cases:**

- Performing gene-disease enrichment analysis on a list of differentially expressed genes.
- Building a disease gene network for a complex disease to identify hub genes.
- Comparing the genetic basis of related diseases to identify shared mechanisms.

**Input Data Accepted:** DisGeNET accepts gene names, Entrez Gene identifiers, UniProt accession numbers, disease names, UMLS Concept Unique Identifiers (CUIs), MeSH terms, OMIM identifiers, and search queries through its web interface or API.

**Output Data Provided:** Outputs include gene-disease association (GDA) tables with quantitative scores and evidence sources, disease-gene interaction networks, enrichment analysis results, and downloadable association datasets in TSV format.

**Strengths:** DisGeNET provides one of the broadest collections of gene-disease associations available, covering both Mendelian and complex multifactorial diseases. It integrates diverse evidence types, including curated databases, genome-wide association studies (GWAS), animal models, and literature text mining. The GDA score offers a quantitative estimate of association strength, and the platform is particularly useful for disease network

analysis, pathway enrichment, and systems biology studies. REST API access supports large-scale computational workflows and integration into bioinformatics pipelines.

**Limitations:** Associations derived from text mining can have relatively high false positive rates and therefore require careful interpretation. DisGeNET GDA scores represent association strength rather than formal clinical evidence grading and should not be interpreted as equivalent to ClinGen or clinical pathogenicity evidence. Coverage of rare Mendelian diseases is generally less detailed than specialized resources such as OMIM. Full access to some datasets and features requires registration, and certain uses may involve licensing restrictions.

**Common Beginner Mistakes:** A common error is treating all DisGeNET associations as equally reliable without considering their evidence source. Users often fail to filter results by evidence category, particularly when high-confidence curated evidence is needed. Another frequent misunderstanding is confusing DisGeNET's broad association data with clinically validated evidence from ClinGen or OMIM.

**When to Use It:** Use DisGeNET when you need a broad view of gene-disease associations for complex diseases, for enrichment analysis of gene lists, or for network analysis of disease genetics. It is particularly useful for complex multifactorial diseases not well covered by OMIM.

**When NOT to Use It:** For Mendelian disease genetics, OMIM provides more detailed and reliable information. For clinical variant interpretation, ClinVar and ClinGen are more appropriate. Do not use DisGeNET text mining associations as clinical evidence.

**Related databases / alternatives:** Related resources include OMIM for Mendelian disease genetics, GWAS Catalog for genome-wide association studies, Open Targets for therapeutic target and disease associations, and MalaCards for integrated disease information.

**How It Connects to Other Resources:** DisGeNET integrates data from OMIM, ClinVar, Orphanet, UniProt, GWAS Catalog, and other sources. It links to Ensembl, UniProt, and disease ontologies. DisGeNET data is used in network medicine analyses and is integrated into tools such as Cytoscape plugins and R/Bioconductor packages.

**API / FTP / programmatic access:** DisGeNET provides programmatic access through the REST API at [DisGeNET API](#). Data downloads are available through [DisGeNET downloads](#), which requires registration. Computational access is also supported through the `disgenet2r` Bioconductor package and standard Python REST API workflows using requests.

**Evidence/curation level:** Mixed: curated associations from expert databases (high quality); GWAS associations (moderate quality); text mining associations (lower quality, higher false positive rate).

**Data Update Status:** Updated periodically; check the DisGeNET website for current version.

**Licensing / access restrictions:** Free registration required for data access. Academic use is free; commercial use requires a license.

**Citation / Recommended Reference:** Pinero J, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 2017;45(D1):D833-D839. doi:10.1093/nar/gkw943

**Beginner-Friendly Explanation:** DisGeNET is a database that collects information about which genes are linked to which diseases, drawing from many different sources including expert-curated databases, genetic studies, animal



experiments, and automated reading of scientific papers. It covers both rare inherited diseases and common complex diseases like diabetes and heart disease. Each gene-disease association has a score that tells you how strong the evidence is. DisGeNET is particularly useful when you have a list of genes and want to know what diseases they are associated with, or vice versa.

**Advanced Technical Explanation:** DisGeNET integrates gene-disease associations from 24 sources using a unified data model with UMLS CUIs as disease identifiers and Entrez Gene IDs as gene identifiers. The GDA (Gene-Disease Association) score is calculated as a weighted sum of evidence from different source types, with curated database evidence weighted more heavily than text mining. The VDA (Variant-Disease Association) score similarly weights evidence for variant-disease associations. DisGeNET provides semantic similarity measures for comparing disease profiles and gene profiles, enabling network medicine analyses. The `disgenet2r` R package provides a comprehensive interface for programmatic access and analysis.

#### **One Practical Workflow Example: Gene-disease enrichment analysis using DisGeNET:**

Step 1: Install the `disgenet2r` R package: `install.packages("disgenet2r")`

Step 2: Query DisGeNET for genes associated with your disease of interest: `library(disgenet2r) results <- disease2gene(disease = "C0002395", # Alzheimer's database = "CURATED", score = c(0.3, 1))`

Step 3: Extract the gene list and perform enrichment analysis against your experimental gene list.

Step 4: Alternatively, query for diseases associated with your gene list: `results <- gene2disease(gene = c("APOE", "APP", "PSEN1"), database = "ALL")`

Step 5: Visualize the gene-disease network using Cytoscape or `ggplot2`.

Step 6: Filter results by evidence type (curated only) for higher confidence associations.

## M5: MalaCards (Human Disease Database)

---

**Official Website URL:** <https://www.malacards.org>

**Resource Type:** Database / Portal

**Main Biological Domain:** Diseases / Clinical genomics

**What It Is Used For:** MalaCards is an integrated human disease database that aggregates information about diseases from dozens of data sources into unified disease cards. Each disease card provides a comprehensive overview including aliases, summaries, associated genes, drugs, publications, clinical trials, and links to specialized resources. MalaCards is used as a one-stop resource for disease information, particularly for researchers who want a broad overview of what is known about a disease from multiple perspectives.

**What Data It Contains:** MalaCards contains information on over 20,000 human diseases, integrating data from sources including OMIM, Orphanet, ClinVar, DisGeNET, DrugBank, ClinicalTrials.gov, GeneCards, and many others. Each disease card includes disease aliases, a summary, associated genes (with evidence scores), drugs and compounds, publications, clinical trials, and links to specialized databases. MalaCards is part of the GeneCards Suite.

**Main question it helps answer:** What is known about this disease from multiple data sources, including its genetic basis, associated drugs, and available clinical trials?

**Typical user:** Researcher / Clinician / Student

**Example scientific questions:**

- What genes, drugs, and clinical trials are associated with this disease?
- What are all the alternative names for this disease across different classification systems?
- What is the integrated evidence for the genetic basis of this disease?

**Example use cases:**

- Getting a rapid overview of a disease before starting a research project.
- Finding all alternative disease names and identifiers for use in literature searches.
- Identifying drugs and compounds associated with a disease for drug repurposing analysis.

**Input data accepted:**

- Disease names (any alias)
- MalaCards disease IDs
- Gene names (to find associated diseases)

**Output data provided**

- Integrated disease cards with summaries, genes, drugs, and publications
- Disease aliases and cross-references to other databases
- Gene-disease association scores
- Links to clinical trials and specialized resources

**Strengths:**

- Comprehensive integration of information from many sources



- Useful for rapid disease overview
- Covers disease aliases and cross-references
- Links to clinical trials and drug information
- Part of the well-maintained GeneCards Suite

**Limitations:**

- Integration of many sources means variable evidence quality
- Not a primary data source; all information is aggregated from other databases
- Full access to some features requires a subscription
- Less rigorous than OMIM or ClinGen for clinical variant interpretation

**Common beginner mistakes:**

- Treating MalaCards as a primary data source rather than an aggregator
- Not following links to primary sources to verify information

**When to Use It:** Use MalaCards when you need a rapid, comprehensive overview of a disease from multiple perspectives, or when you need to find disease aliases and cross-references. It is a good starting point for disease research.

**When NOT to Use It:** For rigorous clinical variant interpretation, use ClinVar and ClinGen. For detailed Mendelian disease genetics, OMIM is more authoritative. MalaCards is best used as a discovery and navigation tool, not as a primary evidence source.

**Related databases / alternatives:**

**OMIM:** more authoritative for Mendelian disease genetics

**Orphanet:** more detailed for rare diseases

**DisGeNET:** more comprehensive for gene-disease associations

**Open Targets:** drug target and disease association platform

**How it connects to other resources:**

MalaCards integrates data from OMIM, Orphanet, ClinVar, DisGeNET, DrugBank, GeneCards, and many other sources. It is part of the GeneCards Suite and links to GeneCards for gene information.

**API / FTP / programmatic access:**

**MalaCards API:** available with subscription

**GeneCards Suite API:** <https://www.genecards.org/Guide/Api>

**Evidence/curation level:** Aggregated from multiple sources with varying curation levels; not independently curated.

**Data Update Status:** Regularly updated as source databases are updated.

**Licensing / access restrictions:** Free access for basic use; full API access and some features require subscription.

**Citation / Recommended Reference:** Rappaport N, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017;45(D1):D877-D887. doi:10.1093/nar/gkw1012

**Beginner-Friendly Explanation:** MalaCards is like a comprehensive disease encyclopedia that pulls together information from dozens of different databases into one place. For each disease, it shows you what genes are associated with it, what drugs are used to treat it, what clinical trials are ongoing, and what other names the disease goes by. It is a great starting point when you want to quickly learn about a disease from many different angles. However, because it aggregates information from many sources, you should always follow the links to the original databases to verify important information.

**Advanced Technical Explanation:** MalaCards uses a disease-centric data integration approach, mapping disease entities from multiple sources to a unified disease ontology using automated and manual curation. Gene-disease associations are scored using a composite score that weights evidence from different source types. The GeneCards Suite infrastructure provides the underlying data integration and API framework. MalaCards disease IDs (e.g., MALA:000001) are stable identifiers that can be used for cross-database integration.

### **One practical workflow example:**

#### **Using MalaCards for disease research overview:**

Step 1: Go to <https://www.malacards.org> and search for your disease of interest by name.

Step 2: Review the disease card for a summary, aliases, and key statistics.

Step 3: Check the "Genes" section for associated genes with evidence scores.

Step 4: Review the "Drugs & Compounds" section for therapeutic associations.

Step 5: Check "Clinical Trials" for ongoing studies.

Step 6: Follow links to primary databases (OMIM, Orphanet, ClinVar) for more detailed and authoritative information on specific aspects.

## BEGINNER EXAMPLE (Category M):

---

A medical student is preparing a presentation on a rare genetic syndrome. They start with MalaCards to get a broad overview, then go to Orphanet for detailed clinical information and prevalence data, then to OMIM for the molecular genetics and literature history. They use ClinGen to check the gene validity classification and find that the causative gene has "Definitive" evidence. They use the ORPHA code and MIM number to cross-reference with other databases.

## ADVANCED EXAMPLE (Category M):

---

A computational biologist is studying the genetic architecture of autoimmune diseases. They download DisGeNET curated associations for 20 autoimmune conditions, build a bipartite gene-disease network, and identify hub genes shared across multiple diseases. They cross-reference with OMIM to distinguish Mendelian from complex disease associations and use ClinGen gene validity classifications to filter for high-confidence gene-disease pairs. They then perform pathway enrichment analysis on the shared hub genes.

## CONFUSION POINTS (Category M):

---

OMIM vs. Orphanet: OMIM focuses on Mendelian genetics with molecular detail; Orphanet focuses on rare disease clinical information and resources.

ClinGen vs. DisGeNET: ClinGen provides rigorous evidence-graded gene-disease validity; DisGeNET provides broad associations from many sources with variable quality. ClinGen is for clinical decisions; DisGeNET is for research exploration.

Disease identifiers: MIM numbers (OMIM), ORPHA codes (Orphanet), UMLS CUIs (DisGeNET), and MalaCards IDs are all different. Use ontology mapping tools when integrating across databases.

## DECISION GUIDE (Category M):

---

For Mendelian disease genetics and molecular mechanisms: OMIM

For rare disease clinical information and resources: Orphanet

For evidence-graded gene-disease validity: ClinGen

For broad gene-disease associations including complex diseases: DisGeNET

For rapid disease overview from multiple sources: MalaCards

For clinical variant interpretation: ClinVar + ClinGen (not in this category but essential complement)

## Category N: Pathway and Systems Biology Databases

### CATEGORY OVERVIEW

Pathway and systems biology databases organize biological knowledge at the level of molecular interactions, biochemical reactions, and cellular processes, moving beyond individual genes and proteins to describe how they work together in coordinated networks. These databases are essential for interpreting the results of high-throughput experiments such as RNA-seq, proteomics, and metabolomics, where the goal is not just to identify which genes or proteins change, but to understand what biological processes are affected. Pathway enrichment analysis — testing whether genes in a pathway are over-represented in a list of differentially expressed genes — is one of the most commonly performed analyses in genomics and relies entirely on the quality and coverage of pathway databases.

The major pathway databases differ significantly in their scope, curation approach, and biological focus. KEGG (Kyoto Encyclopedia of Genes and Genomes) provides manually drawn pathway maps covering metabolism, signaling, and disease, with a strong emphasis on metabolic pathways and cross-species coverage. Reactome provides highly detailed, literature-curated reaction-level pathway descriptions with explicit molecular participants and evidence codes, with a focus on human biology. BioCyc provides organism-specific metabolic pathway databases with particular depth for microbial metabolism. WikiPathways is a community-curated resource where researchers contribute and maintain pathway diagrams. STRING focuses on protein-protein interaction networks rather than traditional pathways.

Each database has distinct strengths that make it more or less appropriate for different analytical contexts. A critical consideration when using pathway databases for enrichment analysis is the choice of gene set collection and the statistical method used. Different pathway databases use different gene identifiers, different pathway boundaries, and different levels of granularity. The same biological process may be represented as a single large pathway in one database and as multiple smaller sub-pathways in another, affecting the statistical power and interpretation of enrichment results. Tools such as clusterProfiler (R/Bioconductor), GSEA, and Enrichr provide interfaces to multiple pathway databases, but researchers must understand the underlying database characteristics to interpret results correctly. It is generally recommended to use multiple pathway databases and compare results, as different databases may highlight different aspects of the same biological response.

## N1: KEGG (Kyoto Encyclopedia of Genes and Genomes)

**Official Website URL:** <https://www.kegg.jp>

**Resource Type:** Knowledgebase / Database

**Main Biological Domain:** Pathways / Systems biology

**What It Is Used For:** KEGG is a comprehensive bioinformatics resource for understanding high-level functions and utilities of biological systems from molecular-level information. It is used for pathway analysis and enrichment analysis of genomics and transcriptomics data, metabolic pathway reconstruction for newly sequenced organisms, drug target identification, and understanding disease mechanisms at the pathway level. KEGG pathway maps are among the most widely used resources in bioinformatics for visualizing and interpreting biological processes.

**What Data It Contains:** KEGG contains multiple integrated databases: KEGG PATHWAY (manually drawn pathway maps for metabolism, signaling, and disease), KEGG GENES (gene catalogs for sequenced organisms), KEGG COMPOUND (chemical compounds), KEGG REACTION (biochemical reactions), KEGG DRUG (drug information), KEGG DISEASE (disease information), and KEGG ORTHOLOGY (KO system for cross-species functional annotation). KEGG covers over 500 organisms and provides organism-specific pathway reconstructions.

**Main question it helps answer:** What metabolic and signaling pathways are represented in this gene list, and how do these genes function together in biological systems?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What KEGG pathways are significantly enriched in my list of differentially expressed genes?
- What is the complete metabolic pathway for the biosynthesis of this compound in this organism?
- What KEGG ORTHOLOGY (KO) terms can I assign to the genes in my newly sequenced genome?

**Example use cases:**

- Performing KEGG pathway enrichment analysis on RNA-seq differential expression results using clusterProfiler or GSEA.
- Reconstructing the metabolic pathways present in a newly sequenced bacterial genome using KEGG ORTHOLOGY assignments.
- Visualizing a pathway of interest using the KEGG pathway viewer with gene expression data overlaid.

**Input Data Accepted:** KEGG accepts gene identifiers such as KEGG gene IDs, Entrez Gene IDs, and UniProt accession numbers; KEGG pathway identifiers (e.g., hsa04110 for the human cell cycle pathway); KEGG Orthology (KO) identifiers; and chemical compound identifiers.

**Output Data Provided:** KEGG provides pathway maps with gene and compound annotations, KEGG Orthology (KO) assignments, enrichment analysis outputs through the KEGG API or compatible software tools, organism-specific pathway reconstructions, and integrated drug–target and disease–gene association information.

**Strengths:** KEGG offers comprehensive coverage of metabolic and signaling pathways across hundreds of organisms and is widely used in genomics, transcriptomics, and systems biology research. Its manually drawn pathway maps are visually intuitive and easy to interpret. The KEGG Orthology (KO) framework enables pathway comparisons across species by assigning homologous genes to shared functional categories. KEGG also integrates



pathway information with drug, disease, and chemical compound data and is supported by many widely used enrichment and visualization tools such as clusterProfiler, GSEA, Enrichr, and pathview.

**Limitations:** Programmatic access to KEGG has licensing and usage restrictions, with commercial use requiring a subscription and some limitations applying to academic use. Because pathways are manually curated and drawn, they may not always reflect the most recent scientific findings. KEGG pathway representations are often less detailed at the reaction level compared with databases such as Reactome, and some pathways simplify or merge distinct biological processes. Additionally, many analyses require mapping standard gene identifiers to KEGG-specific IDs before use.

**Common Beginner Mistakes:** A common mistake is failing to map gene identifiers to KEGG-compatible IDs before pathway enrichment analysis. Users may also overlook the fact that KEGG pathways are manually curated representations and should not be assumed to be exhaustive or fully up to date. Another frequent misunderstanding is confusing KEGG pathway maps, which are visual representations, with KEGG gene sets used for statistical enrichment analysis. Programmatic users also often underestimate API access restrictions and licensing considerations.

**When to Use It:** Use KEGG for pathway enrichment analysis of genomics data, metabolic pathway reconstruction, and cross-species pathway comparison using the KEGG ORTHOLOGY system. It is particularly strong for metabolic pathways and provides excellent visual pathway maps.

**When NOT to Use It:** For detailed reaction-level pathway information with explicit molecular participants and literature evidence, Reactome is more appropriate. For protein-protein interaction networks, STRING is better suited.

**Related databases / alternatives:** Related resources include Reactome for detailed human signaling and reaction-level pathways, WikiPathways for community-curated pathway information, BioCyc for organism-specific metabolic pathways, and MetaCyc for highly curated metabolic pathway data.

**How It Connects to Other Resources:** KEGG links to UniProt, NCBI Gene, PubChem, and ChEBI for molecular information. KEGG ORTHOLOGY assignments are used to map genes from any organism to KEGG pathways. KEGG is integrated into R packages (clusterProfiler, KEGGREST, pathview) and web tools (DAVID, Enrichr).

**API / FTP / programmatic access:** KEGG provides REST-based access through [KEGG REST API](#). Example pathway retrieval can be performed using [KEGG human cell cycle pathway example \(hsa04110\)](#). Programmatic analysis is supported through the KEGGREST R package for API access, clusterProfiler for enrichment analysis, and pathview for pathway visualization with expression data. Users should note that KEGG API usage is subject to licensing and access limitations, particularly for commercial applications.

**Evidence/curation level:** Manually curated pathway maps; literature-based; regularly updated by KEGG curators.

**Data Update Status:** Updated regularly; KEGG releases new versions periodically. Check the KEGG website for current release information.

**Licensing / access restrictions:** Academic use of the KEGG website is free. The KEGG API has usage limitations for academic users and requires a license for commercial use. FTP access to KEGG data requires a subscription.

**Citation / Recommended Reference:** Kanehisa M, et al. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587-D592. doi:10.1093/nar/gkac963

**Beginner-Friendly Explanation:** KEGG is a large database that maps out the molecular "circuits" of living cells — the pathways by which cells carry out metabolism, respond to signals, and perform other biological functions. It contains hundreds of pathway diagrams that show how genes and molecules work together, covering everything from how cells break down glucose to how they respond to growth signals. When you have a list of genes that are turned on or off in an experiment, KEGG can tell you which biological pathways those genes belong to, helping you understand what biological processes are affected.

**Advanced Technical Explanation:** KEGG pathways are represented as KGML (KEGG Markup Language) files encoding nodes (genes, compounds, maps) and edges (interactions, reactions) with graphical coordinates for rendering. The KEGG ORTHOLOGY (KO) system assigns functional identifiers to genes based on sequence similarity and functional equivalence, enabling cross-species pathway mapping. KEGG BRITE provides hierarchical classifications of genes, compounds, and reactions. The KEGG REST API supports retrieval of pathway information, gene-pathway mappings, and compound data in text format. The pathview R package renders KEGG pathway maps with user-supplied expression data overlaid on gene nodes.

**One Practical Workflow Example: KEGG pathway enrichment analysis of RNA-seq results:**

- Step 1: Obtain your list of differentially expressed genes with Entrez Gene IDs or gene symbols.
- Step 2: Install clusterProfiler in R: `BiocManager::install("clusterProfiler")`
- Step 3: Map gene symbols to Entrez IDs if needed: `library(org.Hs.eg.db) gene_ids <- bitr(gene_symbols, fromType="SYMBOL", toType="ENTREZID", OrgDb=org.Hs.eg.db)`
- Step 4: Run KEGG enrichment analysis: `library(clusterProfiler) kegg_results <- enrichKEGG(gene = gene_ids$ENTREZID, organism = "hsa", pvalueCutoff = 0.05)`
- Step 5: Visualize results: `dotplot(kegg_results, showCategory=20)`
- Step 6: Visualize specific pathways with expression data using pathview: `library(pathview) pathview(gene.data = fold_changes, pathway.id = "hsa04110", species = "hsa")`

## N2: Reactome

**Official Website URL:** <https://reactome.org>

**Resource Type:** Knowledgebase / Database

**Main Biological Domain:** Pathways / Systems biology

**What It Is Used For:** Reactome is a free, open-source, curated and peer-reviewed pathway database that provides detailed, reaction-level descriptions of biological processes in humans and other organisms. It is used for pathway enrichment analysis, understanding the molecular details of biological processes, identifying pathway members for a gene of interest, and performing network analysis of biological systems.

**What Data It Contains:** Reactome contains over 15,000 human reactions organized into over 2,500 pathways (as of 2024), covering metabolism, signaling, gene expression, DNA repair, cell cycle, immune system, and many other processes. Each reaction is described with explicit input and output molecules, catalysts, regulators, and literature references. Reactome also provides ortholog projections for over 20 non-human species.

**Main question it helps answer:** What are the detailed molecular reactions and participants in this biological pathway, and which pathways are enriched in my gene list?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What are all the molecular participants and reactions in the MAPK signaling pathway?
- Which Reactome pathways are significantly enriched in my proteomics data?
- What upstream regulators and downstream effectors are connected to this protein in Reactome?

**Example use cases:**

- Performing Reactome pathway enrichment analysis on RNA-seq or proteomics data using the Reactome pathway analysis tool or R packages.
- Exploring the detailed molecular mechanism of a signaling pathway using the Reactome pathway browser.
- Downloading Reactome gene sets for use in GSEA analysis.

**Input Data Accepted:** Reactome accepts gene and protein identifiers including UniProt accessions, Ensembl IDs, Entrez Gene IDs, and gene symbols, as well as Reactome pathway identifiers in R-HSA format, gene expression datasets for pathway analysis, and identifier lists for pathway enrichment studies.

**Output Data Provided:** Reactome provides detailed pathway diagrams with reaction-level biological information, enrichment analysis results including p-values and false discovery rate (FDR), downloadable gene set files in GMT format for Gene Set Enrichment Analysis (GSEA), pathway hierarchy and relationship information, and API outputs in JSON format suitable for computational workflows.

**Strengths:** Reactome is widely recognized for providing some of the most detailed reaction-level pathway descriptions available. Pathways explicitly identify molecular participants and are supported by literature evidence codes. The database is fully open access without major usage restrictions and includes a highly interactive pathway browser for exploration and visualization. Its hierarchical organization enables analysis at multiple biological scales, from broad biological systems to highly specific molecular events, and ortholog projection methods extend

pathway interpretation to multiple species. Reactome is particularly strong for human signaling and mechanistic pathway analysis.

**Limitations:** Reactome is primarily centered on human biology, and while ortholog-based projections exist, non-human pathway coverage is less comprehensive. Metabolic pathway representation is generally less extensive than KEGG. Pathway definitions can be somewhat subjective, with some biological processes divided across multiple related pathways. For beginners, the extensive hierarchy and detailed structure may initially feel complex or difficult to navigate.

**Common Beginner Mistakes:** Users frequently analyze only top-level Reactome pathways and overlook the hierarchical structure, thereby missing biologically meaningful subpathway detail. Another common oversight is failing to download GMT gene set files needed for GSEA workflows. Confusing Reactome pathway identifiers (R-HSA-XXXXXX) with KEGG pathway identifiers is also a frequent source of error.

**When to Use It:** Use Reactome when you need detailed, reaction-level pathway information with explicit molecular participants and literature evidence. It is particularly strong for human signaling pathways and is the preferred database when mechanistic detail is important.

**When NOT to Use It:** For metabolic pathway analysis across many organisms, KEGG provides better coverage. For protein-protein interaction networks, STRING is more appropriate. For community-curated pathways with broad coverage, WikiPathways may complement Reactome.

**Related databases / alternatives:** Related pathway and interaction resources include KEGG for broad organism and metabolic pathway coverage, WikiPathways for community-curated pathways, BioCyc for organism-specific metabolic databases, and STRING for protein interaction networks.

**How It Connects to Other Resources:** Reactome links to UniProt, Ensembl, ChEBI, PubChem, and PubMed. Reactome pathway IDs are used in UniProt cross-references and in many enrichment analysis tools. The ReactomePA R package provides programmatic access to Reactome enrichment analysis.

**API / FTP / programmatic access:** Reactome provides REST-based access through [Reactome Content Service API](#). Example retrieval of pathway events is available at [Reactome pathway API example](#). Downloadable files and datasets are available through [Reactome downloads and FTP](#), including ReactomePathways.gmt.zip for GSEA workflows. Programmatic analysis is supported through ReactomePA in R/Bioconductor and Python workflows using REST API requests.

**Evidence/curation level:** Manually curated by expert biologists; each reaction has literature references and evidence codes; peer reviewed.

**Data Update Status:** Updated quarterly; current version available at [reactome.org](https://reactome.org).

**Licensing / access restrictions:** Fully open access under Creative Commons Attribution 4.0 (CC BY 4.0). All data freely downloadable and reusable.

**Citation / Recommended Reference:** Gillespie M, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D1):D687-D692. doi:10.1093/nar/gkab1028

**Beginner-Friendly Explanation:** Reactome is a detailed, carefully checked database of biological pathways — the step-by-step molecular processes that happen inside cells. Unlike some pathway databases that show simplified

diagrams, Reactome describes each individual reaction with the exact molecules involved and cites the scientific papers that support each step. This makes it very reliable but also very detailed. When you have a list of genes from an experiment, Reactome can tell you which biological processes those genes are involved in, and you can then explore those processes in detail using Reactome's interactive pathway browser.

**Advanced Technical Explanation:** Reactome uses a formal data model where pathways are composed of reactions, each with defined inputs, outputs, catalysts, and regulators represented as PhysicalEntity objects (proteins, complexes, small molecules). The hierarchical pathway structure uses Event objects (Pathway, ReactionLikeEvent) with parent-child relationships. Each reaction is annotated with literature references and evidence codes. The Reactome pathway analysis tool uses a binomial test with Benjamini-Hochberg FDR correction for enrichment analysis. The GMT file format provides pathway gene sets compatible with GSEA and other enrichment tools. Ortholog projections use Ensembl compara data to map human reactions to other species.

**One Practical Workflow Example:** Reactome pathway enrichment analysis:

- Step 1: Go to <https://reactome.org/PathwayBrowser/#TOOL=AT> and upload your gene list (UniProt IDs, Ensembl IDs, or gene symbols).
- Step 2: Review the enrichment results, which show pathways ranked by p-value with the number of submitted genes in each pathway.
- Step 3: Click on enriched pathways to explore them in the pathway browser.
- Step 4: For programmatic analysis, use ReactomePA in R: `BiocManager::install("ReactomePA")`  
`library(ReactomePA)`  
`results <- enrichPathway(gene = entrez_ids, organism = "human", pvalueCutoff = 0.05, readable = TRUE)`  
`dotplot(results)`
- Step 5: Download the GMT file for GSEA analysis: `wget https://reactome.org/download/current/ReactomePathways.gmt.zip`
- Step 6: Run GSEA with the Reactome GMT file and your ranked gene list.

### N3: BioCyc (Collection of Pathway/Genome Databases) NOTE: BioCyc has tiered access; some features and databases require a subscription. EcoCyc (E. coli) and MetaCyc (metabolic pathways) have more open access.

**Official Website URL:** <https://biocyc.org>

**Resource Type:** Knowledgebase / Database

**Main Biological Domain:** Pathways / Systems biology

**What It Is Used For:** BioCyc is a collection of organism-specific Pathway/Genome Databases (PGDBs) that integrate genome sequence data with metabolic and regulatory pathway information. It is used for metabolic pathway analysis in specific organisms, metabolic flux analysis, comparative genomics of metabolic capabilities, and understanding gene function in the context of metabolic networks. BioCyc is particularly valuable for microbial genomics and metabolic engineering research.

**What Data It Contains:** BioCyc contains over 20,000 organism-specific databases, each integrating the genome sequence with metabolic pathways, enzymes, reactions, compounds, and regulatory information. The flagship databases include EcoCyc (E. coli K-12, highly curated), HumanCyc (human metabolism), and MetaCyc (curated metabolic pathways from all organisms). Each database provides metabolic pathway maps, enzyme-reaction associations, compound structures, and regulatory network information.

**Main question it helps answer:** What metabolic pathways are present in this organism, and how are they organized and regulated?

**Typical user:** Researcher (microbiology, metabolic engineering) / Bioinformatician

#### Example scientific questions:

- What metabolic pathways are predicted to be present in this newly sequenced bacterial genome?
- What is the complete biosynthetic pathway for this compound in E. coli?
- How does the metabolic network of this pathogen differ from a related non-pathogenic species?

#### Example use cases:

- Using the Pathway Tools software to build a metabolic model for a newly sequenced organism.
- Comparing metabolic pathway content across multiple bacterial genomes using BioCyc's comparative genomics tools.
- Performing metabolic flux analysis using the BioCyc metabolic network as a constraint-based model.

**Input Data Accepted:** BioCyc accepts organism names or genome identifiers, gene names, enzyme names, compound names, and pathway names or BioCyc-specific pathway identifiers.

**Output Data Provided:** BioCyc provides organism-specific metabolic pathway maps, enzyme–reaction–compound associations, regulatory network information, comparative pathway analysis results, and downloadable pathway datasets in multiple formats suitable for computational analysis.

**Strengths:** BioCyc offers some of the most comprehensive organism-specific metabolic pathway databases available. EcoCyc, the Escherichia coli component of BioCyc, is among the most detailed and extensively curated biological databases for any organism. MetaCyc serves as one of the most comprehensive manually curated



references for metabolic pathways across diverse species. The integrated Pathway Tools software enables automated metabolic model reconstruction and pathway prediction, making BioCyc particularly valuable for microbial systems biology and metabolic engineering applications.

**Limitations:** BioCyc operates under a tiered access model, with some databases and advanced features requiring subscription access. While EcoCyc and MetaCyc receive extensive manual curation, many organism-specific BioCyc databases are computationally predicted rather than experimentally curated. BioCyc is generally less comprehensive for signaling pathways compared with Reactome, and the Pathway Tools software environment can present a steep learning curve for new users.

**Common Beginner Mistakes:** A common issue is failing to recognize BioCyc's tiered access model and expecting unrestricted access to subscription-only resources. Users may also incorrectly assume that all BioCyc databases are curated to the same standard as EcoCyc, despite most organism-specific databases relying heavily on computational prediction. Another frequent mistake is using BioCyc for signaling pathway analysis, where Reactome or KEGG are often more appropriate choices.

**When to Use It:** Use BioCyc when you need organism-specific metabolic pathway information, particularly for microbial organisms. EcoCyc is the best resource for *E. coli* metabolism. MetaCyc is the best reference for curated metabolic pathways across all organisms.

**When NOT to Use It:** For signaling pathways in human biology, Reactome or KEGG are more appropriate. For pathway enrichment analysis of genomics data, KEGG or Reactome are more commonly used and better supported by analysis tools.

**Related databases / alternatives:** Related resources include KEGG for broad metabolic and signaling pathway coverage, MetaCyc as the curated metabolic reference component of BioCyc, Reactome for detailed human signaling pathways, and BRENDA for enzyme-specific information.

**How It Connects to Other Resources:** BioCyc links to UniProt, NCBI Gene, PubChem, and ChEBI. MetaCyc is integrated into KEGG and other metabolic databases. Pathway Tools software uses BioCyc databases for metabolic model construction.

**API / FTP / programmatic access:** BioCyc provides web services and API access through [BioCyc web services and API](#) documentation, although some capabilities require subscription access. Pathway Tools software is downloadable for academic use, and MetaCyc data downloads are available with registration.

**Evidence/curation level:** Mixed: EcoCyc and MetaCyc are highly manually curated; most other BioCyc databases are computationally predicted using Pathway Tools.

**Data Update Status:** Updated regularly; EcoCyc and MetaCyc updated continuously.

**Licensing / access restrictions:** Tiered access: basic web access is free; full database access and API require subscription. Academic licenses available. Pathway Tools software free for academic use.

**Citation / Recommended Reference:** Karp PD, et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform.* 2019;20(4):1085-1093. doi:10.1093/bib/bbx085

**Beginner-Friendly Explanation:** BioCyc is a collection of databases that describe the metabolic pathways of specific organisms — essentially, the complete set of chemical reactions that an organism can perform. The most



detailed database in the collection is EcoCyc, which describes the metabolism of the bacterium *E. coli* in extraordinary detail. BioCyc also includes MetaCyc, which catalogs metabolic pathways from all organisms and is useful as a reference. Note that while some parts of BioCyc are freely accessible, full access to all features requires a subscription.

**Advanced Technical Explanation:** BioCyc databases are built using the Pathway Tools software, which implements a frame-based knowledge representation system (the Pathway/Genome Database ontology) for encoding metabolic and regulatory networks. EcoCyc is manually curated from the primary literature with explicit evidence codes for each assertion. Computationally predicted BioCyc databases are generated by PathoLogic, which uses MetaCyc as a reference to predict pathway presence based on enzyme homology. The BioCyc API supports SPARQL queries and programmatic retrieval of pathway, gene, and compound data.

**One Practical Workflow Example: Using MetaCyc to find metabolic pathways for a compound:**

- Step 1: Go to <https://metacyc.org> and search for your compound of interest (e.g., "L-tryptophan biosynthesis").
- Step 2: Review the pathway page showing all reactions, enzymes, and intermediates.
- Step 3: Check which organisms have this pathway using the "Taxonomic Distribution" feature.
- Step 4: For *E. coli* specifically, go to EcoCyc (<https://ecocyc.org>) for detailed, curated information about the pathway.
- Step 5: Download the pathway data for use in metabolic modeling.
- Step 6: Use Pathway Tools to build a genome-scale metabolic model for your organism of interest.

## N4: WikiPathways

**Official Website URL:** <https://www.wikipathways.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Pathways / Systems biology

**What It Is Used For:** WikiPathways is an open, collaborative platform for biological pathway curation, where researchers contribute and maintain pathway diagrams using a wiki-style editing system. It is used for pathway enrichment analysis, pathway visualization, and as a community resource for sharing pathway knowledge. WikiPathways is particularly valuable for pathways not well covered by other databases, including disease-specific pathways, drug mechanism pathways, and pathways from non-model organisms.

**What Data It Contains:** WikiPathways contains over 3,000 pathways (as of 2024) for over 30 species, contributed and maintained by the research community. Each pathway is represented as a diagram with gene/protein nodes, interactions, and annotations. Pathways are available in multiple formats (GPML, GMT, SVG) and are linked to gene identifiers from multiple databases. WikiPathways covers signaling, metabolic, regulatory, and disease pathways.

**Main question it helps answer:** What community-curated pathways are available for this biological process, and which pathways are enriched in my gene list?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- Is there a WikiPathways pathway for this specific disease mechanism that I can use for enrichment analysis?
- What genes are included in the WikiPathways representation of this signaling pathway?
- How can I contribute a new pathway to WikiPathways for a process not covered by other databases?

**Example use cases:**

- Using WikiPathways gene sets for enrichment analysis of genomics data, particularly for pathways not in KEGG or Reactome.
- Downloading WikiPathways GMT files for GSEA analysis.
- Contributing a new pathway diagram for a recently characterized biological process.

**Input Data Accepted:** WikiPathways accepts gene identifiers from multiple identifier systems, pathway names or WikiPathways accession identifiers in WP format, and general search queries for pathway discovery and retrieval.

**Output Data Provided:** WikiPathways provides pathway diagrams in GPML, SVG, and PNG formats, downloadable gene sets in GMT format for enrichment analysis, pathway metadata and associated gene lists, and API responses in JSON or XML formats suitable for computational workflows.

**Strengths:** WikiPathways is a community-curated pathway resource that often includes pathways absent from other databases. It is fully open access with no licensing restrictions and supports multiple species. An active contributor community enables regular pathway updates and expansion. GMT files are readily available for enrichment analysis, and pathway diagrams are editable, allowing researchers to improve or extend pathway

knowledge collaboratively. WikiPathways is particularly useful for disease-specific pathways, drug mechanism pathways, and non-model organism biology.

**Limitations:** The quality and completeness of WikiPathways entries can vary substantially because of community-based curation. The overall pathway collection is smaller than KEGG or Reactome, and some pathways may be outdated or incomplete. Evidence tracking and curation standards are generally less formal and less rigorous than those used in Reactome.

**Common Beginner Mistakes:** A frequent mistake is failing to review the pathway's curation status and last update date before relying on it for analysis. Another common issue is treating WikiPathways as a standalone pathway resource without cross-checking findings against KEGG or Reactome for broader coverage and validation.

**When to Use It:** Use WikiPathways as a complement to KEGG and Reactome, particularly for disease-specific pathways, drug mechanism pathways, or pathways from non-model organisms. It is also the appropriate resource when you want to contribute a new pathway to the community.

**When NOT to Use It:** For the most rigorously curated pathway information, Reactome is preferable. For metabolic pathway analysis across many organisms, KEGG is more comprehensive.

**Related databases / alternatives:** Comparable resources include KEGG for extensive metabolic and organism-wide pathway coverage, Reactome for detailed reaction-level pathways and mechanistic evidence, and BioCyc for organism-specific metabolic pathway databases.

**How It Connects to Other Resources:** WikiPathways links to Ensembl, UniProt, ChEBI, and PubChem for molecular identifiers. WikiPathways data is integrated into enrichment analysis tools (Enrichr, clusterProfiler) and the Cytoscape WikiPathways app. The WikiPathways REST API provides programmatic access to pathway data.

**API / FTP / programmatic access:** WikiPathways offers programmatic access through the [WikiPathways REST API](#). GMT gene set downloads are available at [WikiPathways GMT](#) downloads, while pathway diagrams in GPML format can be obtained from [WikiPathways GPML downloads](#). Programmatic access is also supported through the rWikiPathways R package and Python workflows using REST API requests.

**Evidence/curation level:** Community-curated; quality varies by pathway and curator; no formal evidence-grading system.

**Data Update Status:** Continuously updated by the community; monthly data releases available for download.

**Licensing / access restrictions:** Fully open access under Creative Commons Attribution 4.0 (CC BY 4.0). All pathway data freely downloadable and reusable.

**Citation / Recommended Reference:** Martens M, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021;49(D1):D613-D621. doi:10.1093/nar/gkaa1024

**Beginner-Friendly Explanation:** WikiPathways is like Wikipedia for biological pathways — it is a community-maintained database where researchers from around the world contribute and update pathway diagrams. This means it often has pathways for specific diseases or processes that are not covered by other databases, because the researchers who study those processes have contributed their knowledge directly. The quality of pathways varies, so it is good to check when a pathway was last updated and by whom. WikiPathways is completely free to use and all data can be downloaded without restrictions.

**Advanced Technical Explanation:** WikiPathways uses the GPML (Graphical Pathway Markup Language) format to encode pathway diagrams, with nodes representing biological entities (genes, proteins, metabolites) and edges representing interactions and reactions. Each node is annotated with identifiers from multiple databases (Ensembl, UniProt, ChEBI, etc.) enabling cross-database integration. The WikiPathways REST API supports pathway retrieval, gene-pathway mapping, and pathway search in JSON and XML formats. GMT files are generated monthly from the current pathway collection and are compatible with GSEA, Enrichr, and other enrichment tools.

#### One Practical Workflow Example: WikiPathways enrichment analysis using clusterProfiler:

- Step 1: Download the WikiPathways GMT file for your organism: `wget https://data.wikipathways.org/current/gmt/ wikipathways-YYYYMMDD-gmt-Homo_sapiens.gmt`
- Step 2: Load the GMT file in R: `library(clusterProfiler) wp_gmt <- read.gmt("wikipathways-YYYYMMDD-gmt-Homo_sapiens.gmt")`
- Step 3: Run enrichment analysis: `results <- enricher(gene = gene_list, TERM2GENE = wp_gmt, pvalueCutoff = 0.05)`
- Step 4: Visualize results: `dotplot(results, showCategory=20)`
- Step 5: For pathways of interest, view the pathway diagram at [https://www.wikipathways.org/pathways/WP\[ID\]](https://www.wikipathways.org/pathways/WP[ID])
- Step 6: Compare results with KEGG and Reactome enrichment to identify consistent findings.

## N5 – STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) — Cross-reference Entry

**Official Website URL:** <https://string-db.org>

**Entry type:** Cross-reference entry — full database card provided in Category O, O1.

STRING is mentioned in Category N because pathway and systems biology workflows often require protein interaction context. However, STRING is primarily a protein–protein interaction and functional association resource. For full coverage, see Category O: Protein–Protein Interaction Databases, O1 – STRING.

### BEGINNER EXAMPLE (Category N):

A graduate student performs RNA-seq on cells treated with a drug and identifies 200 differentially expressed genes. They run KEGG enrichment analysis using clusterProfiler and find enrichment in "MAPK signaling pathway" and "PI3K-Akt signaling pathway." They then run Reactome enrichment and find more specific sub-pathways enriched. They use STRING to visualize the interaction network of the top 50 differentially expressed proteins and identify a cluster of interacting proteins involved in cell cycle regulation.

### ADVANCED EXAMPLE (Category N):

A systems biologist is studying metabolic reprogramming in cancer. They integrate RNA-seq, proteomics, and metabolomics data from tumor samples. They use KEGG for metabolic pathway

mapping, Reactome for detailed reaction-level analysis of altered pathways, and STRING for protein interaction network analysis. They build a multi-layer network integrating pathway membership (KEGG/Reactome), protein interactions (STRING), and gene expression changes, then apply network centrality analysis to identify key regulatory nodes.

## CONFUSION POINTS (Category N):

---

KEGG vs. Reactome: Both are pathway databases but with different strengths. KEGG is better for metabolic pathways and cross-species analysis; Reactome is better for detailed human signaling pathways with reaction-level information. Use both for comprehensive analysis.

STRING interactions vs. pathways: STRING provides protein interaction networks (who interacts with whom), not traditional pathways (what reactions occur in what order). These are complementary but different types of biological knowledge.

BioCyc access: BioCyc has a tiered access model. EcoCyc and MetaCyc have more open access; other organism databases may require subscription.

WikiPathways quality: WikiPathways is community-curated, so quality varies. Always check the pathway's curation history and last update date.

## DECISION GUIDE (Category N):

---

1. For metabolic pathway analysis across many organisms: KEGG
2. For detailed human signaling pathway information: Reactome
3. For organism-specific microbial metabolic pathways: BioCyc/MetaCyc
4. For community-curated pathways including disease-specific: WikiPathways
5. For protein interaction network analysis: STRING
6. For comprehensive enrichment analysis: use KEGG + Reactome + WikiPathways and compare results
7. For network visualization: STRING + Cytoscape

## Category O: Protein–Protein Interaction Databases

### OVERVIEW

Protein–protein interactions (PPIs) are fundamental to virtually all biological processes, from signal transduction and gene regulation to metabolic coordination and immune response. Mapping the interactome—the complete network of protein interactions within a cell or organism—has become a central goal of systems biology. Dedicated PPI databases aggregate interaction data from diverse experimental methods including yeast two-hybrid (Y2H) screens, affinity purification coupled to mass spectrometry (AP-MS), co-immunoprecipitation, and proximity labeling approaches, as well as from computational prediction methods based on sequence co-evolution, domain interactions, and text mining of the scientific literature.

The PPI database landscape is characterized by a tension between breadth and quality. Databases like STRING integrate all available evidence—experimental, computational, and text-mined—to provide the most comprehensive coverage of potential interactions, at the cost of including many low-confidence predictions. In contrast, databases like IntAct and BioGRID focus on curating experimentally validated interactions from primary literature, providing higher confidence data at the cost of completeness. DIP and MINT represent earlier efforts in this space that have had limited updates in recent years, though their curated data remains valuable. Researchers must choose the appropriate database based on whether they prioritize coverage or confidence.

A critical consideration when working with PPI databases is the distinction between binary interactions (direct physical contact between two proteins) and co-complex membership (proteins that are part of the same complex but may not directly contact each other). Many experimental methods, particularly AP-MS, detect co-complex membership rather than direct binary interactions, and this distinction is important for interpreting network topology and inferring functional relationships. Additionally, PPI data is heavily biased toward well-studied proteins—hub proteins in the network often appear highly connected simply because they have been studied more intensively, not because they genuinely have more interaction partners than other proteins.

## O1 – STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)

**Official Website URL:** <https://string-db.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Proteins / Systems biology

**What It Is Used For:** STRING is used to retrieve and visualize protein–protein interaction networks, integrating evidence from experimental data, computational predictions, co-expression, text mining, and database imports. It is widely used for network analysis, functional enrichment, and understanding the cellular context of proteins of interest. Researchers use STRING to explore the interaction neighborhood of a protein, to identify functional modules in a network, and to perform network-based enrichment analysis.

**What Data It Contains:** STRING contains interaction data for over 67 million proteins from over 14,000 organisms, with each interaction scored on a confidence scale from 0 to 1000 based on the strength of evidence from multiple channels: neighborhood (genomic context), gene fusion, co-occurrence, co-expression, experimental data, database imports, and text mining. The database provides both direct (physical) and indirect (functional) associations.

**Main question it helps answer:** What proteins interact with my protein of interest, and what is the evidence supporting each interaction?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What proteins interact with TP53, and what functional modules do they form?
- What is the interaction network of the proteins in my differentially expressed gene list?
- Which proteins are predicted to interact based on genomic context in bacteria?

**Example use cases:**

- Visualizing the interaction network of a set of proteins from a proteomics experiment
- Performing network-based functional enrichment analysis
- Identifying hub proteins in a disease-associated interaction network

**Input Data Accepted:** Protein names, UniProt accessions, gene names, lists of proteins

**Output Data Provided:** Interaction networks (visual and tabular), confidence scores, functional enrichment results, network statistics

**Strengths:**

- Broadest coverage of organisms and proteins of any PPI database
- Integrates multiple evidence types with transparent scoring
- Excellent visualization tools
- Functional enrichment analysis built in
- Freely accessible with comprehensive API

**Limitations:**

- Includes many low-confidence and computationally predicted interactions





- Text-mined interactions may reflect co-mention rather than direct interaction
- Network is biased toward well-studied proteins
- Functional associations are not equivalent to direct physical interactions
- Default confidence threshold (0.4) may be too permissive for some analyses

**Common beginner mistakes:**

- Using the default confidence threshold (0.4) without considering whether it is appropriate
- Treating all STRING interactions as direct physical interactions
- Not distinguishing between physical and functional associations
- Interpreting network hubs as biologically important without considering study bias

**When to Use It:** Use STRING when you need a broad overview of the interaction landscape for a protein or set of proteins, when you want to perform network-based functional enrichment, or when working with non-model organisms where experimental data is limited.

**When NOT to Use It:** Do not use STRING as the sole source for high-confidence physical interactions; use IntAct or BioGRID for experimentally validated interactions. Do not treat STRING scores as probabilities of direct physical interaction.

**Related databases / alternatives:**

**BioGRID:** Curated experimental interactions

**IntAct:** High-quality curated interactions

**DIP:** Legacy curated interactions

**MINT:** Legacy curated interactions

**How It Connects to Other Resources:** STRING integrates data from BioGRID, IntAct, DIP, MINT, and other databases. Proteins are linked to UniProt, Ensembl, and NCBI Gene. Functional enrichment uses GO, KEGG, and Reactome annotations.

**API / FTP / programmatic access:** REST API at <https://string-db.org/api/>; returns TSV, JSON, or image. Python package string-db available. Bulk downloads at <https://string-db.org/cgi/download>.

**Evidence/curation level:** Mixed; experimental, computational, text-mined, and database-imported

**Data Update Status:** Regular releases; STRING v12 current; actively maintained as of 2024

**Licensing / access restrictions:** Freely available for academic use; commercial use requires license

**Citation / Recommended Reference:** Szklarczyk D et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646. <https://doi.org/10.1093/nar/gkac1000>

**Beginner-Friendly Explanation:** STRING is a database that shows you which proteins interact with each other in a cell. It collects information from many different sources—laboratory experiments, computer predictions, scientific papers, and other databases—and combines them into a single network. Each connection between proteins gets a confidence score that tells you how strong the evidence is. STRING is particularly useful for visualizing the "neighborhood" of a protein you are studying and for understanding how a group of proteins work together. It covers proteins from thousands of different organisms, making it useful for researchers working with any species.

**Advanced Technical Explanation:** STRING implements a Bayesian integration framework that combines evidence from seven channels: genomic neighborhood (conserved gene order), gene fusion events, phylogenetic co-occurrence, co-expression (from microarray and RNA-seq data), experimental interactions (from primary databases), curated pathway databases, and text mining of PubMed abstracts. Each channel contributes a prior probability of interaction, and these are combined using a naive Bayes approach to produce a combined score. The network is stored as a weighted graph, and STRING provides tools for network clustering (MCL algorithm), functional enrichment (hypergeometric test with FDR correction), and network visualization using a force-directed layout.

**One practical workflow example:**

- Step 1: Navigate to <https://string-db.org> and enter your protein of interest (e.g., TP53).
- Step 2: Select the correct organism (e.g., Homo sapiens) and click "Search."
- Step 3: In the network view, set the confidence threshold to 0.7 (high confidence) to filter for stronger interactions.
- Step 4: Click "Analysis" to perform functional enrichment analysis on the network proteins.
- Step 5: Export the interaction table (TSV format) for downstream network analysis in Cytoscape or R.
- Step 6: Use the STRING API to retrieve interactions programmatically: [https://string-db.org/api/tsv/network?identifiers=TP53&species=9606&required\\_score=700](https://string-db.org/api/tsv/network?identifiers=TP53&species=9606&required_score=700).

## O2 – BioGRID (Biological General Repository for Interaction Datasets)

**Official Website URL:** <https://thebiogrid.org>

**Resource Type:** Database / Repository

**Main Biological Domain:** Proteins / Systems biology

**What It Is Used For:** BioGRID is used to access curated protein and genetic interaction data from primary literature, providing a high-quality reference for experimentally validated interactions. It is used for network analysis, functional studies, and as a gold standard dataset for benchmarking computational interaction prediction methods. BioGRID is particularly valuable for its comprehensive coverage of genetic interactions in model organisms, especially yeast.

**What Data It Contains:** BioGRID contains over 2.4 million curated interactions from over 75,000 publications, covering protein–protein interactions, genetic interactions, and chemical interactions for over 70 organisms. Each interaction record includes the experimental method, publication reference, and interaction type. BioGRID is particularly comprehensive for *S. cerevisiae*, *S. pombe*, *C. elegans*, *D. melanogaster*, and human interactions.

**Main question it helps answer:** What experimentally validated protein–protein and genetic interactions have been reported for my protein of interest?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What proteins have been shown to physically interact with BRCA1 in experimental assays?
- What genetic interactions have been reported for my yeast gene of interest?
- What experimental methods have been used to study interactions of my protein?

**Example use cases:**

- Retrieving experimentally validated interactions for a protein of interest
- Downloading the complete human interactome for network analysis
- Using BioGRID as a gold standard for benchmarking PPI prediction algorithms

**Input Data Accepted:** Gene names, protein names, organism names, BioGRID IDs

**Output Data Provided:** Interaction records with experimental evidence, publication references, downloadable datasets

**Strengths:**

- High-quality curated experimental interactions
- Comprehensive coverage of model organism interactions
- Detailed experimental method annotations
- Freely accessible with comprehensive downloads
- Regularly updated with new publications

**Limitations:**

- Coverage biased toward well-studied organisms and proteins
- Does not include computational predictions (by design)



- Some interaction types (e.g., proximity ligation) may not reflect direct physical contact
- Curation lag means very recent publications may not be included
- Less comprehensive than STRING for non-model organisms

**Common beginner mistakes:**

- Not filtering by interaction type (physical vs. genetic) for specific analyses
- Not checking the experimental method when interpreting interaction data
- Assuming BioGRID is comprehensive for all organisms (coverage varies significantly)
- Not using the multi-validated interactions filter for high-confidence analyses

**When to Use It:** Use BioGRID when you need experimentally validated interactions with detailed evidence annotations, when working with model organisms (especially yeast), or when you need a high-quality dataset for benchmarking or network analysis.

**When NOT to Use It:** Do not use BioGRID for comprehensive interaction coverage of non-model organisms; use STRING instead. BioGRID does not include computational predictions.

**Related databases / alternatives:**

- STRING: Comprehensive interactions including predictions
- IntAct: High-quality curated interactions with molecular details
- DIP: Legacy curated interactions
- MINT: Legacy curated interactions

**How It Connects to Other Resources:** BioGRID data is imported by STRING and other databases. Proteins are cross-referenced to UniProt, Ensembl, and NCBI Gene. BioGRID participates in the IMEx consortium for data sharing with IntAct and other databases.

**API / FTP / programmatic access:** REST API at <https://webservice.thebiogrid.org/>; returns JSON or tab-delimited. Bulk downloads at <https://downloads.thebiogrid.org/>. Python package biogrid available.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available for academic use; commercial use requires license

**Citation / Recommended Reference:** Oughtred R et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200. <https://doi.org/10.1002/pro.3978>

**Beginner-Friendly Explanation:** BioGRID is a database that collects information about protein interactions that have been experimentally demonstrated in the laboratory. Unlike some databases that include computer predictions, BioGRID focuses on interactions that scientists have actually measured using techniques like yeast two-hybrid assays, co-immunoprecipitation, or mass spectrometry. For each interaction, BioGRID records which experimental method was used and which paper reported it, so you can evaluate the quality of the evidence. It is particularly comprehensive for interactions in model organisms like yeast and fruit flies.

**Advanced Technical Explanation:** BioGRID implements a comprehensive data model that captures interaction type (physical, genetic, chemical), experimental system (with controlled vocabulary for over 40 experimental methods), throughput (low-throughput vs. high-throughput), modification (post-translational modifications affecting the interaction), and phenotype (for genetic interactions). The genetic interaction data from BioGRID, particularly the yeast genetic interaction network from the Boone and Andrews labs, represents the most comprehensive genetic interaction dataset for any organism. BioGRID participates in the IMEx (International Molecular Exchange) consortium, sharing data with IntAct, MINT, and other databases using the PSI-MI XML format.

**One practical workflow example:**

- Step 1: Navigate to <https://thebiogrid.org> and search for your protein of interest.
- Step 2: Filter by "Interaction Type" to select physical interactions only.
- Step 3: Filter by "Experimental System" to focus on specific methods (e.g., "Co-immunoprecipitation").
- Step 4: Download the interaction table for your protein.
- Step 5: Use the multi-validated filter to identify interactions supported by multiple independent experiments.
- Step 6: Cross-reference with STRING to identify additional predicted interactions not yet experimentally validated.

## O3 – IntAct

**Official Website URL:** <https://www.ebi.ac.uk/intact>

**Resource Type:** Database / Repository

**Main Biological Domain:** Proteins / Systems biology

**What It Is Used For:** IntAct is used to access high-quality, manually curated molecular interaction data from primary literature, providing detailed annotations of protein–protein, protein–nucleic acid, and protein–small molecule interactions. It is used for network analysis, functional studies, and as a reference for high-confidence interaction data. IntAct is particularly valued for its detailed molecular annotations and its role as a central hub for the IMEx consortium.

**What Data It Contains:** IntAct contains over 1.2 million binary interactions curated from primary literature, with detailed annotations including interaction type, experimental method, interactor roles, binding regions, mutations affecting interaction, and post-translational modifications. The database covers interactions for thousands of organisms, with a focus on human, mouse, and model organism proteins.

**Main question it helps answer:** What are the detailed molecular characteristics of the experimentally validated interactions for my protein of interest?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What are the binding domains involved in the interaction between my two proteins of interest?
- What mutations have been shown to disrupt the interaction between these proteins?
- What experimental methods have been used to validate this interaction?

**Example use cases:**

- Retrieving detailed molecular interaction data for structural biology studies
- Finding mutations that disrupt specific protein interactions
- Downloading high-quality interaction data for network analysis

**Input Data Accepted:** Protein names, UniProt accessions, gene names, IntAct IDs

**Output Data Provided:** Interaction records with detailed molecular annotations, PSICQUIC-compatible data, downloadable datasets

**Strengths:**

- Highest level of molecular detail for interaction annotations
- IMEx consortium member ensuring data quality and standardization
- PSICQUIC-compatible for programmatic access
- Covers protein–nucleic acid and protein–small molecule interactions
- Freely accessible

**Limitations:**

- Smaller dataset than BioGRID or STRING

- Curation is labor-intensive, leading to coverage gaps
- Less comprehensive for non-human organisms
- Interface can be complex for new users
- Curation lag for very recent publications

#### Common beginner mistakes:

- Not using the PSICQUIC interface for programmatic access
- Not filtering by interaction detection method
- Overlooking the binding region and mutation annotations
- Not using the complex portal for protein complex data

**When to Use It:** Use IntAct when you need detailed molecular annotations for protein interactions, when studying binding domains and interaction-disrupting mutations, or when you need high-quality data for structural biology or drug discovery.

**When NOT to Use It:** Do not use IntAct for comprehensive interaction coverage; use STRING or BioGRID for broader coverage. IntAct is best for detailed molecular characterization of specific interactions.

**Related databases / alternatives:** BioGRID: Comprehensive curated interactions; STRING: Broad coverage including predictions; Complex Portal: Protein complex data (EBI); MINT: Legacy curated interactions

**How It Connects to Other Resources:** IntAct is a member of the IMEx consortium and shares data with BioGRID, MINT, and other databases. Proteins are cross-referenced to UniProt, Ensembl, and ChEBI (for small molecules). IntAct data is imported by STRING.

**API / FTP / programmatic access:** PSICQUIC REST API at <https://www.ebi.ac.uk/Tools/webservices/psicquic/intact/webservices/current/search/>; returns MITAB or PSI-MI XML. FTP downloads at <https://ftp.ebi.ac.uk/pub/databases/intact/>. Python package pyintact available.

**Evidence/curation level:** Manually curated from primary literature; IMEx-compliant; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Orchard S et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363. <https://doi.org/10.1093/nar/gkt1115>

**Beginner-Friendly Explanation:** IntAct is a database maintained by the European Bioinformatics Institute that collects detailed information about molecular interactions—primarily between proteins, but also between proteins and DNA, RNA, or small molecules. What makes IntAct special is the level of detail it provides: for each interaction, it records not just which proteins interact, but also which parts of the proteins are involved, what experimental method was used, and whether any mutations affect the interaction. This makes it particularly valuable for researchers who want to understand the molecular details of how proteins interact.

**Advanced Technical Explanation:** IntAct implements the PSI-MI (Proteomics Standards Initiative Molecular Interactions) data model, which provides a standardized vocabulary for describing molecular interactions at the molecular level. Each interaction record captures interactor identifiers (with cross-references to UniProt, ChEBI,



etc.), interaction type (physical association, direct interaction, colocalization, etc.), experimental method (with MI ontology terms), binding regions (feature annotations with sequence ranges), mutations affecting interaction, and post-translational modifications. IntAct is a founding member of the IMEx consortium, which coordinates curation efforts across 11 molecular interaction databases to avoid duplication and ensure data quality.

#### One practical workflow example:

- Step 1: Navigate to <https://www.ebi.ac.uk/intact> and search for your protein of interest using its UniProt accession.
- Step 2: Filter interactions by "Interaction type" to select direct interactions.
- Step 3: Review the feature annotations to identify binding regions and interaction-disrupting mutations.
- Step 4: Download the interaction data in MITAB format for downstream analysis.
- Step 5: Use the PSICQUIC API for programmatic access: <https://www.ebi.ac.uk/Tools/webservices/psicquic/intact/webservices/current/search/query/P04637>.
- Step 6: Cross-reference with BioGRID to identify additional experimentally validated interactions.

## O4 – DIP (Database of Interacting Proteins)

---

**Official Website URL:** <https://dip.doe-mbi.ucla.edu>

**Resource Type:** Database

**Main Biological Domain:** Proteins / Systems biology

**What It Is Used For:** DIP is used to access a curated database of experimentally determined protein–protein interactions. NOTE: DIP has had limited updates since approximately 2017 and should be considered a legacy resource. For current interaction data, BioGRID or IntAct are recommended. DIP's historical data remains valuable for benchmarking and for interactions curated before 2017.

**What Data It Contains:** DIP contains over 80,000 interactions between over 30,000 proteins from over 200 organisms, curated from primary literature. The database focuses on binary protein–protein interactions with experimental evidence, with quality scores based on the number of independent experimental validations.

**Main question it helps answer:** What experimentally validated binary protein–protein interactions were reported in the literature up to approximately 2017?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What binary interactions have been reported for my protein of interest (historical data)?
- What is the DIP quality score for this interaction?
- What interactions are available for benchmarking PPI prediction algorithms?

**Example use cases:**

- Accessing historical interaction data for benchmarking studies
- Retrieving interactions curated before 2017 that may not be in other databases
- Using DIP as a reference dataset for computational method development

**Input Data Accepted:** Protein names, UniProt accessions, DIP IDs

**Output Data Provided:** Interaction records with experimental evidence, quality scores

**Strengths:** Historical curated interaction data; Quality scoring system for interactions; Freely accessible; Useful for benchmarking

**Limitations:** Limited updates since approximately 2017 (legacy resource); Smaller and less comprehensive than BioGRID or IntAct; May not reflect current understanding of interactions; Website availability may be intermittent; Not recommended as primary resource for current research

**Common beginner mistakes:** Using DIP as a primary interaction resource without recognizing its limited update status; Not cross-referencing with BioGRID or IntAct for current data; Not recognizing that DIP data may be outdated

**When to Use It:** Use DIP only for historical benchmarking studies or when specifically looking for interactions curated before 2017. For current research, use BioGRID or IntAct.

**When NOT to Use It:** Do not use DIP as the primary source for current protein interaction data. Use BioGRID or IntAct instead.

**Related databases / alternatives:** BioGRID: Current comprehensive curated interactions (recommended); IntAct: Current high-quality curated interactions (recommended); STRING: Comprehensive interactions including predictions

**How It Connects to Other Resources:** DIP data has been imported by STRING and other databases. Proteins are cross-referenced to UniProt and NCBI.

**API / FTP / programmatic access:** FTP downloads available from the DIP website when accessible. Limited API access.

**Evidence/curation level:** Manually curated from primary literature; limited updates since ~2017

**Data Update Status:** Limited updates since approximately 2017; legacy resource

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Salwinski L et al. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Research, 32(Database issue):D449–D451. <https://doi.org/10.1093/nar/gkh086>

**Beginner-Friendly Explanation:** DIP (Database of Interacting Proteins) is one of the original databases for protein–protein interaction data, collecting experimentally validated interactions from the scientific literature. However, it is important to know that DIP has not been actively updated since around 2017, which means it does not contain interactions discovered in recent years. For current research, databases like BioGRID or IntAct are better choices. DIP's historical data is still useful for certain purposes, such as benchmarking computational methods, but it should not be used as the primary source for up-to-date interaction information.

**Advanced Technical Explanation:** DIP implements a quality scoring system that assigns confidence scores to interactions based on the number of independent experimental validations (the "core" dataset contains interactions validated by at least two independent methods or by a single high-throughput study with additional validation). The database uses the PSI-MI data model for interaction representation. DIP's historical data has been widely used as a benchmark dataset for PPI prediction algorithms, and many published methods report performance on the DIP dataset for comparison purposes.

**One practical workflow example:**

- Step 1: Verify that <https://dip.doe-mbi.ucla.edu> is accessible.
- Step 2: If accessible, search for your protein of interest.
- Step 3: Note the DIP quality score for each interaction.
- Step 4: Download the interaction data for your protein.
- Step 5: Cross-reference all DIP interactions with BioGRID and IntAct for current validation status.
- Step 6: For current research, use BioGRID or IntAct as the primary resource.

## O5 – MINT (Molecular INTERaction database)

**Official Website URL:** <https://mint.bio.uniroma2.it>

**Resource Type:** Database

**Main Biological Domain:** Proteins / Systems biology

**What It Is Used For:** MINT is used to access a curated database of experimentally verified protein–protein interactions, with a focus on mammalian interactions. NOTE: MINT has had limited updates in recent years and should be considered a legacy resource. For current interaction data, BioGRID or IntAct are recommended.

**What Data It Contains:** MINT contains over 240,000 interactions curated from primary literature, with a focus on mammalian (particularly human and mouse) protein–protein interactions. The database uses the PSI-MI data model and is a member of the IMEx consortium.

**Main question it helps answer:** What experimentally validated mammalian protein–protein interactions were reported in the literature up to the last update?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What mammalian protein interactions are available in MINT for my protein of interest?
- What is the experimental evidence for this interaction in MINT?
- How does MINT coverage compare to IntAct for my protein?

**Example use cases:**

- Accessing historical mammalian interaction data
- Cross-referencing with IntAct for comprehensive IMEx data
- Using MINT data for benchmarking studies

**Input Data Accepted:** Protein names, UniProt accessions, MINT IDs

**Output Data Provided:** Interaction records with experimental evidence

**Strengths:** Focus on mammalian interactions; IMEx consortium member; PSI-MI compliant data model; Freely accessible

**Limitations:** Limited updates in recent years (legacy resource); Smaller dataset than BioGRID or IntAct; Not recommended as primary resource for current research; Website availability may be intermittent

**Common beginner mistakes:**

- Using MINT as a primary interaction resource without recognizing its limited update status
- Not cross-referencing with IntAct for current IMEx data

**When to Use It:** Use MINT only as a supplementary resource or for historical benchmarking. For current research, use BioGRID or IntAct.

**When NOT to Use It:** Do not use MINT as the primary source for current protein interaction data.

**Related databases / alternatives:**

- IntAct: Current IMEx-compliant interactions (recommended)

- BioGRID: Current comprehensive curated interactions (recommended)
- STRING: Comprehensive interactions including predictions

**How It Connects to Other Resources:** MINT is a member of the IMEx consortium and shares data with IntAct. Proteins are cross-referenced to UniProt.

**API / FTP / programmatic access:** PSICQUIC-compatible API when accessible. FTP downloads available.

**Evidence/curation level:** Manually curated from primary literature; limited recent updates

**Data Update Status:** Limited updates in recent years; legacy resource

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Licata L et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Research, 40(D1):D857–D861. <https://doi.org/10.1093/nar/gkr930>

**Beginner-Friendly Explanation:** MINT (Molecular INTERaction database) is a database of protein–protein interactions that was developed at the University of Rome. It focuses particularly on interactions between mammalian proteins and has been curated from the scientific literature. Like DIP, MINT has not been actively updated in recent years, so it should be considered a legacy resource. For current research, IntAct or BioGRID are better alternatives. However, MINT's historical data, particularly for mammalian interactions, can still be useful for certain analyses.

**Advanced Technical Explanation:** MINT implements the PSI-MI data model and is a member of the IMEx consortium, which means its data is shared with IntAct and other IMEx partners. The database uses controlled vocabularies from the PSI-MI ontology for experimental methods, interaction types, and interactor roles. MINT's focus on mammalian interactions made it a valuable complement to DIP in the early days of PPI databases. The IMEx data sharing means that MINT interactions are accessible through the IntAct PSICQUIC interface.

#### One practical workflow example:

- Step 1: Verify that <https://mint.bio.uniroma2.it> is accessible.
- Step 2: If accessible, search for your protein of interest.
- Step 3: Review the interaction records and experimental evidence.
- Step 4: Note that MINT data is also accessible through IntAct (as an IMEx partner).
- Step 5: For current research, use IntAct as the primary resource (which includes MINT data).
- Step 6: Cross-reference with BioGRID for additional experimentally validated interactions.

## Beginner EXAMPLE (Category O):

---

A graduate student wants to understand the protein interaction network around their gene of interest, PTEN. They navigate to STRING (<https://string-db.org>), enter "PTEN" and select "Homo sapiens." They set the confidence threshold to 0.7 and see a network of ~20 high-confidence interactors including PIK3CA, AKT1, and TP53. They click "Analysis" and find enrichment for the PI3K-Akt signaling pathway. They then check BioGRID to confirm which of these interactions have direct experimental evidence.

## ADVANCED EXAMPLE (Category O):

---

A systems biologist is building a disease-specific protein interaction network for Parkinson's disease. They download the complete human interactome from BioGRID, then supplement with high-confidence STRING interactions (score > 0.9) not present in BioGRID. They use IntAct to retrieve binding domain information for key interactions. They build the network in Cytoscape, apply the MCODE algorithm to identify functional modules, and perform GO enrichment on each module. They identify a module enriched for mitochondrial function and autophagy, consistent with known Parkinson's disease biology.

## CONFUSION POINTS (Category O):

---

STRING "interactions" include functional associations (co-expression, text co-mention) that are not physical interactions. Always check the evidence type.

BioGRID genetic interactions are not protein-protein interactions; they describe epistatic relationships between genes. IntAct "co-localization" is not the same as direct physical interaction.

High STRING confidence scores do not guarantee experimental validation.

DIP and MINT are legacy databases; their absence of an interaction does not mean the interaction does not exist.

## DECISION GUIDE (Category O):

---

Need broad network overview for any organism? → STRING; Need experimentally validated interactions with evidence details? → BioGRID; Need molecular-level interaction details (binding domains, mutations)? → IntAct; Need historical benchmarking dataset? → DIP (with caution); Need mammalian interactions from IMEx? → IntAct (includes MINT data); Need genetic interactions in yeast? → BioGRID (most comprehensive)

## Category P: Drug, Compound, and Target Databases

### OVERVIEW

Drug, compound, and target databases form the backbone of computational drug discovery and pharmacology research. These resources aggregate information about small molecules, their biological targets, mechanisms of action, pharmacokinetic properties, clinical indications, and adverse effects. The field has grown enormously with the rise of high-throughput screening, combinatorial chemistry, and the need to repurpose existing drugs for new indications. Modern drug databases integrate chemical structure information with biological activity data, clinical trial results, and regulatory approval status, providing a comprehensive view of the drug-target landscape.

The major databases in this space serve different but complementary purposes. PubChem, maintained by NCBI, is the largest open-access repository of chemical information, containing data on over 100 million compounds with biological activity data from thousands of assays. ChEMBL, maintained by the EBI, focuses on bioactive molecules with drug-like properties and provides curated binding affinity data from primary literature. DrugBank provides the most comprehensive drug-specific information including pharmacokinetics, drug interactions, and clinical data, but operates on a tiered access model. BindingDB specializes in measured binding affinities between proteins and small molecules, while TTD (Therapeutic Target Database) focuses on the target side of the drug-target relationship.

A critical consideration when using these databases is the distinction between chemical compounds (any molecule with a defined structure), bioactive compounds (molecules with measured biological activity), approved drugs (molecules with regulatory approval for clinical use), and investigational compounds (molecules in clinical trials). Different databases focus on different parts of this spectrum, and researchers must choose the appropriate resource based on their specific question. Additionally, binding affinity data from different assays ( $IC_{50}$ ,  $K_i$ ,  $K_d$ ,  $EC_{50}$ ) are not directly comparable, and the assay conditions (cell type, buffer, temperature) can significantly affect measured values.



## P1 – DrugBank

**Official Website URL:** <https://go.drugbank.com>

**Resource Type:** Knowledgebase

**Main Biological Domain:** Drugs / Clinical genomics

**What It Is Used For:** DrugBank is used to access comprehensive information about approved drugs, experimental drugs, and nutraceuticals, including their chemical properties, pharmacology, mechanisms of action, targets, enzymes, transporters, and clinical information. NOTE: DrugBank operates on a tiered access model; the free academic version provides access to most data, but some features (bulk downloads, advanced API access) require a commercial license. It is widely used in drug repurposing, pharmacogenomics, and drug interaction studies.

**What Data It Contains:** DrugBank contains detailed records for over 14,000 drug entries (including ~2,700 approved small molecule drugs, ~1,500 approved biotech drugs, and ~6,700 experimental drugs), with information on chemical structure, pharmacology, mechanism of action, pharmacokinetics (ADMET), drug-drug interactions, drug-food interactions, targets (with binding affinities where available), enzymes, transporters, and carriers. Clinical information includes indications, contraindications, and adverse effects.

**Main question it helps answer:** What is the complete pharmacological profile of this drug, including its targets, mechanism of action, and clinical properties?

**Typical user:** Researcher / Clinician / Bioinformatician / Data analyst

**Example scientific questions:**

- What are all the known targets of metformin?
- What drugs target the EGFR kinase domain?
- What are the known drug-drug interactions for warfarin?

**Example use cases:**

- Drug repurposing: finding new indications for approved drugs
- Pharmacogenomics: identifying drugs metabolized by specific CYP enzymes
- Drug interaction analysis: checking for potential interactions between co-administered drugs

**Input Data Accepted:** Drug names, DrugBank IDs, InChI keys, SMILES strings, target names

**Output Data Provided:** Comprehensive drug records, target lists, interaction data, pharmacokinetic parameters

**Strengths:**

- Most comprehensive drug-specific information available
- Covers approved drugs, experimental drugs, and nutraceuticals
- Detailed ADMET information
- Drug-drug and drug-food interaction data
- Regularly updated

**Limitations:**

- Tiered access model; some features require commercial license

- Bulk downloads require registration and may require license
- Coverage of experimental compounds less comprehensive than ChEMBL
- Some binding affinity data may be from secondary sources
- Not ideal for large-scale computational screening

#### Common beginner mistakes:

- Not recognizing the tiered access model and attempting to access restricted features
- Confusing DrugBank IDs with other identifiers (e.g., PubChem CIDs)
- Not distinguishing between approved drugs and experimental compounds
- Using DrugBank for comprehensive bioactivity data (use ChEMBL instead)

**When to Use It:** Use DrugBank when you need comprehensive pharmacological information about approved drugs, including clinical data, drug interactions, and ADMET properties. Ideal for drug repurposing and pharmacogenomics studies.

**When NOT to Use It:** Do not use DrugBank for comprehensive bioactivity data on experimental compounds; use ChEMBL or PubChem instead. Not ideal for large-scale computational screening due to access restrictions.

#### Related databases / alternatives:

- ChEMBL: Bioactivity data for drug-like molecules
- PubChem: Comprehensive chemical information
- BindingDB: Binding affinity data
- TTD: Therapeutic target information

**How It Connects to Other Resources:** DrugBank cross-references PubChem, ChEMBL, UniProt (for targets), KEGG, and clinical databases. Drug structures are linked to PubChem CIDs and ChEMBL IDs.

**API / FTP / programmatic access:** REST API at <https://go.drugbank.com/releases/latest> (requires registration); returns XML or JSON. Bulk downloads require license agreement. Python package drugbank-downloader available.

**Evidence/curation level:** Manually curated from primary literature and regulatory documents; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Tiered access; free academic version available with registration; commercial license required for bulk downloads and advanced API

**Citation / Recommended Reference:** Wishart DS et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1):D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>

**Beginner-friendly explanation:** DrugBank is like an encyclopedia for drugs. For each drug, it tells you what the drug is made of, how it works in the body, what proteins it targets, how the body processes it, and what side effects or interactions it might have. It covers thousands of approved drugs as well as experimental compounds being tested in clinical trials. DrugBank is particularly useful if you want to understand the complete pharmacological profile of a drug or if you are looking for drugs that target a specific protein. Note that while basic access is free, some advanced features require a paid license.

**Advanced Technical Explanation:** DrugBank implements a comprehensive pharmacological data model that captures drug-target interactions with binding affinity data ( $K_i$ ,  $IC_{50}$ ,  $K_d$ ), pharmacokinetic parameters (bioavailability, half-life, protein binding, volume of distribution), metabolic pathways (CYP enzyme substrates, inhibitors, inducers), transporter interactions (P-glycoprotein, OATP, etc.), and drug-drug interactions with mechanism annotations. The database uses standardized chemical identifiers (InChI, SMILES, InChIKey) and cross-references to UniProt for target proteins. DrugBank's ADMET data is particularly valuable for computational ADMET prediction model training and validation.

**One practical workflow example:**

- Step 1: Navigate to <https://go.drugbank.com> and search for your drug of interest (e.g., imatinib).
- Step 2: Review the "Pharmacology" section for mechanism of action and targets.
- Step 3: Click on each target to see binding affinity data and the target protein's UniProt entry.
- Step 4: Check the "Drug Interactions" section for known interactions with other drugs.
- Step 5: Download the drug record in XML format for computational analysis.
- Step 6: Use the DrugBank API to retrieve all drugs targeting a specific protein:  
<https://go.drugbank.com/unearth/q?searcher=drugs&query=EGFR>.

## P2 – ChEMBL

**Official Website URL:** <https://www.ebi.ac.uk/chembl>

**Resource Type:** Database / Repository

**Main Biological Domain:** Drugs / Proteins

**What It Is Used For:** ChEMBL is used to access bioactivity data for drug-like molecules, providing curated binding affinity, functional activity, and ADMET data from primary medicinal chemistry literature. It is widely used in computational drug discovery, QSAR modeling, target identification, and drug repurposing. ChEMBL is the primary resource for large-scale bioactivity data mining.

**What Data It Contains:** ChEMBL contains over 2.4 million distinct compounds with over 20 million bioactivity measurements from over 90,000 assays, curated from over 90,000 publications. Data includes binding affinities (K<sub>i</sub>, K<sub>d</sub>, IC<sub>50</sub>), functional activities (EC<sub>50</sub>, E<sub>max</sub>), ADMET properties, and clinical trial information. The database covers targets from multiple organisms.

**Main question it helps answer:** What is the measured bioactivity of this compound against its targets, and what other compounds have been tested against this target?

**Typical user:** Bioinformatician / Data analyst / Researcher

**Example scientific questions:**

- What is the IC<sub>50</sub> of this compound against EGFR?
- What compounds have been tested against BRAF, and what are their activities?
- What is the selectivity profile of this kinase inhibitor?

**Example use cases:**

- Building QSAR models for target activity prediction
- Identifying lead compounds for a new drug target
- Analyzing the selectivity profile of a compound series

**Input Data Accepted:** Compound names, SMILES, InChI keys, ChEMBL IDs, target names, UniProt accessions

**Output Data Provided:** Bioactivity data, compound structures, target information, assay details

**Strengths:**

- Largest curated bioactivity database
- Freely accessible with comprehensive API
- Standardized activity data with units and assay details
- Covers both binding and functional assays
- Regularly updated with new publications

**Limitations:**

- Data quality varies across assays and publications
- Activity values from different assays are not directly comparable
- Coverage biased toward kinases and other druggable targets

- Some assay details may be incomplete
- Not focused on clinical drug information (use DrugBank for that)

#### Common beginner mistakes:

- Comparing IC50 values from different assay types without considering assay conditions
- Not filtering by assay type (binding vs. functional) for specific analyses
- Not using the confidence score for target assignments
- Ignoring the data validity flag for activity values

**When to Use It:** Use ChEMBL when you need large-scale bioactivity data for computational drug discovery, QSAR modeling, or target-based drug discovery. Ideal for mining structure-activity relationships.

**When NOT to Use It:** Do not use ChEMBL for comprehensive clinical drug information; use DrugBank instead. ChEMBL is not ideal for approved drug pharmacology.

#### Related databases / alternatives:

- **DrugBank:** Comprehensive drug pharmacology
- **PubChem:** Comprehensive chemical information
- **BindingDB:** Binding affinity data
- **ZINC:** Virtual screening library

**How It Connects to Other Resources:** ChEMBL cross-references PubChem, UniProt (for targets), DrugBank, and clinical databases. Compound structures are linked to PubChem CIDs. Target proteins are linked to UniProt accessions.

**API / FTP / programmatic access:** REST API at <https://www.ebi.ac.uk/chembl/api/data/>; returns JSON or XML. Python client: pip install chembl-webresource-client. FTP downloads at <https://ftp.ebi.ac.uk/pub/databases/chembl/>.

**Evidence/curation level:** Manually curated from primary medicinal chemistry literature; high quality

**Data Update Status:** Regular releases (ChEMBL 34 current); actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution-ShareAlike 3.0

**Citation / Recommended Reference:** Mendez D et al. (2019). ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research, 47(D1):D930–D940. <https://doi.org/10.1093/nar/gky1075>

**Beginner-Friendly Explanation:** ChEMBL is a large database of drug-like molecules and their biological activities, maintained by the European Bioinformatics Institute. For each molecule, it records how strongly it binds to or affects specific proteins, based on measurements reported in scientific papers. This makes it invaluable for drug discovery researchers who want to understand which molecules are active against a particular target, or to build computer models that predict drug activity. ChEMBL is freely available and has an excellent programming interface, making it popular for large-scale computational analyses.

**Advanced Technical Explanation:** ChEMBL implements a comprehensive bioactivity data model that captures assay type (binding, functional, ADMET, toxicity), target type (single protein, protein complex, cell line, organism), activity type (IC50, Ki, Kd, EC50, Emax, etc.) with standardized units, and data validity flags. The

database uses a confidence scoring system for target assignments (0-9 scale) based on the directness of the assay-target relationship. ChEMBL's compound standardization pipeline normalizes SMILES representations and assigns InChIKeys for structure-based searching. The database is widely used for machine learning model training in drug discovery, particularly for QSAR and deep learning-based activity prediction.

#### One practical workflow example:

- Step 1: Navigate to <https://www.ebi.ac.uk/chembl> and search for your target of interest (e.g., EGFR).
- Step 2: Click on the target to see all compounds tested against it.
- Step 3: Filter by assay type (e.g., "Binding") and activity type (e.g., "IC50").
- Step 4: Download the bioactivity data as CSV for QSAR modeling.
- Step 5: Use the Python client to retrieve data programmatically: 

```
from chembl_webresource_client.new_client import new_client
activity = new_client.activity
res = activity.filter(target_chembl_id='CHEMBL203',
standard_type='IC50')
```
- Step 6: Filter for high-quality data (pchembl\_value >= 6, data\_validity\_comment is null).

## P3 – PubChem

**Official Website URL:** <https://pubchem.ncbi.nlm.nih.gov>

**Resource Type:** Database / Repository

**Main Biological Domain:** Drugs / Proteins

**What It Is Used For:** PubChem is used to access the world's largest freely accessible chemical information database, providing chemical structures, properties, biological activities, safety information, and literature references for over 100 million compounds. It is used for chemical structure searching, bioactivity data mining, safety assessment, and as a central hub for chemical information in biomedical research.

**What Data It Contains:** PubChem contains over 100 million unique chemical structures (Compounds), over 290 million substance records from depositors, and over 300 million bioactivity data points from over 1.5 million assays. Data includes chemical properties (molecular weight, logP, etc.), biological activities from high-throughput screening, safety and toxicity information, patent information, and literature references.

**Main question it helps answer:** What is known about the chemical properties and biological activities of this compound?

**Typical user:** Researcher / Bioinformatician / Data analyst / Beginner student

**Example scientific questions:**

- What is the structure and properties of aspirin?
- What compounds have been tested in this high-throughput screening assay?
- What is the safety profile of this industrial chemical?

**Example use cases:**

- Retrieving chemical structures and properties for a list of compounds
- Mining bioactivity data from high-throughput screening campaigns
- Checking safety and toxicity information for a compound

**Input Data Accepted:** Compound names, SMILES, InChI, CAS numbers, PubChem CIDs, SIDs

**Output Data Provided:** Chemical structures, properties, bioactivity data, safety information, literature references

**Strengths:**

- Largest freely accessible chemical database
- Comprehensive coverage of chemical space
- Excellent structure search capabilities
- Integrates data from hundreds of depositors
- Excellent API and programmatic access

**Limitations:**

- Data quality varies widely across depositors
- Bioactivity data from HTS may have high false positive rates
- Less curated than ChEMBL for drug-like molecules



- Compound records may have duplicate or inconsistent data
- Not focused on clinical drug information

#### Common beginner mistakes:

- Confusing Compound (CID) and Substance (SID) records
- Not filtering bioactivity data by assay quality
- Using PubChem for clinical drug information (use DrugBank instead)
- Not recognizing that HTS data may have high false positive rates

**When to Use It:** Use PubChem when you need comprehensive chemical information for any compound, when searching for compounds by structure, or when accessing HTS bioactivity data. Ideal as a starting point for any chemical biology question.

**When NOT to Use It:** Do not use PubChem as the primary source for curated drug pharmacology; use DrugBank or ChEMBL instead. PubChem's bioactivity data requires careful quality filtering.

**Related databases / alternatives:** ChEMBL: Curated bioactivity data for drug-like molecules; DrugBank: Comprehensive drug pharmacology; BindingDB: Binding affinity data; ChemSpider: Alternative chemical database

**How It Connects to Other Resources:** PubChem cross-references ChEMBL, DrugBank, UniProt, and hundreds of other databases. PubChem CIDs are widely used as chemical identifiers across the biomedical literature.

**API / FTP / programmatic access:** PUG REST API at <https://pubchem.ncbi.nlm.nih.gov/rest/pug/>; returns JSON, XML, CSV. PUG View API for detailed records. Python package pubchempy available. FTP downloads at <https://ftp.ncbi.nlm.nih.gov/pubchem/>.

**Evidence/curation level:** Mixed; depositor-submitted with some curation; bioactivity data from HTS campaigns

**Data Update Status:** Continuously updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; public domain

**Citation / Recommended Reference:** Kim S et al. (2023). PubChem 2023 update. Nucleic Acids Research, 51(D1):D1373–D1380. <https://doi.org/10.1093/nar/gkac956>

**Beginner-Friendly Explanation:** PubChem is a massive, freely available database of chemical compounds maintained by the US National Institutes of Health. It contains information on over 100 million different chemical structures, including their physical and chemical properties, biological activities, and safety information. PubChem is often the first place researchers go when they want to find information about a chemical compound—whether it is a drug, a natural product, an industrial chemical, or an experimental molecule. It is particularly useful for finding the structure of a compound, checking its properties, or seeing what biological activities have been reported for it.

**Advanced Technical Explanation:** PubChem implements a three-tier data model: Substances (SIDs, depositor-submitted records), Compounds (CIDs, standardized chemical structures derived from substances), and BioAssays (AIDs, biological activity data). The compound standardization pipeline normalizes SMILES representations, removes salts and solvents, and assigns InChIKeys for structure-based deduplication. PubChem's structure search uses a fingerprint-based similarity search (Tanimoto coefficient) and substructure search (SMARTS patterns). The



PUG REST API provides programmatic access to all data types, and PubChem's integration with NCBI's Entrez system allows cross-database searching.

**One practical workflow example:**

- Step 1: Navigate to <https://pubchem.ncbi.nlm.nih.gov> and search for your compound by name or structure.
- Step 2: Click on the Compound record (CID) to see the standardized structure and properties.
- Step 3: Check the "Biological Test Results" section for bioactivity data.
- Step 4: Use the PUG REST API to retrieve data programmatically:  
<https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/aspirin/JSON>
- Step 5: Download the bioactivity data for a specific assay (AID) in CSV format.
- Step 6: Cross-reference with ChEMBL for curated bioactivity data on drug-like compounds.

## P4 – BindingDB

**Official Website URL:** <https://www.bindingdb.org>

**Resource Type:** Database

**Main Biological Domain:** Drugs / Proteins

**What It Is Used For:** BindingDB is used to access measured binding affinities between proteins and small molecules, providing a curated resource for drug-target interaction data. It is particularly valuable for computational drug discovery, QSAR modeling, and understanding structure-activity relationships. BindingDB focuses specifically on quantitative binding data (K<sub>i</sub>, K<sub>d</sub>, IC<sub>50</sub>, EC<sub>50</sub>) from primary literature.

**What Data It Contains:** BindingDB contains over 2.8 million measured binding affinities for over 1.2 million small molecules against over 9,000 protein targets, curated from primary literature and patent databases. Data includes binding affinity values with units, assay conditions, and literature references.

**Main question it helps answer:** What is the measured binding affinity of this compound against this protein target?

**Typical user:** Bioinformatician / Data analyst / Researcher

**Example scientific questions:**

- What is the K<sub>i</sub> of this compound against thrombin?
- What compounds have been measured against HIV protease, and what are their affinities?
- What is the selectivity profile of this compound across related kinases?

**Example use cases:**

- Building machine learning models for binding affinity prediction
- Analyzing structure-activity relationships for a compound series
- Identifying selective compounds for a target of interest

**Input Data Accepted:** Compound names, SMILES, protein names, UniProt accessions

**Output Data Provided:** Binding affinity data, compound structures, target information

**Strengths:** Focus on quantitative binding data; Curated from primary literature; Freely accessible; Good API access; Useful for machine learning model training

**Limitations:** Smaller than ChEMBL or PubChem; Coverage biased toward well-studied drug targets; Assay conditions vary across entries; Less comprehensive than ChEMBL for overall bioactivity data

**Common beginner mistakes:**

- Not distinguishing between K<sub>i</sub>, K<sub>d</sub>, IC<sub>50</sub>, and EC<sub>50</sub> values
- Not considering assay conditions when comparing values
- Using BindingDB as the sole source for bioactivity data

**When to Use It:** Use BindingDB when you specifically need quantitative binding affinity data for drug-target interactions, particularly for machine learning model training or QSAR studies.

**When NOT to Use It:** Do not use BindingDB as the primary source for comprehensive bioactivity data; use ChEMBL instead. BindingDB is best for specific binding affinity queries.

**Related databases / alternatives:** ChEMBL: Comprehensive bioactivity data; PubChem: Comprehensive chemical information; DrugBank: Drug pharmacology

**How It Connects to Other Resources:** BindingDB cross-references PubChem, ChEMBL, and UniProt. Data is also available through ChEMBL.

**API / FTP / programmatic access:** REST API at <https://www.bindingdb.org/axis2/services/BDBService>; returns XML. Bulk downloads available. Python package bindingdb available.

**Evidence/curation level:** Manually curated from primary literature; moderate quality

**Data Update Status:** Regular updates; actively maintained as of 2024

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Gilson MK et al. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Research, 44(D1):D1045–D1053. <https://doi.org/10.1093/nar/gkv1072>

**Beginner-Friendly Explanation:** BindingDB is a database that focuses specifically on how tightly small molecules (like drugs) bind to proteins. For each pair of a molecule and a protein, it records the measured binding strength (expressed as  $K_i$ ,  $K_d$ ,  $IC_{50}$ , or  $EC_{50}$  values) from laboratory experiments. This makes it particularly useful for researchers who want to understand how strongly a drug candidate binds to its target, or for building computer models that predict binding strength. BindingDB is freely available and is widely used in computational drug discovery.

**Advanced Technical Explanation:** BindingDB implements a data model focused on quantitative binding measurements, capturing the assay type (equilibrium binding, enzyme inhibition, cell-based), measurement type ( $K_i$ ,  $K_d$ ,  $IC_{50}$ ,  $EC_{50}$ ,  $K_{cat}/K_i$ ), and assay conditions. The database includes data from both primary literature and patent databases, providing broader coverage than literature-only databases. BindingDB's data has been widely used for training and benchmarking machine learning models for binding affinity prediction, including deep learning models like DeepDTA and AttentionDTA.

#### One practical workflow example:

- Step 1: Navigate to <https://www.bindingdb.org> and search for your target protein.
- Step 2: Filter by measurement type (e.g.,  $K_i$ ) for consistent data.
- Step 3: Download the binding data for your target in CSV format.
- Step 4: Filter for high-quality data (remove entries with  $>$  or  $<$  qualifiers).
- Step 5: Use the data to build a QSAR model or analyze structure-activity relationships.
- Step 6: Cross-reference with ChEMBL for additional bioactivity data.

## P5 –TTD (Therapeutic Target Database)

**Official Website URL:** <https://db.idrblab.net/ttd>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Drugs / Proteins / Diseases

**What It Is Used For:** TTD is used to access information about therapeutic targets and the drugs that act on them, providing a target-centric view of the drug-target landscape. It is particularly useful for target identification, drug repurposing, and understanding the therapeutic relevance of specific proteins. TTD covers approved drugs, clinical trial drugs, and experimental compounds.

**What Data It Contains:** TTD contains information on over 3,600 therapeutic targets (including successful, clinical trial, and research targets) and over 38,000 drugs/compounds, with information on target-drug relationships, disease associations, target sequences, and drug structures. The database provides a classification of targets by their clinical development stage.

**Main question it helps answer:** Is this protein a validated therapeutic target, and what drugs have been developed against it?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- Is PCSK9 a validated therapeutic target, and what drugs target it?
- What targets are associated with type 2 diabetes?
- What is the clinical development stage of drugs targeting this protein?

**Example use cases:**

- Target validation: checking if a protein is a known therapeutic target
- Drug repurposing: finding drugs that target proteins in a disease pathway Understanding the therapeutic landscape for a disease area

**Input Data Accepted:** Target names, drug names, disease names, UniProt accessions

**Output Data Provided:** Target-drug relationships, disease associations, clinical development status

**Strengths:** Target-centric view of the drug-target landscape; Clinical development stage information; Disease association data; Freely accessible; Covers both approved and investigational drugs

**Limitations:** Smaller than DrugBank or ChEMBL; Less comprehensive bioactivity data than ChEMBL; Less clinical detail than DrugBank; Update frequency may lag behind primary literature

**Common beginner mistakes:**

- Using TTD as the sole source for drug information (supplement with DrugBank)
- Not checking the clinical development stage of targets
- Not cross-referencing with ChEMBL for bioactivity data

**When to Use It:** Use TTD when you want a target-centric view of the drug-target landscape, particularly for target validation and understanding the clinical development status of drugs against a target.

**When NOT to Use It:** Do not use TTD as the primary source for comprehensive drug pharmacology or bioactivity data; use DrugBank or ChEMBL instead.

**Related databases / alternatives:** DrugBank: Comprehensive drug pharmacology; ChEMBL: Comprehensive bioactivity data; PubChem: Comprehensive chemical information

**How It Connects to Other Resources:** TTD cross-references UniProt, DrugBank, ChEMBL, and disease databases. Target sequences are linked to UniProt.

**API / FTP / programmatic access:** Download files available from the TTD website. Limited API access.

**Evidence/curation level:** Manually curated; moderate quality

**Data Update Status:** Regular updates; actively maintained as of 2024

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Wang Y et al. (2020). Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Research, 48(D1):D1031–D1041. <https://doi.org/10.1093/nar/gkz981>

**Beginner-Friendly Explanation:** TTD (Therapeutic Target Database) is a database that focuses on the proteins that drugs target in the body. For each protein target, it tells you what drugs have been developed against it, whether those drugs are approved or still in clinical trials, and what diseases the target is associated with. This makes TTD particularly useful for researchers who want to know whether a protein they are studying is already a drug target, or for finding drugs that might be repurposed for a new disease. TTD is freely available and provides a useful complement to databases like DrugBank and ChEMBL.

**Advanced Technical Explanation:** TTD implements a target classification system that categorizes targets as "successful" (with approved drugs), "clinical trial" (with drugs in clinical development), "research" (with experimental compounds), or "abandoned" (targets that have been discontinued). This classification is valuable for target validation and drug repurposing analyses. TTD's disease association data links targets to disease ontology terms, enabling systematic analysis of the therapeutic landscape for specific disease areas.

#### One practical workflow example:

- Step 1: Navigate to <https://db.idrblab.net/ttd> and search for your target protein.
- Step 2: Check the target classification (successful, clinical trial, research, abandoned).
- Step 3: Review the list of drugs targeting this protein and their clinical development status.
- Step 4: Check the disease associations for the target.
- Step 5: Download the target-drug data for downstream analysis.
- Step 6: Cross-reference with DrugBank for detailed pharmacological information on the drugs.

## BEGINNER EXAMPLE (Category P):

---

A pharmacology student wants to understand how imatinib (Gleevec) works. They search DrugBank for "imatinib" and find its mechanism of action (BCR-ABL tyrosine kinase inhibitor), its targets (ABL1, KIT, PDGFRA), and its clinical indications (CML, GIST). They then check ChEMBL for the IC50 values against each target and find that imatinib is most potent against ABL1 (IC50 ~1 nM). Finally, they check PubChem for the chemical structure and properties.

## ADVANCE EXAMPLE (Category P):

---

A computational chemist is performing a drug repurposing analysis for a new kind of target. They download all compounds tested against related kinases from ChEMBL (using the Python client), filter for high-quality IC50 data (`pchembl_value >= 6`), and build a QSAR model using molecular fingerprints. They then screen the DrugBank approved drug set against the model to identify potential repurposing candidates. They validate top candidates using BindingDB data and check TTD for the clinical development status of the target.

## CONFUSION POINTS (Category P):

---

IC50, Ki, Kd, and EC50 are different measurements and cannot be directly compared.

PubChem Substance (SID) and Compound (CID) records are different; CIDs are standardized structures.

DrugBank's free version has access restrictions; some features require a commercial license.

ChEMBL confidence scores refer to target assignment confidence, not data quality.

TTD "research targets" may not be validated therapeutic targets.

## DECISION GUIDE (Category P):

---

Need comprehensive drug pharmacology (mechanism, ADMET, interactions)? → DrugBank

Need large-scale bioactivity data for QSAR or machine learning? → ChEMBL

Need chemical structure and properties for any compound? → PubChem

Need quantitative binding affinity data? → BindingDB

Need target validation and clinical development status? → TTD

Need to check if a protein is a known drug target? → TTD or DrugBank



## Category Q: Ontologies and Controlled Vocabularies

### OVERVIEW

Ontologies and controlled vocabularies are formal systems for organizing biological knowledge using standardized terms and defined relationships. In bioinformatics, ontologies serve as the backbone for data integration, enabling researchers to annotate biological entities (genes, proteins, diseases, anatomical structures) with consistent, machine-readable terms that can be compared across experiments, species, and databases. Without ontologies, the same concept might be described using dozens of different terms across different databases and publications, making systematic analysis impossible.

Gene Ontology (GO) is the most widely used biological ontology, providing a controlled vocabulary for describing gene product functions across three domains: molecular function, biological process, and cellular component. GO annotations are used in virtually every functional genomics analysis, from RNA-seq differential expression studies to proteomics experiments. Other ontologies address different aspects of biology: the Human Phenotype Ontology (HPO) describes clinical phenotypes for disease genetics, the Disease Ontology (DO) provides standardized disease terms, the Sequence Ontology (SO) describes sequence features, and Uberon provides a multi-species anatomy ontology. MeSH (Medical Subject Headings) is a controlled vocabulary used for indexing biomedical literature.

A key concept in working with ontologies is hierarchical structure: terms are organized in a directed acyclic graph (DAG) where more specific terms are children of more general parent terms. This structure enables "ontology propagation" if a gene is annotated with a specific term, it is implicitly also annotated with all parent terms. This is important for enrichment analysis, where the choice of ontology level (specific vs. general terms) significantly affects the results. Researchers should also be aware that ontology annotations vary in evidence quality, from experimentally validated annotations to computationally inferred ones, and filtering by evidence code is important for high-confidence analyses.

## Q1 – Gene Ontology (GO)

**Official Website URL:** <https://geneontology.org>

**Resource Type:** Ontology / Database

**Main Biological Domain:** DNA sequences / RNA/transcriptomics / Proteins

**What It Is Used For:** Gene Ontology is used to annotate gene products with standardized terms describing their molecular functions, biological processes, and cellular components, enabling systematic functional analysis across species and experiments. GO annotations are used in functional enrichment analysis (GO enrichment), gene set analysis, and comparative genomics.

**What Data It Contains:** GO contains over 43,000 terms organized in three ontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The GO Annotation (GOA) database contains over 700 million annotations for proteins from UniProt and other databases, with evidence codes indicating the basis for each annotation (experimental, computational, literature-based, etc.).

**Main question it helps answer:** What biological functions, processes, and cellular locations are associated with my gene or protein of interest?

**Typical user:** Researcher / Bioinformatician / Data analyst / Beginner student

**Example scientific questions:**

- What biological processes are enriched in my differentially expressed gene list?
- What molecular functions are associated with my protein of interest?
- What GO terms are shared between my gene list and a known disease gene set?

**Example use cases:** GO enrichment analysis of differentially expressed genes; Annotating proteins with functional terms for database submission; Comparing functional profiles of gene sets across conditions

**Input Data Accepted:** Gene names, UniProt accessions, Ensembl IDs, GO term IDs.

**Output Data Provided:** GO term annotations, ontology hierarchy, enrichment analysis results.

**Strengths:** Universal standard for gene function annotation; Covers all organisms with sequenced genomes; Multiple evidence types with quality codes; Excellent tools for enrichment analysis; Freely accessible

**Limitations:** Annotations biased toward well-studied genes and organisms; Many annotations are computationally inferred (IEA) and may be inaccurate; GO terms can be very broad or very specific, affecting enrichment results; Ontology structure can be complex for new users; Redundancy between related GO terms can complicate interpretation

**Common beginner mistakes:** Not filtering by evidence code (including IEA annotations may reduce specificity); Not considering the ontology level when interpreting enrichment results; Treating all GO annotations as experimentally validated; Not accounting for multiple testing correction in enrichment analysis; Using GO enrichment without considering gene set size biases.

**When to Use It:** Use GO for functional annotation and enrichment analysis of gene or protein lists. GO is the standard for functional genomics analyses and should be used whenever you need to characterize the biological functions of a gene set.

**When NOT to Use It:** Do not use GO for disease-specific phenotype annotation (use HPO or DO instead). GO is not ideal for clinical or phenotypic descriptions.

**Related databases / alternatives:** HPO: Clinical phenotype ontology; DO: Disease ontology; KEGG: Pathway-based functional annotation; Reactome: Pathway-based functional annotation

**How It Connects to Other Resources:** GO annotations are provided by UniProt, Ensembl, NCBI Gene, and model organism databases. GO enrichment tools include DAVID, g:Profiler, clusterProfiler, and Enrichr.

**API / FTP / programmatic access:** AmiGO API at <https://api.geneontology.org/>; returns JSON. OWL/OBO format downloads at <https://geneontology.org/docs/download-ontology/>. Python package goatools available.

**Evidence/curation level:** Mixed; experimental (EXP, IDA, IPI, IMP, IGI, IEP), computational (IEA, ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD), and literature-based (TAS, NAS, IC, ND)

**Data Update Status:** Regular releases; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Gene Ontology Consortium (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031. <https://doi.org/10.1093/genetics/iyad031>

**Beginner-Friendly Explanation:** Gene Ontology (GO) is a standardized system for describing what genes and proteins do in cells. It organizes biological knowledge into three categories: what the protein does at the molecular level (Molecular Function), what biological process it participates in (Biological Process), and where in the cell it is located (Cellular Component). Each category is organized as a hierarchy from general to specific terms. GO is used in almost every functional genomics analysis—for example, after identifying differentially expressed genes in an RNA-seq experiment, researchers use GO enrichment analysis to find which biological processes are over-represented in their gene list.

**Advanced Technical Explanation:** The Gene Ontology is implemented as a directed acyclic graph (DAG) using the OBO (Open Biomedical Ontologies) format, with terms connected by "is\_a," "part\_of," "regulates," "positively\_regulates," and "negatively\_regulates" relationships. GO annotations use a controlled vocabulary of evidence codes that indicate the basis for each annotation: experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP) indicate direct experimental support, while electronic annotation (IEA) indicates computational inference. GO enrichment analysis typically uses the hypergeometric test or Fisher's exact test with multiple testing correction (Benjamini-Hochberg FDR), and the choice of background gene set significantly affects results.

### One practical workflow example:

- Step 1: Obtain your gene list (e.g., differentially expressed genes from RNA-seq).
- Step 2: Navigate to g:Profiler (<https://biit.cs.ut.ee/gprofiler/>) or use clusterProfiler in R.
- Step 3: Input your gene list and select the appropriate organism.
- Step 4: Run GO enrichment analysis for all three GO domains (MF, BP, CC).
- Step 5: Filter results by adjusted p-value < 0.05 and review enriched terms.
- Step 6: Visualize results using a dot plot or enrichment map to identify key biological themes.

## Q2 – Sequence Ontology (SO)

**Official Website URL:** <http://www.sequenceontology.org>

**Resource Type:** Ontology

**Main Biological Domain:** DNA sequences / RNA/transcriptomics

**What It Is Used For:** The Sequence Ontology is used to annotate sequence features in genome annotations, providing standardized terms for describing genomic elements such as genes, exons, introns, regulatory regions, and variants. SO is used in genome annotation pipelines, variant annotation, and data exchange formats (GFF3, VCF). It is the standard vocabulary for describing sequence features in genome databases.

**What Data It Contains:** SO contains over 2,400 terms describing sequence features, including gene structures (gene, mRNA, exon, intron, CDS, UTR), regulatory elements (promoter, enhancer, TFBS), repeat elements, variants (SNP, insertion, deletion, inversion), and sequence attributes. Terms are organized hierarchically with defined relationships.

**Main question it helps answer:** What is the standardized term for this type of genomic sequence feature?

**Typical user:** Bioinformatician / Data analyst

**Example scientific questions:**

- What is the correct SO term for a missense variant?
- What SO terms describe regulatory sequence features?
- How should I annotate this novel sequence feature in GFF3 format?

**Example use cases:**

- Annotating genome features in GFF3 format
- Standardizing variant annotations in VCF files
- Developing genome annotation pipelines

**Input Data Accepted:** Feature names, SO term IDs.

**Output Data Provided:** Standardized term definitions, hierarchical relationships.

**Strengths:** Standard vocabulary for sequence feature annotation; Used in GFF3 and VCF formats; Covers all types of genomic features; Freely accessible

**Limitations:** Primarily useful for bioinformaticians developing annotation pipelines; Less relevant for end-user biological analysis; Some terms may be ambiguous or overlapping; Less frequently updated than GO

**Common beginner mistakes:**

- Using non-standard feature names in GFF3 files instead of SO terms
- Not checking SO for the correct term before creating custom annotations
- Confusing SO with GO (different domains)

**When to Use It:** Use SO when developing genome annotation pipelines, creating GFF3 files, or standardizing variant annotations. Essential for bioinformaticians working with genome annotation.

**When NOT Use It:** Do not use SO for functional annotation of genes (use GO instead). SO describes sequence features, not biological functions.

**Related databases / alternatives:** GO: Gene function ontology; Ensembl: Uses SO terms in genome annotations; NCBI: Uses SO terms in RefSeq annotations.

**How It Connects to Other Resources:** SO terms are used in GFF3 genome annotation files, VCF variant files, and by Ensembl, NCBI, and other genome databases.

**API / FTP / programmatic access:** OBO format downloads at <http://www.sequenceontology.org/resources/download.html>. Python package bioontologies available.

**Evidence/curation level:** Community-curated; regularly reviewed.

**Data Update Status:** Periodic updates; actively maintained.

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 3.0

**Citation / Recommended Reference:** Eilbeck K et al. (2005). Sequence Ontology: a tool for the unification of genome annotations. Genome Biology, 6(5):R44. <https://doi.org/10.1186/gb-2005-6-5-r44>

**Beginner-Friendly Explanation:** The Sequence Ontology (SO) is a standardized vocabulary for describing the features found in DNA and RNA sequences. When scientists annotate a genome—marking where genes, exons, regulatory regions, and other features are located—they need to use consistent terminology so that different databases and software tools can understand each other. SO provides this standardized vocabulary. For example, instead of one database calling something a "coding sequence" and another calling it a "CDS," SO provides the official term "CDS" (SO:0000316) that everyone agrees on. SO is particularly important for bioinformaticians who work with genome annotation files.

**Advanced Technical Explanation:** The Sequence Ontology is implemented in OBO format and defines terms for sequence features at multiple levels of abstraction, from broad categories (region, sequence\_feature) to specific feature types (five\_prime\_UTR, splice\_donor\_site, etc.). SO is used as the standard vocabulary for the GFF3 (General Feature Format version 3) genome annotation format, where the "type" field must contain a valid SO term. SO also defines variant types used in VCF (Variant Call Format) files and by variant annotation tools like SnpEff and VEP. The ontology uses "is\_a" and "part\_of" relationships to organize terms hierarchically.

#### One practical workflow example:

- Step 1: Navigate to <http://www.sequenceontology.org> and browse or search for the feature type you need.
- Step 2: Find the correct SO term and its accession number (e.g., SO:0000316 for CDS).
- Step 3: Use the SO term in your GFF3 annotation file in the "type" column.
- Step 4: Validate your GFF3 file using a GFF3 validator that checks SO term compliance.
- Step 5: Use the SO hierarchy to understand the relationships between feature types.
- Step 6: Reference SO terms in your methods section when describing genome annotation.

## Q3 – Human Phenotype Ontology (HPO)

**Official Website URL:** <https://hpo.jax.org>

**Resource Type:** Ontology / Database

**Main Biological Domain:** Diseases / Clinical genomics

**What It Is Used For:** The Human Phenotype Ontology is used to annotate clinical phenotypes in human genetic diseases, providing standardized terms for describing symptoms, signs, and clinical findings. HPO is used in clinical genomics for variant prioritization, disease gene discovery, and patient phenotype matching. It is the standard vocabulary for clinical phenotype description in rare disease genetics.

**What Data It Contains:** HPO contains over 18,000 terms describing human clinical phenotypes, organized hierarchically from general (Abnormality of the nervous system) to specific (Focal-onset impaired awareness seizure). The HPO Annotation database contains over 156,000 disease-phenotype associations linking HPO terms to diseases in OMIM, Orphanet, and DECIPHER.

**Main question it helps answer:** What standardized clinical phenotype terms describe the symptoms of this disease, and what genes are associated with this phenotype?

**Typical user:** Clinician / Researcher / Bioinformatician

**Example scientific questions:**

- What HPO terms describe the phenotype of Marfan syndrome?
- What genes are associated with intellectual disability and seizures?
- What diseases share the phenotype of elevated serum creatine kinase?

**Example use cases:** Variant prioritization in clinical exome/genome sequencing; Patient phenotype matching for rare disease diagnosis; Disease gene discovery using phenotype-genotype associations.

**Input Data Accepted:** HPO term IDs, disease names, gene names, clinical phenotype descriptions.

**Output Data Provided:** Phenotype annotations, disease-gene associations, phenotype similarity scores.

**Strengths:** Standard vocabulary for clinical phenotype description; Comprehensive coverage of rare disease phenotypes; Disease-gene association data; Used by major clinical genomics tools; Freely accessible

**Limitations:** Focused on rare diseases; less comprehensive for common diseases; Phenotype annotations may be incomplete for rare diseases; Clinical terminology may be unfamiliar to non-clinicians; Annotations based on published literature may not reflect full phenotypic spectrum

**Common beginner mistakes:** Using HPO for common disease phenotypes (HPO is primarily for rare diseases); Not using HPO terms in clinical genomics tools that require them.

**Confusing HPO with DO (different purposes:** phenotypes vs. diseases; Not considering phenotype frequency when interpreting disease associations

**When to Use It:** Use HPO for clinical phenotype annotation in rare disease genetics, variant prioritization in clinical genomics, and patient phenotype matching. Essential for clinical genomics workflows.

**When NOT to Use It:** Do not use HPO for common disease phenotypes or for disease classification (use DO instead). HPO describes phenotypes, not diseases.



**Related databases / alternatives:** DO: Disease ontology; OMIM: Disease-gene associations; Orphanet: Rare disease database; MeSH: Medical subject headings

**How It Connects to Other Resources:** HPO is linked to OMIM, Orphanet, DECIPHER, and ClinVar. HPO terms are used by clinical genomics tools including Exomiser, PhenIX, and Phenomizer.

**API / FTP / programmatic access:** REST API at <https://hpo.jax.org/api/hpo/>; returns JSON. OBO/OWL downloads at <https://hpo.jax.org/app/data/ontology>. Python package hpo3 available.

**Evidence/curation level:** Manually curated from primary literature and clinical databases; high quality.

**Data Update Status:** Regular releases; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; hpo.jax.org license (free for academic use)

**Citation / Recommended Reference:** Köhler S et al. (2021). The Human Phenotype Ontology in 2021. Nucleic Acids Research, 49(D1):D1207–D1217. <https://doi.org/10.1093/nar/gkaa1043>

**Beginner-Friendly Explanation:** The Human Phenotype Ontology (HPO) is a standardized vocabulary for describing the clinical features (symptoms and signs) of human diseases. When a patient has a genetic disease, doctors observe specific clinical features, for example, tall stature, long fingers, and heart problems in Marfan syndrome. HPO provides standardized terms for all these clinical features, organized from general to specific. This standardization is crucial for clinical genomics: when a patient's DNA is sequenced, computer programs use HPO terms to match the patient's phenotype to known disease-gene associations and prioritize which genetic variants are most likely to be causing the disease.

**Advanced Technical Explanation:** HPO implements a DAG structure with terms organized under the root "Phenotypic abnormality" (HP:0000118), with major branches for abnormalities of different organ systems. The ontology uses "is\_a" relationships for hierarchical organization and "part\_of" relationships for anatomical relationships. HPO annotations use a frequency vocabulary (obligate, very frequent, frequent, occasional, very rare, excluded) to describe how often a phenotype occurs in a disease. HPO is used by phenotype-driven variant prioritization tools like Exomiser, which uses semantic similarity measures (Resnik, Lin, Jiang-Conrath) to compare patient phenotypes to disease phenotype profiles.

#### One practical workflow example:

- Step 1: Navigate to <https://hpo.jax.org> and search for the clinical features of your patient.
- Step 2: Select the most specific HPO terms that describe the patient's phenotype.
- Step 3: Use the HPO term IDs as input to a variant prioritization tool like Exomiser.
- Step 4: Review the disease-gene associations for the top-ranked diseases.
- Step 5: Cross-reference with OMIM and ClinVar for additional disease-gene information.
- Step 6: Use the HPO API to retrieve phenotype-disease associations programmatically.



## Q4 – Disease Ontology (DO)

**Official Website URL:** <https://disease-ontology.org>

**Resource Type:** Ontology / Database

**Main Biological Domain:** Diseases

**What It Is Used For:** Disease Ontology is used to provide standardized disease terms for annotating biomedical data, integrating disease information across databases, and enabling systematic disease-gene and disease-pathway analyses. DO provides a hierarchical classification of human diseases with cross-references to other disease vocabularies (ICD, MeSH, OMIM, SNOMED CT). It is used in bioinformatics for disease enrichment analysis and data integration.

**What Data It Contains:** DO contains over 11,000 disease terms organized hierarchically, with cross-references to ICD-9, ICD-10, MeSH, OMIM, SNOMED CT, and other disease vocabularies. The database provides disease-gene associations and disease-pathway associations.

**Main question it helps answer:** What is the standardized disease term for this condition, and how does it relate to other diseases in the hierarchy?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What is the DO term for type 2 diabetes mellitus?
- What diseases are classified under "cardiovascular disease" in DO?
- What genes are associated with this disease in DO?

**Example use cases:**

- Standardizing disease annotations in bioinformatics databases
- Disease enrichment analysis of gene sets
- Integrating disease information across databases using DO cross-references.

**Input Data Accepted:** Disease names, DO term IDs, ICD codes, OMIM IDs

**Output Data Provided:** Disease terms, hierarchical relationships, cross-references, disease-gene associations.

**Strengths:** Comprehensive disease classification; Cross-references to multiple disease vocabularies; Freely accessible; Used for disease enrichment analysis; Integrates with major bioinformatics tools

**Limitations:** Less detailed clinical information than HPO; Disease-gene associations may be incomplete; Some disease classifications may not reflect current clinical understanding; Less comprehensive for rare diseases than Orphanet

**Common beginner mistakes:**

- Confusing DO with HPO (DO classifies diseases; HPO describes phenotypes)
- Not using DO cross-references to integrate with other disease databases.
- Using DO for clinical phenotype description (use HPO instead)

**When to Use It:** Use DO for standardizing disease annotations, disease enrichment analysis, and integrating disease information across databases. Useful for any analysis that requires consistent disease terminology.

**When NOT Use It:** Do not use DO for clinical phenotype description (use HPO instead). DO classifies diseases, not symptoms.

**Related databases / alternatives:** HPO: Clinical phenotype ontology; MeSH: Medical subject headings; OMIM: Disease-gene associations; Orphanet: Rare disease database

**How It Connects to Other Resources:** DO cross-references ICD, MeSH, OMIM, SNOMED CT, and other disease vocabularies. DO is used by bioinformatics tools for disease enrichment analysis.

**API / FTP / programmatic access:** OBO/OWL downloads at <https://disease-ontology.org/downloads/>. Python package pronto available for OBO parsing.

**Evidence/curation level:** Manually curated; regularly reviewed.

**Data Update Status:** Regular releases; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 3.0

**Citation / Recommended Reference:** Schriml LM et al. (2022). The Human Disease Ontology 2022 update. Nucleic Acids Research, 50(D1):D1255–D1261. <https://doi.org/10.1093/nar/gkab1063>

**Beginner-Friendly Explanation:** The Disease Ontology (DO) is a standardized vocabulary for human diseases, organized in a hierarchy from general categories (like "cardiovascular disease") to specific conditions (like "coronary artery disease"). It provides a common language for describing diseases across different databases and research tools. One of DO's most useful features is its cross-references to other disease classification systems like ICD (used in clinical medicine) and MeSH (used for literature indexing), which allows researchers to connect information from different sources. DO is particularly useful for bioinformaticians who need to annotate data with disease terms or perform disease enrichment analysis.

**Advanced Technical Explanation:** DO implements a DAG structure using OBO format, with terms organized under the root "disease" (DOID:4). The ontology uses "is\_a" relationships for hierarchical classification and "has\_symptom" and other relationships for phenotype associations. DO's cross-reference system maps DO terms to ICD-9, ICD-10, MeSH, OMIM, SNOMED CT, and NCI Thesaurus identifiers, enabling data integration across clinical and research databases. DO is used by bioinformatics tools like DOSE (R package) for disease enrichment analysis, which uses the hypergeometric test with the DO hierarchy for gene set enrichment.

#### One practical workflow example:

- Step 1: Navigate to <https://disease-ontology.org> and search for your disease of interest.
- Step 2: Note the DO term ID (DOID) for use in downstream analyses.
- Step 3: Check the cross-references to find the corresponding ICD, MeSH, and OMIM identifiers.
- Step 4: Use the DO hierarchy to identify related diseases for comparative analysis.
- Step 5: Use the DOSE R package for disease enrichment analysis of your gene list.
- Step 6: Cross-reference with HPO for clinical phenotype information.

## Q5 – Uberon

**Official Website URL:** <https://obophenotype.github.io/uberon>

**Resource Type:** Ontology

**Main Biological Domain:** Omics / Systems biology

**What It Is Used For:** Uberon is used to annotate anatomical structures across multiple species using a unified vocabulary, enabling cross-species comparison of gene expression, phenotype, and developmental data. It is used in developmental biology, comparative genomics, and single-cell atlas projects to standardize tissue and organ annotations. Uberon bridges species-specific anatomy ontologies (FMA for human, MA for mouse, ZFA for zebrafish, etc.).

**What Data It Contains:** Uberon contains over 15,000 terms describing anatomical structures across multiple species, with cross-references to species-specific anatomy ontologies (FMA, MA, ZFA, FBbt, etc.) and developmental stage ontologies. Terms are organized hierarchically with "is\_a," "part\_of," and "develops\_from" relationships.

**Main question it helps answer:** What is the standardized cross-species term for this anatomical structure, and how does it relate to structures in other species?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- What is the Uberon term for the mouse hippocampus?
- What anatomical structures are part of the mammalian kidney?
- How does the zebrafish pronephros correspond to the human kidney?

**Example use cases:**

- Standardizing tissue annotations in single-cell atlas projects
- Cross-species comparison of gene expression data
- Annotating developmental biology data with consistent anatomical terms

**Input Data Accepted:** Anatomical structure names, Uberon term IDs, species-specific anatomy term IDs.

**Output Data Provided:** Standardized anatomical terms, cross-species mappings, hierarchical relationships.

**Strengths:**

- Cross-species anatomical vocabulary
- Bridges species-specific anatomy ontologies
- Used in major single-cell atlas projects.
- Freely accessible
- Comprehensive coverage of metazoan anatomy

**Limitations:**

- Complex ontology structure can be difficult to navigate.
- Coverage of invertebrate anatomy less comprehensive
- Some cross-species mappings may be approximate.

- Less relevant for plant or microbial biology

### Common beginner mistakes:

- Using species-specific anatomy terms instead of Uberon for cross-species analyses
- Not checking cross-references to species-specific ontologies
- Confusing Uberon with species-specific anatomy ontologies
- **When to Use It:** Use Uberon when you need to compare anatomical data across species, when annotating single-cell atlas data, or when integrating gene expression data from multiple organisms.
- **When NOT Use It:** Do not use Uberon for species-specific detailed anatomy; use the appropriate species-specific ontology (FMA for human, MA for mice, etc.) instead.

### Related databases / alternatives:

- **FMA:** Foundational Model of Anatomy (human)
- **MA:** Mouse Anatomy ontology
- **ZFA:** Zebrafish Anatomy ontology
- **BRENDA:** Tissue ontology

**How It Connects to Other Resources:** Uberon is used by the Human Cell Atlas, Allen Brain Atlas, and other single-cell atlas projects. It cross-references FMA, MA, ZFA, and other species-specific anatomy ontologies.

**API / FTP / programmatic access:** OBO/OWL downloads at <https://obophenotype.github.io/uberon/>. Python package pronto available for OBO parsing. OLS (Ontology Lookup Service) API at <https://www.ebi.ac.uk/ols/ontologies/uberon>.

**Evidence/curation level:** Community-curated; regularly reviewed.

**Data Update Status:** Regular releases; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 3.0

**Citation / Recommended Reference:** Mungall CJ et al. (2012). Uberon, an integrative multi-species anatomy ontology. Genome Biology, 13(1):R5. <https://doi.org/10.1186/gb-2012-13-1-r5>

**Beginner-Friendly Explanation:** Uberon is an anatomy ontology that works across multiple species. When scientists study gene expressions in different animals' say, comparing the brain of a mouse to the brain of a zebrafish—they need a common vocabulary to describe anatomical structures. Uberon provides this vocabulary, with terms that apply across species and cross-references to species-specific anatomy terms. For example, Uberon has a term for "hippocampus" that applies to all mammals, with links to the specific mouse anatomy term and the human anatomy term. This makes Uberon particularly valuable for comparative genomics and for projects like the Human Cell Atlas that aim to map gene expression across tissues.

**Advanced Technical Explanation:** Uberon implements a multi-species anatomy ontology using OBO format, with terms organized using "is\_a" (taxonomic classification), "part\_of" (anatomical composition), "develops\_from" (developmental relationships), and "homologous\_to" (cross-species homology) relationships. Uberon uses a "taxon constraints" system to specify which taxa a term applies to, enabling species-specific queries. The ontology bridges species-specific anatomy ontologies through "xref" cross-references and "equivalent\_to" axioms in OWL format. Uberon is used by Cell Ontology (CL) for tissue context annotations in single-cell data.

**One practical workflow example:**

- Step 1: Navigate to <https://www.ebi.ac.uk/ols/ontologies/uberon> and search for your anatomical structure.
- Step 2: Find the Uberon term ID (UBERON:XXXXXXX) for your structure.
- Step 3: Check the cross-references to find the corresponding species-specific terms.
- Step 4: Use the Uberon term to annotate your data consistently across species.
- Step 5: Use the OLS API to retrieve term information programmatically.
- Step 6: For single-cell data, use Uberon terms for tissue annotations in AnnData objects.

## MeSH

**Official Website URL:** <https://www.nlm.nih.gov/mesh>

**Resource Type:** Ontology / Controlled vocabulary

**Main Biological Domain:** Literature / Diseases

**What It Is Used For:** MeSH is used as the controlled vocabulary for indexing biomedical literature in PubMed and other NLM databases, enabling systematic literature searches and text mining. MeSH terms are assigned to PubMed articles by trained indexers, allowing researchers to find all articles on a topic regardless of the specific terminology used by authors. MeSH is also used for disease annotation in bioinformatics databases.

**What Data It Contains:** MeSH contains over 30,000 descriptors organized in a hierarchical tree structure, covering diseases, drugs, anatomy, organisms, biological phenomena, and other biomedical concepts. Each descriptor may have multiple entry terms (synonyms) and is assigned to one or more tree locations.

**Main question it helps answer:** What is the standardized MeSH term for this biomedical concept, and what PubMed articles are indexed with this term?

**Typical user:** Researcher / Clinician / Bioinformatician / Beginner student

**Example scientific questions:**

- What is the MeSH term for type 2 diabetes?
- What PubMed articles are indexed with the MeSH term "Breast Neoplasms"?
- What MeSH terms are used for CRISPR-related articles?

**Example use cases:**

- Systematic literature searches in PubMed using MeSH terms
- Text mining of biomedical literature using MeSH annotations
- Disease annotation in bioinformatics databases

**Input Data Accepted:** Biomedical concept names, MeSH term IDs.

**Output Data Provided:** MeSH terms, hierarchical relationships, PubMed article links.

**Strengths:** Standard vocabulary for biomedical literature indexing; Enables systematic literature searches; Covers all biomedical domains; Freely accessible; Regularly updated

**Limitations:** Indexing lag means very recent articles may not have MeSH terms; MeSH terms may not capture all nuances of a concept; Hierarchical structure can be complex; Not designed for computational analysis (use DO or HPO for that)

**Common beginner mistakes:** Not using MeSH terms in PubMed searches (relying only on free text); Not using the MeSH tree to find related terms; Confusing MeSH with disease ontologies like DO or HPO; Not using the "explode" function to include all narrower MeSH terms in searches.

**When to Use It:** Use MeSH for systematic literature searches in PubMed, for text mining of biomedical literature, and for disease annotation in databases that use MeSH terminology.

**When NOT to Use It:** Do not use MeSH for computational disease analysis; use DO or HPO instead. MeSH is primarily a literature indexing vocabulary.

**Related databases / alternatives:** DO: Disease ontology for computational analysis; HPO: Clinical phenotype ontology; SNOMED CT: Clinical terminology; ICD: International Classification of Diseases

**How It Connects to Other Resources:** MeSH terms are used in PubMed, MEDLINE, and other NLM databases. MeSH cross-references are included in DO and other disease ontologies.

**API / FTP / programmatic access:** E-utilities API at <https://eutils.ncbi.nlm.nih.gov/>; MeSH RDF at <https://id.nlm.nih.gov/mesh/>. Bulk downloads at <https://www.nlm.nih.gov/databases/download/mesh.html>.

**Evidence/curation level:** Manually curated by NLM indexers; high quality.

**Data Update Status:** Annual updates; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; public domain

**Citation / Recommended Reference:** Lipscomb CE (2000). Medical Subject Headings (MeSH). Bulletin of the Medical Library Association, 88(3):265–266. PMID: 10928714

**Beginner-Friendly Explanation:** MeSH (Medical Subject Headings) is the vocabulary used by the US National Library of Medicine to index articles in PubMed. When a new article is added to PubMed, trained indexers assign MeSH terms to describe what the article is about. This means that if you search PubMed using MeSH terms instead of just keywords, you will find all relevant articles even if they use different words to describe the same concept. For example, searching for the MeSH term "Neoplasms" will find articles about cancer, tumors, malignancies, and other related terms. MeSH is organized as a hierarchy, so you can search for a broad term and include all more specific terms automatically.

**Advanced Technical Explanation:** MeSH implements a hierarchical tree structure with 16 main categories (A: Anatomy, B: Organisms, C: Diseases, D: Chemicals and Drugs, etc.), each with multiple levels of specificity. Each descriptor has a unique MeSH ID (e.g., D003924 for Diabetes Mellitus, Type 2), entry terms (synonyms), scope notes, and tree numbers indicating its position in the hierarchy. MeSH also includes Supplementary Concept Records (SCRs) for specific chemicals, drugs, and rare diseases that are not main descriptors. The MeSH RDF (Resource Description Framework) representation enables SPARQL queries for programmatic access.

### One practical workflow example:

- Step 1: Navigate to <https://www.nlm.nih.gov/mesh> and search for your topic.
- Step 2: Find the most specific MeSH term that describes your topic.
- Step 3: Note the MeSH tree number to understand the hierarchical context.
- Step 4: Use the MeSH term in a PubMed search with the [MeSH Terms] tag: "Diabetes Mellitus, Type 2"[MeSH Terms].
- Step 5: Use the "explode" function (default in PubMed) to include all narrower terms.
- Step 6: Combine MeSH terms with Boolean operators for complex searches.



## Beginner EXAMPLE (Category Q):

---

A biology student has a list of 200 differentially expressed genes from an RNA-seq experiment comparing cancer cells to normal cells. They use g:Profiler (<https://biit.cs.ut.ee/gprofiler/>) to perform GO enrichment analysis. They find significant enrichment for "cell cycle" (GO:0007049), "DNA repair" (GO:0006281), and "apoptosis" (GO:0006915) biological processes. They then check HPO to see if any of these processes are associated with cancer phenotypes.

## ADVANCE EXAMPLE (Category Q):

---

A clinical genomicist is analyzing exome sequencing data from a patient with an undiagnosed rare disease. They describe the patient's phenotype using HPO terms (HP:0001250 Seizures, HP:0001263 Global developmental delay, HP:0000252 Microcephaly) and use Exomiser to prioritize variants. Exomiser uses semantic similarity between the patient's HPO terms and disease phenotype profiles to rank candidate genes. They cross-reference top candidates with DO for disease classification and with GO for functional annotation of the candidate genes.

## CONFUSION POINTS (Category Q):

---

GO describes gene functions; HPO describes clinical phenotypes; DO classifies diseases. These are complementary, not interchangeable.

GO IEA (Inferred from Electronic Annotation) annotations are computationally predicted and may be inaccurate.

MeSH is for literature indexing, not for computational disease analysis.

Uberon is for cross-species anatomy; use species-specific ontologies for detailed anatomy.

GO enrichment results depend heavily on the background gene set and the ontology level used.

## DECISION GUIDE (Category Q):

---

Need to annotate gene functions? → GO; Need to annotate genome sequence features? → SO; Need to describe clinical phenotypes for rare disease genetics? → HPO; Need standardized disease terms for bioinformatics? → DO; Need cross-species anatomical terms? → Uberon; Need to search for biomedical literature systematically? → MeSH; Need disease-gene associations for clinical genomics? → HPO + OMIM

## Category R: Epigenomics Databases

### OVERVIEW

Epigenomics databases store and provide access to data describing heritable changes in gene expression that do not involve alterations to the DNA sequence itself. These changes include DNA methylation, histone modifications, chromatin accessibility, and non-coding RNA regulation. Epigenomic data is inherently cell-type and tissue-specific, making it far more complex to catalog than genomic sequence data. The major epigenomics databases have been built around large-scale consortium efforts—ENCODE (Encyclopedia of DNA Elements) and the Roadmap Epigenomics project—that systematically profiled epigenomic marks across hundreds of cell types and tissues.

The ENCODE project, launched in 2003, aimed to identify all functional elements in the human genome, including transcription factor binding sites, histone modification patterns, methylation, and chromatin accessibility. ENCODE has generated over 19,000 experiments across hundreds of cell types, providing an unprecedented view of the regulatory landscape of the human genome. The Roadmap Epigenomics project complemented ENCODE by focusing on primary human tissues and cells, generating reference epigenomes for 111 human tissues and cell types. While Roadmap data collection is complete (no new data is being generated), the existing dataset remains a valuable reference. Cistrome, JASPAR, and ChIP-Atlas provide complementary resources for transcription factor binding data and ChIP-seq data analysis.

A key challenge in epigenomics is the interpretation of regulatory elements. Transcription factor binding sites identified by ChIP-seq represent potential regulatory elements, but their functional significance must be validated experimentally. JASPAR provides a database of transcription factor binding motifs (position weight matrices) that can be used to predict binding sites from sequence alone. ChIP-Atlas aggregates publicly available ChIP-seq data from multiple sources, providing a comprehensive view of transcription factors binding across cell types. Researchers working with epigenomics data must be aware of the cell-type specificity of epigenomic marks and the importance of using appropriate reference datasets for their specific biological context.

## R1 – ENCODE (Encyclopedia of DNA Elements)

**Official Website URL:** <https://www.encodeproject.org>

**Resource Type:** Database / Repository / Dataset Collection

**Main Biological Domain:** Epigenomics / DNA sequences / RNA/transcriptomics

**What Is Used For:** ENCODE is used to access a comprehensive catalog of functional elements in the human (and mouse) genome, including transcription factor binding sites, histone modifications, DNA methylation, chromatin accessibility (ATAC-seq, DNase-seq), and RNA expression data. It is used for regulatory element annotation, transcription factor binding analysis, and understanding the regulatory landscape of the genome. ENCODE data is widely used as a reference for epigenomic analyses.

**What Data It Contains:** ENCODE contains over 19,000 experiments from over 1,300 cell types and tissues, including ChIP-seq (histone modifications and transcription factors), ATAC-seq, DNase-seq, RRBS/WGBS (DNA methylation), RNA-seq, RAMPAGE, and other assays. Data is available for human (GRCh38) and mouse (mm10) genomes, with processed data (peaks, signal tracks) and raw data (FASTQ files).

**Main question it helps answer:** What regulatory elements (transcription factor binding sites, histone modifications, open chromatin regions) are present in my genomic region of interest?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:** What transcription factors bind to the promoter of my gene of interest? | What histone modifications are present at this enhancer in different cell types? | What regions of the genome are accessible (open chromatin) in this cell type?

**Example use cases:** Annotating regulatory elements in a genomic region of interest; Identifying transcription factor binding sites near differentially expressed genes; Comparing chromatin accessibility across cell types.

**Input Data Accepted:** Genomic coordinates, gene names, cell type names, assay types.

**Output Data Provided:** Processed data files (BED, bigWig, BAM), raw FASTQ files, metadata

**Strengths:** Largest collection of functional genomics data; Standardized experimental protocols and data processing; Comprehensive metadata and quality metrics; Freely accessible; Covers hundreds of cell types

**Limitations:** Primarily focused on human and mouse; Data volume can be overwhelming for new users; Cell line data may not reflect primary tissue biology; Some assays have limited cell type coverage; Data processing pipelines may differ from user's preferred methods

**Common beginner mistakes:** Not filtering by cell type when downloading data; Not checking data quality metrics before using data; Downloading raw FASTQ files when processed data is sufficient; Not using the ENCODE data portal's search filters effectively; Confusing different assay types (ChIP-seq vs. ATAC-seq vs. DNase-seq)

**When to Use It:** Use ENCODE when you need reference epigenomic data for human or mouse cell types, when annotating regulatory elements, or when studying transcription factor binding. ENCODE is the primary reference for functional genomics in human cells.

**When NOT to Use It:** Do not use ENCODE for organisms other than human and mouse. For primary tissue data, consider Roadmap Epigenomics. For specific transcription factor motifs, use JASPAR.

**Related databases / alternatives:** Roadmap Epigenomics: Primary tissue epigenomes; Cistrome: ChIP-seq data analysis; JASPAR: Transcription factor binding motifs; ChIP-Atlas: Aggregated ChIP-seq data

**How It Connects to Other Resources:** ENCODE data is cross-referenced to UCSC Genome Browser, Ensembl, and NCBI. ENCODE experiments are linked to GEO and SRA for raw data access.

**API / FTP / programmatic access:** REST API at <https://www.encodeproject.org/search/?format=json>; returns JSON. FTP downloads at <https://www.encodeproject.org/help/download/>.

**Evidence/curation level:** Experimentally generated with standardized protocols; quality-controlled

**Data Update Status:** Regularly updated with new experiments; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; ENCODE data use policy applies

**Citation / Recommended Reference:** ENCODE Project Consortium (2020). Perspectives on ENCODE. Nature, 583(7818):693–698. <https://doi.org/10.1038/s41586-020-2449-8>

**Beginner-Friendly Explanation:** ENCODE (Encyclopedia of DNA Elements) is a large research project that has systematically mapped all the functional elements in the human genome—not just genes, but also the regulatory regions that control when and where genes are turned on or off. ENCODE has generated thousands of experiments measuring things like which proteins bind to DNA (ChIP-seq), which regions of DNA are accessible (ATAC-seq), and how DNA is chemically modified (methylation). All this data is freely available through the ENCODE portal, making it an invaluable resource for researchers who want to understand the regulatory landscape of the human genome.

**Advanced Technical Explanation:** ENCODE implements standardized experimental protocols and bioinformatics pipelines for each assay type, ensuring data comparability across experiments. ChIP-seq data is processed using the ENCODE ChIP-seq pipeline (SPP peak caller for transcription factors, MACS2 for histone modifications), with IDR (Irreproducibility Discovery Rate) analysis for replicate consistency. ATAC-seq data is processed using the ENCODE ATAC-seq pipeline. All experiments are assigned quality metrics (NRF, PBC1, PBC2 for ChIP-seq; TSS enrichment for ATAC-seq) and must pass quality thresholds to be released. The ENCODE portal provides a REST API for programmatic data access and download.

**One practical workflow example:**

Step 1: Navigate to <https://www.encodeproject.org> and use the search to find experiments for your cell type and assay of interest.

Step 2: Filter by "Assay type" (e.g., ChIP-seq), "Target" (e.g., H3K27ac), and "Biosample" (e.g., HeLa-S3).

Step 3: Select experiments with "Released" status and check quality metrics.

Step 4: Download the processed peak files (BED format) for your analysis.

Step 5: Use bedtools to intersect ENCODE peaks with your genomic regions of interest.

Step 6: Visualize the data in UCSC Genome Browser using the ENCODE track hub.

## R2 – Roadmap Epigenomics

**Official Website URL:** <https://www.roadmapepigenomics.org>

**Resource Type:** Database / Dataset Collection

**Main Biological Domain:** Epigenomics

**What It Is Used For:** Roadmap Epigenomics is used to access reference epigenomes for 111 primary human tissues and cell types, providing comprehensive histone modification, DNA methylation, and chromatin accessibility data. NOTE: Data collection for the Roadmap Epigenomics project is complete; no new data is being generated. The existing dataset remains a valuable reference for primary tissue epigenomics. It is used for understanding tissue-specific regulatory elements and for interpreting GWAS variants in regulatory context.

**What Data It Contains:** Roadmap Epigenomics contains reference epigenomes for 111 human tissues and cell types, including ChIP-seq data for five core histone marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3), DNA methylation (WGBS/RRBS), and chromatin accessibility (DNase-seq). Chromatin state annotations (15-state and 18-state models) are available for all 111 reference epigenomes.

**Main question it helps answer:** What is the epigenomic state of this genomic region in primary human tissues?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What is the chromatin state of this GWAS variant in different tissues?
- What tissues show active enhancer marks at this genomic region?
- What is the DNA methylation pattern at this locus in primary tissues?

**Example use cases:** Interpreting GWAS variants in regulatory context; Identifying tissue-specific enhancers; Comparing epigenomic states across primary tissues

**Input Data Accepted:** Genomic coordinates, tissue names

**Output Data Provided:** Chromatin state annotations, histone modification tracks, DNA methylation data

**Strengths:** Comprehensive primary tissue coverage (111 tissues); Standardized reference epigenomes; Chromatin state annotations available; Freely accessible; Valuable for GWAS interpretation

**Limitations:** Data collection complete; no new data being generated; Limited to human tissues; Some tissues have limited replicate data; Older data may not reflect current best practices; Website may have limited maintenance

**Common beginner mistakes:** Not recognizing that data collection is complete (no new data); Not using the chromatin state annotations for regulatory element interpretation; Not cross-referencing with ENCODE for cell line data

**When to Use It:** Use Roadmap Epigenomics for primary tissue epigenomic data, particularly for GWAS variant interpretation and tissue-specific regulatory element analysis. The 111 reference epigenomes are a unique resource for primary tissue biology.

**When NOT to Use It:** Do not expect new data from Roadmap Epigenomics. For cell line data, use ENCODE. For current primary tissue data, check for newer datasets in GEO or ArrayExpress.

**Related databases / alternatives:** ENCODE: Cell line epigenomics (actively updated); GEO: Repository for new epigenomics datasets; Cistrome: ChIP-seq data analysis.

**How It Connects to Other Resources:** Roadmap data is available through the UCSC Genome Browser and Ensembl. Data is also deposited in GEO and SRA.

**API / FTP / programmatic access:** FTP downloads at <https://egg2.wustl.edu/roadmap/data/byFileType/>. Data also accessible through UCSC Genome Browser track hubs.

**Evidence/curation level:** Experimentally generated with standardized protocols; quality-controlled.

**Data Update Status:** Data collection complete; no new data being generated; existing data remains available.

**Licensing / access restrictions:** Freely available; NIH data use policy applies.

**Citation / Recommended Reference:** Roadmap Epigenomics Consortium et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330. <https://doi.org/10.1038/nature14248>

**Beginner-Friendly Explanation:** The Roadmap Epigenomics project created a comprehensive map of epigenomic marks (chemical modifications to DNA and histones that affect gene expression) across 111 different human tissues and cell types. Unlike ENCODE, which focuses on cell lines grown in the laboratory, Roadmap Epigenomics used primary tissues—actual tissue samples from human donors. This makes it particularly valuable for understanding how genes are regulated in real tissues. The project has finished collecting data, so no new experiments are being added, but the existing dataset is still widely used as a reference for understanding tissue-specific gene regulation.

**Advanced Technical Explanation:** Roadmap Epigenomics generated reference epigenomes using standardized protocols for ChIP-seq (five core histone marks: H3K4me3 for active promoters, H3K4me1 for enhancers, H3K36me3 for transcribed regions, H3K27me3 for Polycomb repression, H3K9me3 for constitutive heterochromatin), WGBS/RRBS for DNA methylation, and DNase-seq for chromatin accessibility. Chromatin state annotations were generated using ChromHMM, a hidden Markov model that integrates multiple histone marks to define chromatin states (active TSS, flanking active TSS, transcr. at gene 5' and 3', strong transcription, weak transcription, genic enhancers, enhancers, ZNF genes & repeats, heterochromatin, bivalent/poised TSS, flanking bivalent TSS/Enh, bivalent enhancer, repressed PolyComb, weak repressed PolyComb, quiescent/low).

#### One practical workflow example:

Step 1: Navigate to <https://www.roadmapepigenomics.org> and browse the available reference epigenomes.

Step 2: Download the chromatin state annotation BED files for tissues of interest.

Step 3: Use bedtools to intersect your GWAS variants with chromatin state annotations.

Step 4: Identify variants that fall in active enhancer states (state 6/7) in disease-relevant tissues.

Step 5: Download the H3K27ac ChIP-seq signal tracks for visualization in UCSC Genome Browser.

Step 6: Cross-reference with ENCODE for cell line data to complement the primary tissue data.



## R3 – Cistrome

**Official Website URL:** <https://cistrome.org>

**Resource Type:** Database / Tool

**Main Biological Domain:** Epigenomics

**What It Is Used For:** Cistrome is used to access and analyze publicly available ChIP-seq and ATAC-seq data, providing a curated collection of processed ChIP-seq datasets for transcription factors and histone modifications. It is used for transcription factor binding analysis, regulatory element identification, and understanding the cistrome (the set of binding sites for a transcription factor) of specific transcription factors. Cistrome DB provides a searchable database of processed ChIP-seq data.

**What Data It Contains:** Cistrome DB contains over 45,000 processed ChIP-seq and ATAC-seq datasets from human and mouse, with quality metrics and peak calls. Data is organized by transcription factors, histone mark, cell type, and species. The Cistrome Data Browser provides visualization and download tools.

**Main question it helps answer:** What is the genome-wide binding sites for this transcription factor in this cell type?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- Where does CTCF bind in the genome of HeLa cells?
- What transcription factors bind near my gene of interest?
- What is the quality of this publicly available ChIP-seq dataset?

**Example use cases:**

- Identifying transcription factor binding sites near genes of interest
- Comparing transcription factors binding across cell types
- Quality assessment of publicly available ChIP-seq data

**Input Data Accepted:** Transcription factor names, cell type names, genomic coordinates.

**Output Data Provided:** Peak files, signal tracks, quality metrics

**Strengths:** Large collection of processed ChIP-seq data; Quality metrics for each dataset; User-friendly data browser; Freely accessible; Covers human and mouse

**Limitations:** Data quality varies across deposited datasets; Coverage of transcription factors and cell types is uneven; Processing pipelines may differ from user's preferred methods; Less comprehensive than ENCODE for standardized data

**Common beginner mistakes:** Not checking quality metrics before using data; Not filtering by cell type for specific analyses; Confusing Cistrome with ENCODE (different scope and curation level)

**When to Use It:** Use Cistrome when you need processed ChIP-seq data for a specific transcription factor or cell type, particularly when ENCODE does not have the data you need.

**When NOT to Use It:** Do not use Cistrome as a substitute for ENCODE's standardized data. For the highest quality data, use ENCODE.



**Related databases / alternatives:** ENCODE: Standardized functional genomics data; ChIP-Atlas: Alternative aggregated ChIP-seq database; JASPAR: Transcription factor binding motifs

**How It Connects to Other Resources:** Cistrome data is linked to GEO and SRA for raw data access. Transcription factors are cross-referenced to UniProt and NCBI Gene.

**API / FTP / programmatic access:** Data downloads available from <https://cistrome.org/db/#/>. Limited API access.

**Evidence/curation level:** Experimentally generated; quality-controlled with automated metrics

**Data Update Status:** Regularly updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Zheng R et al. (2019). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Research*, 47(D1):D729–D735. <https://doi.org/10.1093/nar/gky1094>

**Beginner-Friendly Explanation:** Cistrome is a database and analysis platform for ChIP-seq data—a type of experiment that identifies where specific proteins (like transcription factors) bind to DNA in the genome. Cistrome collects publicly available ChIP-seq datasets from many different laboratories, processes them using standardized methods, and makes them available through a user-friendly web interface. This means that instead of having to download and process raw data yourself, you can directly access processed results showing where a transcription factor of interest binds in different cell types. Cistrome also provides quality metrics to help you assess the reliability of each dataset.

**Advanced Technical Explanation:** Cistrome processes ChIP-seq data using a standardized pipeline that includes read alignment (BWA), peak calling (MACS2), and quality assessment (FastQC, ENCODE quality metrics). Quality metrics include NRF (Non-Redundant Fraction), PBC1 and PBC2 (PCR Bottleneck Coefficients), FRiP (Fraction of Reads in Peaks), and peak number. Cistrome DB provides a RESTful API for programmatic data access and integrates with the Cistrome-GO tool for functional enrichment analysis of ChIP-seq peaks. The Cistrome toolkit includes tools for motif analysis, peak annotation, and differential binding analysis.

**One practical workflow example:**

- Step 1: Navigate to <https://cistrome.org/db> and search for your transcription factor of interest.
- Step 2: Filter by cell type and check quality metrics (FRiP > 0.01, peak number > 500).
- Step 3: Download the peak files (BED format) for high-quality datasets.
- Step 4: Use bedtools to intersect peaks with your genomic regions of interest.
- Step 5: Use MEME-ChIP or HOMER for motif analysis of the peak sequences.
- Step 6: Cross-reference with ENCODE for additional datasets and standardized data.

## R4 – JASPAR (Transcription Factor Binding Profiles Database)

---

**Official Website URL:** <https://jaspar.elixir.no>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Epigenomics / DNA sequences

**What It Is Used For:** JASPAR is used to access a curated database of transcription factor binding profiles (position weight matrices, PWMs) for eukaryotic transcription factors, enabling prediction of transcription factor binding sites from DNA sequence. It is used for regulatory element analysis, promoter analysis, and understanding transcription factor binding specificity. JASPAR is the primary open-access resource for transcription factor binding motifs.

**What Data It Contains:** JASPAR contains over 1,800 curated transcription factor binding profiles (PWMs) for over 1,000 transcription factors from multiple eukaryotic species, including human, mouse, Drosophila, yeast, and plants. Each profile is derived from experimental data (SELEX, ChIP-seq, HT-SELEX) and includes quality metrics and taxonomic information.

**Main question it helps answer:** What is the DNA binding motif for this transcription factor, and where does it bind in my sequence of interest?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:** What is the binding motif for CTCF?; What transcription factors are predicted to bind to this promoter sequence?; What is the similarity between the binding motifs of these two transcription factors?

**Example use cases:** Predicting transcription factor binding sites in promoter sequences; Motif enrichment analysis in ChIP-seq peaks; Comparing binding motifs across transcription factor families

**Input Data Accepted:** Transcription factor names, DNA sequences, JASPAR matrix IDs

**Output Data Provided:** Position weight matrices, binding site predictions, motif similarity scores

**Strengths:** Curated, high-quality transcription factor binding profiles; Open access (unlike TRANSFAC which is commercial); Covers multiple eukaryotic species; Excellent API and programmatic access; Widely used in bioinformatics tools

**Limitations:** Binding site predictions have high false positive rates; PWMs do not capture all aspects of binding specificity (e.g., DNA shape); Coverage of transcription factors is incomplete; Binding motifs may not reflect in vivo binding in all cell types; Some profiles may be derived from limited experimental data

**Common beginner mistakes:** Using JASPAR predictions without experimental validation; Not setting appropriate p-value thresholds for binding site predictions; Confusing JASPAR with TRANSFAC (JASPAR is open access; TRANSFAC is commercial); Not considering cell-type context when interpreting binding site predictions

**When to Use It:** Use JASPAR when you need transcription factor binding motifs for sequence analysis, motif enrichment in ChIP-seq peaks, or promoter analysis.

**When NOT to Use It:** Do not use JASPAR predictions as definitive evidence of transcription factor binding; validate with ChIP-seq data. For commercial applications, TRANSFAC may provide additional motifs.

**Related databases / alternatives:** TRANSFAC: Commercial TF binding database; ENCODE: Experimental TF binding data; Cistrome: Processed ChIP-seq data; HOCOMOCO: Alternative TF binding motif database

**How It Connects to Other Resources:** JASPAR motifs are used by FIMO, MEME-ChIP, HOMER, and other motif analysis tools. Transcription factors are cross-referenced to UniProt and Ensembl.

**API / FTP / programmatic access:** REST API at <https://jaspar.elixir.no/api/v1/>; returns JSON. Python package jaspar available. Bulk downloads at <https://jaspar.elixir.no/downloads/>.

**Evidence/curation level:** Manually curated from experimental data (SELEX, ChIP-seq, HT-SELEX); high quality

**Data Update Status:** Regular releases (JASPAR 2024 current); actively maintained

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Castro-Mondragon JA et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles.

**Beginner-Friendly Explanation:** JASPAR is a database of transcription factor binding motifs—the specific DNA sequences that transcription factors (proteins that control gene expression) prefer to bind to. Each motif is represented as a position weight matrix (PWM), which shows the probability of each DNA base (A, T, G, C) at each position in the binding site. Researchers use JASPAR to predict where a transcription factor might bind in a DNA sequence, or to identify which transcription factors might regulate a gene by analyzing its promoter sequence. JASPAR is freely available, unlike some competing databases that require a commercial license.

**Advanced Technical Explanation:** JASPAR implements position weight matrices (PWMs) derived from experimental binding data, with each matrix representing the log-odds score for each nucleotide at each position relative to a background model. JASPAR 2024 includes profiles derived from SELEX (Systematic Evolution of Ligands by EXponential enrichment), ChIP-seq, and HT-SELEX experiments. The database provides tools for scanning sequences with PWMs (JASPAR TFBS scanner), comparing motifs (JASPAR motif comparison), and clustering motifs by similarity. JASPAR motifs are used by FIMO (Find Individual Motif Occurrences), MEME-ChIP, HOMER, and other widely used motif analysis tools.

#### One practical workflow example:

- Step 1: Navigate to <https://jaspar.elixir.no> and search for your transcription factor of interest.
- Step 2: Download the PWM in JASPAR format for use with motif analysis tools.
- Step 3: Use FIMO (<https://meme-suite.org/meme/tools/fimo>) to scan your sequence for binding sites.
- Step 4: Set a p-value threshold (e.g.,  $p < 1e-4$ ) to filter for high-confidence predictions.
- Step 5: Validate predicted binding sites with ChIP-seq data from ENCODE or Cistrome.
- Step 6: Use the JASPAR API to retrieve motifs programmatically: <https://jaspar.elixir.no/api/v1/matrix/MA0139.1/>.

## R5 – ChIP-Atlas

**Official Website URL:** <https://chip-atlas.org>

**Resource Type:** Database / Tool

**Main Biological Domain:** Epigenomics

**What It Is Used For:** ChIP-Atlas is used to access and analyze a comprehensive collection of publicly available ChIP-seq, ATAC-seq, and DNase-seq data from multiple organisms, providing processed peak data and tools for enrichment analysis. It is used for identifying transcription factor binding sites, analyzing regulatory elements, and performing enrichment analysis of genomic regions. ChIP-Atlas aggregates data from GEO and SRA.

**What Data It Contains:** ChIP-Atlas contains over 100,000 processed ChIP-seq, ATAC-seq, and DNase-seq datasets from human, mouse, rat, fruit fly, nematode, and yeast, with peak calls and signal tracks. The database provides tools for peak enrichment analysis (Enrichment Analysis) and target gene prediction.

**Main question it helps answer:** What transcription factors and histone modifications are enriched at my genomic regions of interest across publicly available datasets?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- What transcription factors are enriched at my ChIP-seq peaks?
- What publicly available ChIP-seq datasets show binding at my genomic region?
- What are the target genes of this transcription factor based on all available ChIP-seq data?

**Example use cases:**

- Enrichment analysis of genomic regions against all available ChIP-seq data
- Identifying transcription factors that co-bind with your factor of interest
- Predicting target genes of a transcription factor

**Input Data Accepted:** Genomic coordinates (BED format), transcription factor names, cell type names

**Output Data Provided:** Enrichment analysis results, peak files, signal tracks

**Strengths:** Largest collection of processed ChIP-seq data; Covers multiple organisms; Enrichment analysis tool for genomic regions; Freely accessible; Regularly updated

**Limitations:** Data quality varies across deposited datasets; Processing pipelines may differ from user's preferred methods; Some organisms have limited coverage; Enrichment analysis may be slow for large datasets

**Common beginner mistakes:**

- Not filtering by data quality before using results
- Not considering cell-type specificity of results
- Using ChIP-Atlas as a substitute for ENCODE's standardized data

**When to Use It:** Use ChIP-Atlas when you need a comprehensive view of transcription factor binding across all available public data, particularly for enrichment analysis of genomic regions.

**When NOT to Use It:** Do not use ChIP-Atlas as a substitute for ENCODE's standardized data. For the highest quality data, use ENCODE or Cistrome.

**Related databases / alternatives:** ENCODE: Standardized functional genomics data; Cistrome: Alternative aggregated ChIP-seq database; JASPAR: Transcription factor binding motifs

**How It Connects to Other Resources:** ChIP-Atlas aggregates data from GEO and SRA. Transcription factors are cross-referenced to UniProt and NCBI Gene.

**API / FTP / programmatic access:** REST API at <https://chip-atlas.org/api>; returns JSON. Bulk downloads available.

**Evidence/curation level:** Experimentally generated; automated quality control

**Data Update Status:** Regularly updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Oki S et al. (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Reports, 19(12):e46255. <https://doi.org/10.15252/embr.201846255>

**Beginner-Friendly Explanation:** ChIP-Atlas is a database that collects and processes publicly available ChIP-seq data—experiments that show where proteins bind to DNA—from thousands of studies deposited in public repositories. Instead of having to download and process raw data from many different sources, ChIP-Atlas provides ready-to-use processed results for over 100,000 datasets across multiple organisms. One of its most useful features is the "Enrichment Analysis" tool, which lets you upload a list of genomic regions and find out which transcription factors and histone modifications are enriched at those regions across all available public data.

**Advanced Technical Explanation:** ChIP-Atlas processes ChIP-seq data from GEO and SRA using a standardized pipeline (Bowtie2 for alignment, MACS2 for peak calling) and provides peak files in BED format and signal tracks in bigWig format. The Enrichment Analysis tool uses a Fisher's exact test to identify transcription factors and histone modifications that are significantly enriched at user-provided genomic regions compared to a background model. ChIP-Atlas also provides a "Target Genes" tool that predicts target genes of a transcription factor based on the proximity of ChIP-seq peaks to gene promoters.

**One practical workflow example:**

Step 1: Navigate to <https://chip-atlas.org> and select "Enrichment Analysis."

Step 2: Upload your BED file of genomic regions of interest (e.g., ATAC-seq peaks).

Step 3: Select the organism and genome assembly.

Step 4: Run the enrichment analysis and wait for results.

Step 5: Review the top enriched transcription factors and histone modifications.

Step 6: Download the results and cross-reference with ENCODE for validation.

## Beginner EXAMPLE (Category R):

---

A graduate student has identified a set of enhancers near differentially expressed genes in their RNA-seq experiment. They want to know which transcription factors might regulate these enhancers. They upload the enhancer coordinates to ChIP-Atlas Enrichment Analysis and find significant enrichment for CTCF, FOXA1, and ESR1 binding. They then check JASPAR for the binding motifs of these transcription factors and use FIMO to scan the enhancer sequences for predicted binding sites.

## ADVANCED EXAMPLE (Category R):

---

A computational biologist is studying the regulatory landscape of a GWAS locus associated with type 2 diabetes. They download Roadmap Epigenomics chromatin state annotations for pancreatic islets and find that the GWAS variants fall in active enhancer states. They then use ENCODE ChIP-seq data for relevant transcription factors (PDX1, FOXA2, NKX6-1) to identify which variants overlap transcription factor binding sites. They use JASPAR to check if the variants affect predicted binding motifs and prioritize variants that disrupt high-confidence binding sites.

## CONFUSION POINTS (Category R):

---

ENCODE and Roadmap Epigenomics are complementary: ENCODE focuses on cell lines, Roadmap on primary tissues.

Roadmap Epigenomics data collection is complete; no new data is being generated.

JASPAR provides binding motifs (sequence preferences), not actual binding sites; use ChIP-seq data for actual binding sites.

ChIP-seq peaks represent protein binding sites, not necessarily functional regulatory elements.

Histone modification patterns (H3K27ac for active enhancers, H3K4me3 for active promoters) are cell-type specific.

## DECISION GUIDE (Category R):

---

Need reference epigenomic data for human cell lines? → ENCODE

Need primary tissue epigenomic data? → Roadmap Epigenomics

Need processed ChIP-seq data for a specific TF? → Cistrome or ChIP-Atlas

Need transcription factor binding motifs for sequence analysis? → JASPAR

Need enrichment analysis of genomic regions against all public ChIP-seq data? → ChIP-Atlas

Need standardized, quality-controlled data? → ENCODE (highest standard)



## Category S: Single-cell and Spatial Omics Resources

### OVERVIEW

Single-cell omics technologies have transformed biology by enabling the measurement of molecular profiles (gene expression, chromatin accessibility, protein levels) in individual cells rather than bulk populations. This resolution reveals cellular heterogeneity that is invisible in bulk measurements—identifying rare cell types, characterizing cell states along developmental trajectories, and mapping the cellular composition of tissues in health and disease. The rapid growth of single-cell data has driven the development of specialized databases and portals for storing, sharing, and analyzing single-cell datasets. The Human Cell Atlas (HCA) is the most ambitious single-cell project, aiming to create a comprehensive reference map of all human cell types. The HCA Data Portal provides access to single-cell RNA-seq, spatial transcriptomics, and other single-cell datasets from diverse tissues and developmental stages. CellxGene (CELLxGENE), developed by the Chan Zuckerberg Initiative, provides an interactive platform for exploring and downloading single-cell datasets, with a focus on standardized cell type annotations. PanglaoDB provides a curated database of cell type marker genes, while the Single Cell Expression Atlas (SCEA) at EBI provides a standardized collection of single-cell datasets with consistent processing and annotation.

Spatial transcriptomics represents the next frontier in single-cell biology, combining gene expression measurement with spatial information about where cells are located within a tissue. Technologies like 10x Visium, Slide-seq, and MERFISH are generating spatially resolved gene expression data that reveals how cell types are organized in tissues and how they interact with their neighbors. Databases and portals are beginning to incorporate spatial transcriptomics data alongside single-cell RNA-seq data, enabling integrated analyses of cellular identity and spatial organization. Researchers working with single-cell data must be aware of the computational challenges of batch correction, cell type annotation, and trajectory analysis that are specific to this data type.



## S1 – Human Cell Atlas (HCA)

**Official Website URL:** <https://www.humancellatlas.org>

**Resource Type:** Database / Repository / Portal

**Main Biological Domain:** Single-cell / Omics

**What It Is Used For:** The Human Cell Atlas is used to access a comprehensive reference map of human cell types, providing single-cell RNA-seq, spatial transcriptomics, and other single-cell datasets from diverse human tissues and developmental stages. It is used for cell type identification, understanding tissue composition, studying development and disease, and as a reference for cell type annotation. The HCA Data Portal provides access to raw and processed single-cell data.

**What Data It Contains:** The HCA Data Portal contains hundreds of single-cell datasets from diverse human tissues, including brain, heart, lung, kidney, liver, gut, immune system, and reproductive tissues, as well as developmental datasets. Data types include scRNA-seq, snRNA-seq, spatial transcriptomics, scATAC-seq, and CITE-seq. Both raw data (FASTQ) and processed data (count matrices) are available.

**Main question it helps answer:** What cell types are present in this human tissue, and what are their molecular signatures?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What cell types are present in the human lung, and what are their marker genes?
- How does the cellular composition of the kidney change in disease?
- What are the transcriptional signatures of immune cell subtypes in the gut?

**Example use cases:** Using HCA data as a reference for cell type annotation in new datasets; Studying the cellular composition of a tissue of interest; Identifying marker genes for specific cell types

**Input Data Accepted:** Tissue names, cell type names, gene names

**Output Data Provided:** Single-cell count matrices, processed AnnData/Seurat objects, metadata

**Strengths:** Comprehensive human tissue coverage; Standardized data formats and metadata; Freely accessible; Includes spatial transcriptomics data; Active community and ongoing data generation

**Limitations:** Data quality and processing vary across datasets; Cell type annotations may not be consistent across datasets; Large file sizes can be challenging to download; Some tissues have limited coverage; Requires computational expertise to analyze

**Common beginner mistakes:** Not checking data quality metrics before using data; Not considering batch effects when integrating multiple datasets; Using HCA cell type annotations without validation in new datasets; Not using the HCA Data Portal's filtering tools to find relevant datasets

**When to Use It:** Use HCA when you need reference single-cell data for human tissues, when annotating cell types in new datasets, or when studying the cellular composition of human tissues.

**When NOT to Use It:** Do not use HCA for non-human organisms; use species-specific resources. HCA data requires computational expertise to analyze.

**Related databases / alternatives:** CellxGene: Interactive single-cell data exploration; PanglaoDB: Cell type marker genes; SCEA: Standardized single-cell datasets; Allen Brain Atlas: Brain-specific single-cell data

**How It Connects to Other Resources:** HCA data is linked to GEO, ArrayExpress, and EGA for raw data access. Cell type annotations use the Cell Ontology (CL) and Uberon for tissue annotations.

**API / FTP / programmatic access:** HCA Data Portal API at <https://service.azure.data.humancellatlas.org/>; returns JSON. Python package hca-py available. Data also accessible through CellxGene.

**Evidence/curation level:** Experimentally generated; quality-controlled; community-curated annotations

**Data Update Status:** Continuously updated with new datasets; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; HCA data use policy applies; some datasets may have access restrictions

**Citation / Recommended Reference:** Regev A et al. (2017). The Human Cell Atlas. eLife, 6:e27041. <https://doi.org/10.7554/eLife.27041>

**Beginner-Friendly Explanation:** The Human Cell Atlas is an international project that aims to create a comprehensive map of all the different cell types in the human body. Just as the Human Genome Project mapped the DNA sequence of the human genome, the Human Cell Atlas is mapping the gene expression profiles of individual cells from every tissue and organ. The project uses single-cell RNA sequencing to measure which genes are active in each cell, allowing scientists to identify and characterize different cell types. The HCA Data Portal provides free access to all the data generated by this project, making it a valuable resource for researchers studying any human tissue.

**Advanced Technical Explanation:** The HCA Data Portal implements a standardized data model based on the HCA metadata schema, which captures detailed information about donors, specimens, protocols, and data files. Data is stored in cloud-based repositories (AWS, GCP, Azure) and accessed through the Azul search API. Single-cell data is provided in standardized formats (10x Genomics HDF5, Loom, AnnData) with consistent metadata. The HCA uses the Cell Ontology (CL) for cell type annotations and Uberon for tissue annotations, enabling cross-dataset comparisons. The HCA Data Portal integrates with CellxGene for interactive data exploration.

#### One practical workflow example:

Step 1: Navigate to <https://data.humancellatlas.org> and search for datasets from your tissue of interest.

Step 2: Filter by tissue, assay type, and organism to find relevant datasets.

Step 3: Check the dataset metadata and quality metrics.

Step 4: Download the processed count matrix (h5ad format) for your analysis.

Step 5: Load the data in Python using scanpy: `import scanpy as sc; adata = sc.read_h5ad('dataset.h5ad')`.

Step 6: Use the HCA data as a reference for cell type annotation in your own dataset using tools like scANVI or Seurat's label transfer.

## S2 – CellxGene (CELLxGENE)

**Official Website URL:** <https://cellxgene.cziscience.com>

**Resource Type:** Database / Portal / Tool

**Main Biological Domain:** Single-cell / Omics

**What It Is Used For:** CellxGene is used to interactively explore, visualize, and download single-cell RNA-seq datasets, providing a standardized collection of curated single-cell datasets with consistent cell type annotations. It is used for cell type exploration, gene expression analysis, and downloading reference datasets for computational analysis. CellxGene is developed by the Chan Zuckerberg Initiative and provides both a web interface and a programmatic API.

**What Data It Contains:** CellxGene contains over 1,000 curated single-cell datasets with over 50 million cells from human and mouse, with standardized cell type annotations using the Cell Ontology (CL), tissue annotations using Uberon, and disease annotations using the Mondo Disease Ontology. Data is available in AnnData (h5ad) format.

**Main question it helps answer:** What does the gene expression profile look like for specific cell types across different tissues and conditions?

**Typical user:** Researcher / Bioinformatician / Data analyst / Beginner student

**Example scientific questions:**

- What is the expression of my gene of interest across different cell types?
- What single-cell datasets are available for my tissue of interest?
- How does gene expression change between healthy and diseased cells?

**Example use cases:** Exploring gene expression across cell types in a tissue of interest; Downloading reference datasets for cell type annotation; Comparing gene expression between conditions in published datasets

**Input Data Accepted:** Gene names, cell type names, tissue names, disease names

**Output Data Provided:** Interactive visualizations, downloadable AnnData files, gene expression data

**Strengths:** User-friendly interactive interface; Standardized cell type and tissue annotations; Large collection of curated datasets; Freely accessible; Excellent programmatic API

**Limitations:** Focused on human and mouse; Cell type annotations may not capture all subtypes; Some datasets may have limited metadata; Interactive exploration requires internet connection; Large datasets may be slow to load

**Common beginner mistakes:** Not using the standardized cell type annotations for cross-dataset comparisons; Not downloading the full AnnData file for computational analysis; Not checking the dataset metadata for experimental details; Using CellxGene for analysis instead of downloading data for local analysis

**When to Use It:** Use CellxGene for interactive exploration of single-cell data, for finding and downloading reference datasets, and for quick gene expression queries across cell types.

**When NOT to Use It:** Do not use CellxGene for complex computational analyses; download the data and use scanpy or Seurat locally. CellxGene is best for exploration and data access.

**Related databases / alternatives:** HCA Data Portal: Comprehensive human cell atlas data; PanglaoDB: Cell type marker genes; SCEA: Standardized single-cell datasets; Single Cell Portal: Alternative single-cell portal

**How It Connects to Other Resources:** CellxGene uses Cell Ontology (CL) for cell type annotations, Uberon for tissue annotations, and Mondo Disease Ontology for disease annotations. Data is linked to GEO and other repositories.

**API / FTP / programmatic access:** Census API at <https://chanzuckerberg.github.io/cellxgene-census/>; Python package cellxgene-census available. REST API for dataset metadata.

**Evidence/curation level:** Curated from published datasets; standardized annotations; quality-controlled

**Data Update Status:** Regularly updated with new datasets; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0 for most datasets

**Citation / Recommended Reference:** Megill C et al. (2021). cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. bioRxiv. <https://doi.org/10.1101/2021.04.05.438318>

**Beginner-Friendly Explanation:** CellxGene is an interactive platform for exploring single-cell RNA-seq data. It provides a large collection of published single-cell datasets that have been carefully curated and annotated with standardized cell type labels. Using the web interface, you can visualize gene expression across different cell types, compare expression between conditions, and explore the cellular composition of different tissues without needing to write any code. For researchers who want to do more detailed analyses, CellxGene also allows you to download the data in a standard format for use with computational tools like scanpy or Seurat.

**Advanced Technical Explanation:** CellxGene implements a standardized data schema (the CELLxGENE schema) that requires consistent metadata fields including cell type (Cell Ontology), tissue (Uberon), disease (Mondo Disease Ontology), assay type (EFO), organism (NCBI Taxonomy), and sex. This standardization enables cross-dataset queries and comparisons. The CellxGene Census provides programmatic access to the entire CellxGENE collection through a TileDB-backed data store, enabling efficient queries across millions of cells without downloading entire datasets. The Census API supports both Python (tiledb-soma) and R (tiledb-soma) interfaces.

### One practical workflow example:

Step 1: Navigate to <https://cellxgene.cziscience.com> and search for datasets from your tissue of interest.

Step 2: Click on a dataset to open the interactive explorer.

Step 3: Search for your gene of interest and visualize its expression across cell types.

Step 4: Use the "Gene Expression" tab to compare expression across cell types.

Step 5: Download the dataset in h5ad format for local analysis.

Step 6: Use the Census API for programmatic access: 

```
import cellxgene_census census = cellxgene_census.open_soma() adata = cellxgene_census.get_anndata(census, organism="Homo sapiens", obs_value_filter="tissue_general == 'lung'")
```

## S3 – PanglaoDB

**Official Website URL:** <https://panglaodb.se>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Single-cell / Omics

**What It Is Used For:** PanglaoDB is used to access a curated database of cell type marker genes for use in single-cell RNA-seq cell type annotation. It provides marker genes for hundreds of cell types across multiple tissues and species, enabling automated and manual cell type annotation of single-cell datasets. PanglaoDB is widely used as a reference for cell type marker genes in single-cell analysis workflows.

**What Data It Contains:** PanglaoDB contains over 8,000 marker genes for over 180 cell types across human and mouse tissues, curated from primary literature and single-cell datasets. Each marker gene entry includes the cell type, tissue, species, and evidence source. The database also provides a collection of processed single-cell datasets.

**Main question it helps answer:** What are the marker genes for this cell type, and which cell types express my gene of interest?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What are the marker genes for pancreatic beta cells?
- Which cell types express CD3E?
- What cell types are present in the human liver based on marker gene expression?

**Example use cases:** Cell type annotation of single-cell RNA-seq clusters; Identifying marker genes for cell types of interest; Validating cell type assignments in single-cell analyses

**Input Data Accepted:** Cell type names, gene names, tissue names

**Output Data Provided:** Marker gene lists, cell type annotations, processed single-cell datasets

**Strengths:** Curated cell type marker genes; Covers many cell types and tissues; Freely accessible; Useful for automated cell type annotation; Provides processed single-cell datasets

**Limitations:** Marker gene lists may not be comprehensive for all cell types; Some marker genes may not be specific to a single cell type; Coverage of rare cell types may be limited; Database may not be updated as frequently as newer resources; Cell type definitions may not match all classification systems

**Common beginner mistakes:** Using a single marker gene for cell type annotation (use multiple markers); Not validating PanglaoDB annotations with additional evidence; Not considering tissue context when using marker genes; Assuming all PanglaoDB marker genes are equally specific

**When to Use It:** Use PanglaoDB when you need a reference list of cell type marker genes for single-cell RNA-seq annotation. Useful as a starting point for cell type annotation.

**When NOT to Use It:** Do not use PanglaoDB as the sole source for cell type annotation; validate with additional evidence. For comprehensive cell type references, use HCA or CellxGene.

**Related databases / alternatives:** CellxGene: Comprehensive single-cell datasets with annotations; HCA: Human Cell Atlas reference data; CellMarker: Alternative cell type marker database; SCEA: Standardized single-cell datasets

**How It Connects to Other Resources:** PanglaoDB marker genes are cross-referenced to Ensembl and NCBI Gene. The database is used by automated cell type annotation tools.

**API / FTP / programmatic access:** Data downloads available from <https://panglaodb.se/markers.html>. Python package panglaodb available.

**Evidence/curation level:** Manually curated from primary literature and single-cell datasets; moderate quality

**Data Update Status:** Periodic updates; maintained as of 2024

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Franzén O et al. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database, 2019:baz046. <https://doi.org/10.1093/database/baz046>

**Beginner-Friendly Explanation:** PanglaoDB is a database of cell type marker genes—the genes that are specifically expressed in particular cell types and can be used to identify those cells in single-cell RNA-seq data. When you perform single-cell RNA-seq, you get clusters of cells that you need to identify. PanglaoDB helps by providing lists of genes that are known to be expressed in specific cell types (like T cells, B cells, neurons, etc.). By checking which marker genes are expressed in each cluster, you can assign cell type identities to your clusters. PanglaoDB covers hundreds of cell types from human and mouse tissues.

**Advanced Technical Explanation:** PanglaoDB implements a cell type marker database with entries organized by cell type, tissue, species, and evidence source. Marker genes are scored based on their specificity and sensitivity for cell type identification. The database provides a web interface for browsing marker genes and a downloadable TSV file for programmatic use. PanglaoDB marker genes are used by automated cell type annotation tools like scType and CellAssign, which use the marker gene lists to assign cell type labels to single-cell clusters based on gene expression patterns.

**One practical workflow example:**

- Step 1: Navigate to <https://panglaodb.se> and search for your cell type of interest.
- Step 2: Download the marker gene list for your cell types of interest.
- Step 3: In your single-cell analysis (scanpy or Seurat), calculate the average expression of marker genes in each cluster.
- Step 4: Use a dot plot or violin plot to visualize marker gene expression across clusters.
- Step 5: Assign cell type labels to clusters based on marker gene expression patterns.
- Step 6: Validate assignments using additional evidence from CellxGene or HCA reference datasets.



## S4 – Single Cell Expression Atlas (SCEA)

**Official Website URL:** <https://www.ebi.ac.uk/gxa/sc>

**Resource Type:** Database / Portal

**Main Biological Domain:** Single-cell / RNA/transcriptomics

**What It Is Used For:** The Single Cell Expression Atlas is used to access a standardized collection of single-cell RNA-seq datasets with consistent processing and cell type annotation, providing a reference for gene expression across cell types and tissues. It is used for exploring gene expression in single-cell data, comparing expression across datasets, and downloading standardized single-cell data for computational analysis. SCEA is maintained by the EBI and provides consistent data processing using standardized pipelines.

**What Data It Contains:** SCEA contains hundreds of single-cell RNA-seq datasets from human, mouse, and other organisms, processed using standardized pipelines (Cell Ranger, STARsolo) and annotated with consistent cell type labels. Data is available in standardized formats (SingleCellExperiment, AnnData) with consistent metadata.

**Main question it helps answer:** What is the expression of my gene of interest across cell types in standardized, consistently processed single-cell datasets?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What is the expression of my gene across cell types in the human pancreas?
- What single-cell datasets are available for my organism and tissue of interest?
- How does gene expression compare between healthy and diseased cells in published datasets?

**Example use cases:**

- Exploring gene expression across cell types in standardized datasets
- Downloading consistently processed single-cell data for computational analysis
- Comparing gene expression across multiple datasets

**Input Data Accepted:** Gene names, cell type names, tissue names, organism names

**Output Data Provided:** Gene expression visualizations, downloadable count matrices, metadata

**Strengths:** Standardized data processing and annotation; Consistent metadata across datasets; Covers multiple organisms; Freely accessible; Integration with other EBI resources

**Limitations:** Smaller collection than CellxGene; Processing pipelines may not match user's preferred methods; Some datasets may have limited cell type annotations; Less interactive than CellxGene

**Common beginner mistakes:** Not using SCEA's standardized annotations for cross-dataset comparisons; Not downloading the full count matrix for computational analysis; Not checking the processing pipeline used for each dataset

**When to Use It:** Use SCEA when you need standardized, consistently processed single-cell data, particularly for cross-dataset comparisons or when working with non-human organisms.

**When NOT to Use It:** Do not use SCEA as the primary resource for human single-cell data; CellxGene or HCA provide larger collections. SCEA is best for standardized cross-dataset analyses.



**Related databases / alternatives:** CellxGene: Larger collection with interactive interface | HCA: Comprehensive human cell atlas data | PanglaoDB: Cell type marker genes | GEO: Raw single-cell data repository

**How It Connects to Other Resources:** SCEA is linked to ArrayExpress and GEO for raw data access. Cell type annotations use the Cell Ontology (CL) and Uberon for tissue annotations. SCEA integrates with the Expression Atlas for bulk RNA-seq data.

**API / FTP / programmatic access:** REST API at <https://www.ebi.ac.uk/gxa/sc/json/>; returns JSON. FTP downloads at [https://ftp.ebi.ac.uk/pub/databases/microarray/data/atlas/sc\\_experiments/](https://ftp.ebi.ac.uk/pub/databases/microarray/data/atlas/sc_experiments/).

**Evidence/curation level:** Curated from published datasets; standardized processing; quality-controlled

**Data Update Status:** Regularly updated with new datasets; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Papatheodorou I et al. (2020). Expression Atlas update: from tissues to single cells. Nucleic Acids Research, 48(D1):D77–D83. <https://doi.org/10.1093/nar/gkz947>

**Beginner-Friendly Explanation:** The Single Cell Expression Atlas (SCEA) is a database maintained by the European Bioinformatics Institute that provides a collection of single-cell RNA-seq datasets that have been processed and annotated using consistent, standardized methods. This standardization is important because it allows you to compare gene expression data across different datasets without worrying about differences in how the data was processed. SCEA covers multiple organisms (not just human) and provides a user-friendly web interface for exploring gene expression across cell types, as well as downloadable data for computational analysis.

**Advanced Technical Explanation:** SCEA implements a standardized data processing pipeline that includes read alignment (Cell Ranger or STARsolo), quality control (cell filtering based on UMI count, gene count, and mitochondrial fraction), normalization (scraper or scater), and dimensionality reduction (PCA, UMAP). Cell type annotations are performed using a combination of automated methods (SingleR, scmap) and manual curation, with annotations mapped to the Cell Ontology (CL). SCEA uses the SingleCellExperiment (SCE) data format for R and AnnData (h5ad) for Python, with consistent metadata fields following the HCA metadata schema.

#### **One practical workflow example:**

Step 1: Navigate to <https://www.ebi.ac.uk/gxa/sc> and search for datasets from your tissue & organism of interest.

Step 2: Click on a dataset to explore gene expression across cell types.

Step 3: Search for your gene of interest and visualize its expression.

Step 4: Download the processed count matrix in h5ad format.

Step 5: Load the data in Python: `import scanpy as sc; adata = sc.read_h5ad('dataset.h5ad')`.

Step 6: Use the standardized cell type annotations for cross-dataset comparisons.

## Beginner Example (Category S):

---

A biology student has performed single-cell RNA-seq on human lung tissue and has identified 15 clusters. They want to annotate these clusters with cell type labels. They download the PanglaoDB marker gene list and check which marker genes are expressed in each cluster. They find that cluster 3 expresses SFTPC, SFTPB, and ABCA3 (alveolar type 2 cell markers). They then check CellxGene to find published lung datasets and compare their cluster to annotated cell types in the reference data.

## ADVANCED EXAMPLE (Category S):

---

A computational biologist is building a reference atlas for human kidney disease. They download all available kidney single-cell datasets from CellxGene using the Census API, integrate them using scVI (a deep learning-based integration method), and harmonize cell type annotations using the Cell Ontology. They then compare the cellular composition of healthy and diseased kidneys, identifying a disease-specific population of injured proximal tubule cells. They validate their findings using HCA kidney data and deposit their integrated atlas back to CellxGene.

## CONFUSION POINTS (Category S):

---

Single-cell RNA-seq measures gene expression, not protein levels; use CITE-seq for protein-level data.

Cell type annotations in single-cell data are often uncertain and should be validated with multiple marker genes.

Batch effects between datasets can mimic biological differences; always perform batch correction when integrating datasets.

PanglaoDB marker genes are not always specific to a single cell type; use multiple markers for annotation.

Spatial transcriptomics data has lower resolution than single-cell RNA-seq but provides spatial context.

## DECISION GUIDE (Category S):

---

Need comprehensive human cell type reference data? → Human Cell Atlas

Need interactive exploration of single-cell data? → CellxGene

Need cell type marker genes for annotation? → PanglaoDB

Need standardized, consistently processed single-cell data? → SCEA

Need programmatic access to large single-cell collections? → CellxGene Census API

Need spatial transcriptomics data? → HCA Data Portal or specialized spatial databases

## Category T: Microbiome and Metagenomics Databases

### OVERVIEW

Metagenomics and microbiome research have transformed our understanding of microbial communities in diverse environments, from the human gut to ocean sediments. Unlike traditional microbiology, which studies individual organisms in culture, metagenomics analyzes the collective genetic material of all organisms in a sample, enabling the study of unculturable microorganisms that constitute the vast majority of microbial diversity. The field has been driven by advances in high-throughput sequencing and the development of specialized databases and analysis tools for processing and interpreting metagenomic data.

The major microbiome and metagenomics databases serve different but complementary purposes. MGnify (formerly EBI Metagenomics) is the primary repository for metagenomic datasets, providing standardized analysis pipelines and a searchable database of processed results. SILVA and the Ribosomal Database Project (RDP) provide curated databases of ribosomal RNA sequences for taxonomic classification, with SILVA being the more actively maintained resource. The Genome Taxonomy Database (GTDB) provides a phylogenetically consistent taxonomy for bacteria and archaea based on whole-genome data, addressing inconsistencies in traditional taxonomy. MG-RAST is an alternative metagenomics analysis platform that provides automated annotation and analysis of metagenomic datasets.

A key challenge in microbiome research is taxonomic classification—assigning sequences to specific organisms. Different databases use different taxonomic frameworks, and the choice of reference database significantly affects the results of taxonomic classification. SILVA and GTDB use different approaches to bacterial taxonomy, and researchers must be aware of these differences when comparing results across studies. Additionally, the distinction between 16S rRNA amplicon sequencing (which targets a specific marker gene) and shotgun metagenomics (which sequences all DNA in a sample) is important for choosing the appropriate database and analysis approach.

## T1 – MGnify (formerly EBI Metagenomics)

**Official Website URL:** <https://www.ebi.ac.uk/metagenomics>

**Resource Type:** Database / Repository / Tool

**Main Biological Domain:** Microbiome / Omics

**What It Is Used For:** MGnify is used to access a comprehensive repository of metagenomic and metatranscriptomic datasets, with standardized analysis pipelines for taxonomic classification, functional annotation, and diversity analysis. It is used for exploring microbial community composition, comparing microbiomes across environments, and accessing processed metagenomic data. MGnify provides both raw data access and processed analysis results.

**What Data It Contains:** MGnify contains over 700,000 samples from diverse environments (human gut, soil, ocean, etc.), with processed results including taxonomic profiles (16S rRNA and shotgun), functional annotations (GO, KEGG, InterPro), and diversity metrics. Raw data is linked to ENA (European Nucleotide Archive).

**Main question it helps answer:** What is the microbial community composition and functional potential of this environmental sample?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What bacteria are present in this human gut microbiome sample?
- How does the microbial community composition differ between healthy and diseased individuals?
- What metabolic functions are encoded in this soil metagenome?

**Example use cases:** Accessing processed metagenomic data for comparative analyses; Exploring microbial diversity across environments; Downloading taxonomic profiles for meta-analyses

**Input Data Accepted:** Sample IDs, environment types, taxonomic names

**Output Data Provided:** Taxonomic profiles, functional annotations, diversity metrics, downloadable data

**Strengths:** Largest repository of processed metagenomic data; Standardized analysis pipelines; Covers diverse environments; Freely accessible; Integration with ENA for raw data

**Limitations:** Analysis pipelines may not match user's preferred methods; Taxonomic classification depends on reference database version; Some environments have limited coverage; Large datasets can be slow to download; Functional annotations may be incomplete for novel organisms

**Common beginner mistakes:** Not checking the analysis pipeline version used for each dataset; Not considering the reference database used for taxonomic classification; Comparing results from different pipeline versions without normalization; Not downloading raw data for custom analyses

**When to Use It:** Use MGnify when you need processed metagenomic data for comparative analyses, when exploring microbial diversity across environments, or when you need a standardized reference for microbiome studies.

**When NOT to Use It:** Do not use MGnify as a substitute for custom analysis pipelines when specific methods are required. For taxonomic classification, consider using SILVA or GTDB directly.

**Related databases / alternatives:** MG-RAST: Alternative metagenomics analysis platform; SILVA: Ribosomal RNA reference database; GTDB: Genome taxonomy database; NCBI SRA: Raw metagenomic data repository

**How It Connects to Other Resources:** MGnify is linked to ENA for raw data access. Taxonomic classifications use SILVA and NCBI Taxonomy. Functional annotations use GO, KEGG, and InterPro.

**API / FTP / programmatic access:** REST API at <https://www.ebi.ac.uk/metagenomics/api/v1/>; returns JSON. Python package jsonapi-client available. FTP downloads at <https://ftp.ebi.ac.uk/pub/databases/metagenomics/>.

**Evidence/curation level:** Experimentally generated; standardized automated analysis; quality-controlled

**Data Update Status:** Continuously updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; EBI data use policy applies

**Citation / Recommended Reference:** Richardson L et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Research, 51(D1):D753–D759. <https://doi.org/10.1093/nar/gkac1080>

**Beginner-Friendly Explanation:** MGnify is a database and analysis platform for microbiome data, maintained by the European Bioinformatics Institute. It collects metagenomic sequencing data from diverse environments—from the human gut to ocean water to soil—and processes it using standardized methods to identify which microorganisms are present and what metabolic functions they can perform. MGnify is particularly useful because it provides ready-to-use processed results, so you do not need to run the analysis yourself. You can search for samples from specific environments, compare microbial communities across conditions, and download the data for your own analyses.

**Advanced Technical Explanation:** MGnify implements a standardized analysis pipeline (MGnify pipeline v5.0) that includes quality control (Trimmomatic), rRNA identification (Infernal), taxonomic classification (SILVA SSU/LSU for amplicon data; Kraken2/Bracken for shotgun data), and functional annotation (InterProScan for protein families, GO terms, KEGG pathways). The pipeline produces standardized output files including OTU/ASV tables, taxonomic profiles, and functional annotation tables. MGnify uses the SILVA taxonomy for 16S rRNA classification and NCBI Taxonomy for shotgun metagenomics. The MGnify API provides programmatic access to all analysis results.

#### One practical workflow example:

- Step 1: Navigate to <https://www.ebi.ac.uk/metagenomics> and search for studies from your environment of interest.
- Step 2: Filter by biome (e.g., "Human gut") and sequencing type (e.g., "amplicon").
- Step 3: Select a study and browse the sample metadata and analysis results.
- Step 4: Download the taxonomic profiles (TSV format) for comparative analysis.
- Step 5: Use the MGnify API to retrieve data programmatically: `import requests response = requests.get('https://www.ebi.ac.uk/metagenomics/api/v1/studies?biome_name=Human+gut')`
- Step 6: Use R packages (phyloseq, vegan) for diversity analysis of the downloaded data.

## T2 – SILVA

**Official Website URL:** <https://www.arb-silva.de>

**Resource Type:** Database / Reference

**Main Biological Domain:** Microbiology / Taxonomy / Metagenomics

**What It Is Used For:** SILVA is used as a comprehensive reference database of ribosomal RNA (rRNA) sequences for taxonomic classification of bacteria, archaea, and eukaryotes. It is the primary reference database for 16S rRNA amplicon sequencing analysis, providing curated and aligned rRNA sequences with taxonomic annotations. SILVA is used in microbiome research for OTU/ASV classification, phylogenetic analysis, and diversity studies.

**What Data It Contains:** SILVA contains over 9 million ribosomal RNA sequences (SSU: 16S/18S; LSU: 23S/28S) from bacteria, archaea, and eukaryotes, with curated taxonomic annotations and multiple sequence alignments. The database provides both full-length and partial sequences, with quality filtering to remove chimeric and low-quality sequences.

**Main question it helps answer:** What organism does this ribosomal RNA sequence belong to?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What is the taxonomic classification of this 16S rRNA sequence?
- What bacteria are present in this microbiome sample based on 16S amplicon sequencing?
- What is the phylogenetic relationship between these bacterial isolates?

**Example use cases:**

- Taxonomic classification of 16S rRNA amplicon sequencing data
- Building phylogenetic trees from rRNA sequences
- Quality control of rRNA sequences (chimera detection)

**Input Data Accepted:** rRNA sequences (FASTA format), sequence IDs

**Output Data Provided:** Taxonomic classifications, aligned sequences, phylogenetic trees

**Strengths:** Comprehensive rRNA reference database; Curated taxonomic annotations; Multiple sequence alignments available; Widely used standard for microbiome research; Freely accessible

**Limitations:** Taxonomy may not reflect current phylogenomic understanding (use GTDB for genome-based taxonomy); Some taxonomic groups are better represented than others; Large database size requires significant computational resources; Taxonomy updates may lag behind primary literature

**Common beginner mistakes:** Not using the appropriate SILVA release for reproducibility; Not filtering for high-quality sequences before classification; Confusing SILVA taxonomy with GTDB taxonomy; Not using the SILVA-formatted database for specific tools (QIIME2, DADA2)

**When to Use It:** Use SILVA for 16S rRNA amplicon sequencing analysis, phylogenetic analysis of rRNA sequences, and as a reference for microbiome taxonomic classification.

**When NOT to Use It:** Do not use SILVA for whole-genome-based taxonomy; use GTDB instead. SILVA is specifically for rRNA-based classification.



**Related databases / alternatives:** GTDB: Genome-based taxonomy (recommended for whole-genome data); RDP: Alternative rRNA database (limited updates); Greengenes: Alternative 16S database (limited updates); NCBI Taxonomy: General taxonomy reference

**How It Connects to Other Resources:** SILVA sequences are cross-referenced to NCBI GenBank and EMBL. SILVA taxonomy is used by MGnify and other metagenomics platforms.

**API / FTP / programmatic access:** FTP downloads at <https://www.arb-silva.de/download/arb-files/>. SILVA-formatted databases for QIIME2, DADA2, and mothur available. Web-based SINA aligner at <https://www.arb-silva.de/aligner/>.

**Evidence/curation level:** Manually curated and quality-controlled; high quality

**Data Update Status:** Regular releases (SILVA 138.1 current); actively maintained as of 2024

**Licensing / access restrictions:** Open access; freely available at <https://www.arb-silva.de>.

**Citation / Recommended Reference:** Quast C et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596. <https://doi.org/10.1093/nar/gks1219>

**Beginner-Friendly Explanation:** SILVA is a database of ribosomal RNA (rRNA) sequences—a type of RNA that is present in all living organisms and is commonly used to identify and classify microorganisms. The 16S rRNA gene is particularly useful for bacterial identification because it is highly conserved (similar across all bacteria) but also has variable regions that differ between species. SILVA provides a comprehensive, curated collection of rRNA sequences with taxonomic labels, which researchers use as a reference to classify the bacteria in their microbiome samples. When you sequence the 16S rRNA gene from a microbiome sample, you compare your sequences to SILVA to identify which bacteria are present.

**Advanced Technical Explanation:** SILVA implements a comprehensive rRNA sequence database with multiple quality control steps: chimera detection (UCHIME), alignment quality filtering (minimum alignment score), and taxonomic consistency checking. The database provides both the SSU (Small SubUnit: 16S for bacteria/archaea, 18S for eukaryotes) and LSU (Large SubUnit: 23S for bacteria/archaea, 28S for eukaryotes) rRNA databases. SILVA taxonomy is based on a combination of phylogenetic analysis and nomenclatural rules, with regular updates to reflect new taxonomic proposals. SILVA provides formatted databases for use with QIIME2 (silva-138-99-seqs.qza), DADA2 (silva\_nr99\_v138.1\_train\_set.fa.gz), and mothur.

**One practical workflow example:**

- Step 1: Download the SILVA SSU reference database for your analysis tool (e.g., SILVA 138.1 for QIIME2).
- Step 2: Process your 16S amplicon sequencing data to generate ASVs (using DADA2 or denoising).
- Step 3: Classify your ASVs against the SILVA database using the QIIME2 classify-sklearn classifier.
- Step 4: Generate a taxonomy table with SILVA taxonomic assignments.
- Step 5: Use phyloseq or vegan in R for diversity analysis.
- Step 6: Report the SILVA release version used for reproducibility.



## T3 – GTDB (Genome Taxonomy Database)

**Official Website URL:** <https://gtdb.ecogenomic.org>

**Resource Type:** Database / Reference

**Main Biological Domain:** Microbiome / DNA sequences

**What It Is Used For:** GTDB is used to access a phylogenetically consistent taxonomy for bacteria and archaea based on whole-genome data, providing a standardized reference for genome-based taxonomic classification. It is used for taxonomic classification of metagenome-assembled genomes (MAGs), comparative genomics, and understanding the diversity of bacteria and archaea. GTDB addresses inconsistencies in traditional bacterial taxonomy by using a genome-based phylogenetic framework.

**What Data It Contains:** GTDB contains taxonomic classifications for over 400,000 bacterial and archaeal genomes, with a phylogenetically consistent taxonomy based on 120 conserved marker genes. The database provides genome trees, taxonomic assignments, and metadata for all classified genomes.

**Main question it helps answer:** What is the phylogenetically consistent taxonomic classification of this bacterial or archaeal genome?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What is the GTDB taxonomic classification of this metagenome-assembled genome?
- How does the GTDB taxonomy differ from NCBI taxonomy for this organism?
- What is the phylogenetic diversity of bacteria in this environment?

**Example use cases:**

- Taxonomic classification of metagenome-assembled genomes (MAGs)
- Comparative genomics with phylogenetically consistent taxonomy
- Understanding the diversity of uncultured bacteria

**Input Data Accepted:** Genome sequences (FASTA), genome IDs

**Output Data Provided:** Taxonomic classifications, genome trees, metadata

**Strengths:** Phylogenetically consistent taxonomy; Based on whole-genome data (more accurate than 16S-based); Covers uncultured organisms (MAGs); Freely accessible; Regularly updated

**Limitations:** Taxonomy may differ significantly from traditional NCBI taxonomy; Requires whole-genome data (not suitable for 16S amplicon data); Some taxonomic names may not be formally published; Computational resources required for GTDB-Tk classification

**Common beginner mistakes:** Confusing GTDB taxonomy with NCBI taxonomy (they can differ significantly); Using GTDB for 16S amplicon data (use SILVA instead); Not using GTDB-Tk for genome classification; Not reporting which GTDB release was used

**When to Use It:** Use GTDB for taxonomic classification of bacterial and archaeal genomes, particularly metagenome-assembled genomes. GTDB provides the most phylogenetically consistent taxonomy for whole-genome data.

**When NOT to Use It:** Do not use GTDB for 16S rRNA amplicon data; use SILVA instead. GTDB requires whole-genome sequences.

**Related databases / alternatives:** SILVA: rRNA-based taxonomy (for amplicon data); NCBI Taxonomy: Traditional taxonomy reference; RDP: Alternative rRNA database

**How It Connects to Other Resources:** GTDB genomes are cross-referenced to NCBI GenBank. GTDB-Tk uses HMMER and FastTree for genome classification.

**API / FTP / programmatic access:** FTP downloads at <https://data.gtdb.ecogenomic.org/>. GTDB-Tk tool for genome classification: pip install gtdbtk. API at <https://gtdb.ecogenomic.org/api>.

**Evidence/curation level:** Computationally derived from whole-genome data; regularly reviewed

**Data Update Status:** Regular releases (GTDB R220 current); actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Parks DH et al. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794. <https://doi.org/10.1093/nar/gkab776>

**Beginner-Friendly Explanation:** GTDB (Genome Taxonomy Database) is a database that provides a new, more consistent way of classifying bacteria and archaea based on their whole genome sequences. Traditional bacterial taxonomy has many inconsistencies—organisms that look similar may be distantly related, and the same name may be used for different organisms in different databases. GTDB addresses this by using a large set of conserved genes from whole genomes to build a phylogenetic tree and assign consistent taxonomic names. This is particularly important for metagenomics, where researchers often discover new organisms that have never been cultured and need to be classified based on their genome sequences alone.

**Advanced Technical Explanation:** GTDB implements a genome-based taxonomy using 120 conserved bacterial marker genes (bac120) and 53 archaeal marker genes (ar53) for phylogenetic inference. The taxonomy is normalized to ensure consistent rank assignments across the tree, using relative evolutionary divergence (RED) values to define taxonomic ranks. GTDB-Tk (the classification toolkit) uses HMMER for marker gene identification, FastTree or IQ-TREE for phylogenetic placement, and pplacer for query genome placement into the reference tree. GTDB taxonomy often differs significantly from NCBI taxonomy, particularly for recently reclassified groups like Firmicutes (now Bacillota) and Proteobacteria (now Pseudomonadota).

**One practical workflow example:**

- Step 1: Install GTDB-Tk: conda install -c bioconda gtdbtk.
- Step 2: Download the GTDB reference data: download-db.sh.
- Step 3: Run GTDB-Tk on your genome bins: gtdbtk classify\_wf --genome\_dir bins/ --out\_dir gtdbtk\_output/.
- Step 4: Review the taxonomic classifications in the output TSV file.
- Step 5: Compare GTDB classifications with NCBI taxonomy for any discrepancies.
- Step 6: Report the GTDB release version used for reproducibility.

## T4 – RDP (Ribosomal Database Project)

**Official Website URL:** <https://rdp.cme.msu.edu>

**Resource Type:** Database / Tool

**Main Biological Domain:** Microbiome / DNA sequences

**What It Is Used For:** RDP is used to access a curated database of ribosomal RNA sequences and tools for taxonomic classification of bacterial and archaeal 16S rRNA sequences. NOTE: RDP has had limited updates in recent years and SILVA is generally recommended as the primary reference for 16S rRNA classification. RDP's historical data and tools (RDP Classifier) remain useful for certain analyses.

**What Data It Contains:** RDP contains over 3.5 million 16S rRNA sequences from bacteria and archaea, with curated taxonomic annotations. The RDP Classifier is a widely used tool for rapid taxonomic classification of 16S rRNA sequences.

**Main question it helps answer:** What is the taxonomic classification of this 16S rRNA sequence (using the RDP reference database)?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- What is the RDP taxonomic classification of this 16S rRNA sequence?
- How does RDP classification compare to SILVA classification for my sequences?

**Example use cases:**

- Taxonomic classification of 16S rRNA sequences using the RDP Classifier
- Historical benchmarking of classification methods
- Accessing the RDP training set for classifier development

**Input Data Accepted:** 16S rRNA sequences (FASTA format)

**Output Data Provided:** Taxonomic classifications with confidence scores

**Strengths:** RDP Classifier is fast and widely used; Historical reference for 16S classification; Freely accessible

**Limitations:** Limited updates in recent years (legacy resource); SILVA is generally recommended over RDP for current analyses; Smaller database than SILVA; Website availability may be intermittent

**Common beginner mistakes:**

- Using RDP as the primary reference without recognizing its limited update status
- Not cross-referencing with SILVA for current classifications

**When to Use It:** Use RDP primarily for the RDP Classifier tool or for historical benchmarking. For current 16S classification, SILVA is recommended.

**When NOT to Use It:** Do not use RDP as the primary reference for current microbiome analyses; use SILVA instead.

**Related databases / alternatives:**

- SILVA: Current comprehensive rRNA database (recommended)

- GTDB: Genome-based taxonomy
- Greengenes: Alternative 16S database (also limited updates)

**How It Connects to Other Resources:** RDP sequences are cross-referenced to NCBI GenBank. RDP Classifier is used by QIIME and other microbiome analysis tools.

**API / FTP / programmatic access:** RDP Classifier available at <https://rdp.cme.msu.edu/classifier/>. FTP downloads available.

**Evidence/curation level:** Manually curated; limited recent updates

**Data Update Status:** Limited updates in recent years; legacy resource

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Cole JR et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research, 42(D1):D633–D642. <https://doi.org/10.1093/nar/gkt1244>

**Beginner-Friendly Explanation:** The Ribosomal Database Project (RDP) is one of the original databases for ribosomal RNA sequences, providing a reference collection of 16S rRNA sequences for bacterial identification. RDP also developed the RDP Classifier, a widely used tool for rapidly assigning taxonomic labels to 16S rRNA sequences. However, RDP has not been actively updated in recent years, so for current microbiome research, SILVA is generally the better choice. The RDP Classifier tool itself remains useful and is still widely used, even when the underlying reference database is from SILVA.

**Advanced Technical Explanation:** The RDP Classifier uses a naive Bayes classifier trained on the RDP reference database to assign taxonomic labels to 16S rRNA sequences with bootstrap confidence scores. The classifier is fast (can classify thousands of sequences per second) and provides confidence scores at each taxonomic level, allowing users to filter classifications by confidence threshold. The RDP training set is also used to train classifiers in QIIME2 (using the sklearn classifier). RDP's taxonomy follows the Bergey's Manual of Systematic Bacteriology nomenclature.

**One practical workflow example:**

- Step 1: Download the RDP Classifier from <https://rdp.cme.msu.edu/classifier/>.
- Step 2: Run the classifier on your 16S sequences: `java -jar classifier.jar classify -o output.txt sequences.fasta`.
- Step 3: Filter classifications by confidence score (e.g., > 0.8 at genus level).
- Step 4: Compare RDP classifications with SILVA classifications for consistency.
- Step 5: For current analyses, use SILVA as the primary reference.
- Step 6: Report the RDP release version used for reproducibility.

## T5 – MG-RAST (Metagenomics Rapid Annotation using Subsystem Technology)

**Official Website URL:** <https://www.mg-rast.org>

**Resource Type:** Database / Tool / Repository

**Main Biological Domain:** Microbiome / Omics

**What It Is Used For:** MG-RAST is used to submit, analyze, and access metagenomic datasets, providing automated annotation and analysis of metagenomic sequences. It is used for functional annotation of metagenomes, taxonomic classification, and comparative metagenomics. MG-RAST provides a web-based platform for metagenomics analysis without requiring local computational resources.

**What Data It Contains:** MG-RAST contains over 400,000 metagenomic datasets from diverse environments, with automated functional annotations (SEED subsystems, COG, KEGG) and taxonomic classifications. Both raw data and processed results are available.

**Main question it helps answer:** What is the functional and taxonomic composition of this metagenomic sample?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What metabolic functions are present in this soil metagenome?
- How does the functional composition of this microbiome compare to others?
- What organisms are present in this environmental sample?

**Example use cases:**

- Automated annotation of metagenomic datasets
- Comparative metagenomics across environments
- Accessing publicly available metagenomic data

**Input Data Accepted:** Metagenomic sequences (FASTA, FASTQ)

**Output Data Provided:** Functional annotations, taxonomic profiles, comparative analysis results

**Strengths:** Web-based platform (no local installation required); Automated annotation pipeline; Large collection of public metagenomic data; Freely accessible; Comparative analysis tools

**Limitations:** Analysis pipeline may be slower than local tools; Annotation databases may not be the most current; Less standardized than MGnify; Some features may require registration; Data quality varies across deposited datasets

**Common beginner mistakes:** Not checking the annotation database versions used; Not comparing results with MGnify for consistency; Using MG-RAST as the sole analysis platform without validation

**When to Use It:** Use MG-RAST when you need automated metagenomics analysis without local computational resources, or when accessing publicly available metagenomic data with functional annotations.

**When NOT to Use It:** Do not use MG-RAST as the sole analysis platform for publication-quality analyses; use MGnify or custom pipelines for standardized results.

**Related databases / alternatives:** MGnify: More standardized metagenomics platform (recommended); NCBI SRA: Raw metagenomic data repository; SILVA: rRNA reference database



**How It Connects to Other Resources:** MG-RAST data is linked to NCBI GenBank. Functional annotations use SEED, COG, and KEGG databases.

**API / FTP / programmatic access:** REST API at <https://api.mg-rast.org/>; returns JSON. Python package mgrast available.

**Evidence/curation level:** Automated annotation; variable quality

**Data Update Status:** Regularly updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; registration required for submission

**Citation / Recommended Reference:** Meyer F et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9:386. <https://doi.org/10.1186/1471-2105-9-386>

**Beginner-Friendly Explanation:** MG-RAST is a web-based platform for analyzing metagenomic data—DNA sequences from environmental samples that contain genetic material from many different organisms. You can upload your metagenomic sequences to MG-RAST and it will automatically analyze them to identify which organisms are present and what metabolic functions they can perform. MG-RAST also stores a large collection of publicly available metagenomic datasets from diverse environments, which you can browse and compare to your own data.

**Advanced Technical Explanation:** MG-RAST implements an automated annotation pipeline that includes quality control (length filtering, duplicate removal), rRNA identification (BLAT against SILVA), protein prediction (FragGeneScan), and functional annotation (BLAT against M5NR, a comprehensive non-redundant protein database that integrates SEED, COG, KEGG, RefSeq, and other databases). The pipeline produces standardized output files including taxonomic profiles, functional profiles, and diversity metrics. MG-RAST's M5NR database provides a unified protein reference that enables consistent functional annotation across diverse metagenomic datasets.

#### **One practical workflow example:**

- Step 1: Navigate to <https://www.mg-rast.org> and register for a free account.
- Step 2: Upload your metagenomic FASTQ files for analysis.
- Step 3: Monitor the analysis progress and review quality control metrics.
- Step 4: Explore the taxonomic and functional profiles in the web interface.
- Step 5: Download the annotation results for downstream analysis.
- Step 6: Use the MG-RAST API to retrieve data programmatically and compare with MGnify results.



## Short Index Entries — Category T

### T6 – Greengenes2

**Resource Type:** Database (16S rRNA Taxonomy)

**Domain:** Microbiology / Taxonomy / Metagenomics

**Main Purpose:** Updated 16S rRNA reference database for microbial taxonomy, integrating whole-genome phylogeny with 16S rRNA sequences for improved taxonomic classification.

**Best Used For:** 16S rRNA amplicon analysis; microbial taxonomy; integration with whole-genome phylogeny.

**Key Limitation:** Newer database; less widely used than SILVA. Taxonomy may differ from SILVA and GTDB.

**Related Resources:** SILVA, GTDB (genome-based taxonomy), RDP (alternative rRNA database)

**Access / Licensing:** Open access; freely available at <http://greengenes2.ucsd.edu>.

**Citation / Documentation:** McDonald D et al. (2023). Greengenes2 unifies microbial data in a single reference tree. *Nature Biotechnology*, 41(9):1261–1264. doi:10.1038/s41587-023-01845-1

### T7 – IMG/M (Integrated Microbial Genomes and Microbiomes)

**Resource Type:** Database / Analysis Platform (Metagenomics)

**Domain:** Metagenomics / Microbial genomics / Environmental genomics

**Main Purpose:** JGI platform for comparative analysis of microbial genomes and metagenomes, providing integrated genomic and metagenomic data with functional annotations.

**Best Used For:** Comparative microbial genomics; metagenome analysis; functional annotation of microbial genomes.

**Key Limitation:** Access requires registration. JGI-focused; may not include all public metagenomes.

**Related Resources:** MGnify (EMBL-EBI metagenomics), MG-RAST (alternative platform), NCBI SRA (raw data)

**Access / Licensing:** Open access with registration; freely available at <https://img.jgi.doe.gov>.

**Citation / Documentation:** Chen IA et al. (2023). The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Research*, 51(D1):D723–D732. doi:10.1093/nar/gkac976

### Beginner EXAMPLE (Category T):

A microbiology student has performed 16S rRNA amplicon sequencing on human gut microbiome samples from healthy and diseased individuals. They process their sequences using DADA2 to generate ASVs, then classify them against the SILVA 138.1 database using QIIME2. They find that the diseased samples have reduced Firmicutes and increased Proteobacteria. They then upload their data to MGnify to compare with published gut microbiome studies.



## ADVANCED EXAMPLE (Category T):

---

A metagenomics researcher has assembled metagenome-assembled genomes (MAGs) from a novel deep-sea environment. They use GTDB-Tk to classify the MAGs and find several novel lineages with no close relatives in the GTDB database. They use MGnify to compare the functional profiles of their MAGs with published deep-sea metagenomes. They deposit their MAGs in NCBI GenBank and their metagenomic data in MGnify, contributing to the growing catalog of microbial diversity.

## CONFUSION POINTS (Category T):

---

SILVA and GTDB use different taxonomic frameworks; results may differ significantly.

16S amplicon sequencing and shotgun metagenomics require different databases and analysis approaches.

RDP has limited updates; SILVA is generally recommended for current analyses.

MGnify and MG-RAST use different analysis pipelines; results may not be directly comparable.

OTUs (Operational Taxonomic Units) and ASVs (Amplicon Sequence Variants) are different approaches to 16S data analysis.

## DECISION GUIDE (Category T):

---

Need processed metagenomic data from diverse environments? → MGnify

Need a reference database for 16S rRNA classification? → SILVA (recommended) or RDP (legacy)

Need genome-based taxonomy for MAGs? → GTDB

Need automated metagenomics analysis without local resources? → MG-RAST

Need raw metagenomic data? → NCBI SRA or ENA (linked from MGnify)

Need functional annotation of metagenomes? → MGnify or MG-RAST

## Category U: Taxonomy and Organism Databases

### OVERVIEW

Taxonomy databases provide the foundational framework for organizing biological diversity, assigning standardized names and hierarchical classifications to all known organisms. In bioinformatics, consistent taxonomic identifiers are essential for data integration across databases, comparative genomics, and ecological analyses. The major taxonomy databases—NCBI Taxonomy, UniProt Taxonomy, and the Catalogue of Life—serve as reference points for organism identification and classification across the biological sciences.

NCBI Taxonomy is the most widely used taxonomy database in bioinformatics, providing unique numeric identifiers (taxon IDs) for every organism represented in NCBI databases. These taxon IDs are used as cross-references in GenBank, RefSeq, UniProt, and hundreds of other databases, making NCBI Taxonomy the de facto standard for organism identification in computational biology. UniProt Taxonomy provides a similar service specifically for proteins in the UniProt database, with cross-references to NCBI Taxonomy. The Catalogue of Life is a broader initiative that aims to catalog all known species on Earth, providing a comprehensive reference for biodiversity research.

A key challenge in taxonomy is that biological classification is not static—new species are discovered, existing species are reclassified based on new evidence (particularly genomic data), and synonyms accumulate as the same organism is described multiple times under different names. Taxonomy databases must track these changes and provide mechanisms for resolving synonyms and tracking taxonomic history. Researchers must be aware that the same organism may have different names in different databases, and that taxonomic classifications may change over time, potentially affecting the reproducibility of analyses that depend on specific taxonomic assignments.

## U1 – NCBI Taxonomy

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/taxonomy>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** NCBI Taxonomy is used to access standardized taxonomic classifications and unique numeric identifiers (taxon IDs) for all organisms represented in NCBI databases. It is used for organism identification, taxonomic filtering of database searches, and as a cross-reference for integrating data across biological databases. NCBI Taxonomy is the de facto standard for organism identification in bioinformatics.

**What Data It Contains:** NCBI Taxonomy contains taxonomic records for over 2 million organisms, including all organisms with sequences in GenBank, RefSeq, and other NCBI databases. Each record includes the scientific name, common name, taxonomic rank, lineage, and cross-references to other databases.

**Main question it helps answer:** What is the taxonomic classification and NCBI taxon ID for this organism?

**Typical user:** Bioinformatician / Researcher / Data analyst

**Example scientific questions:**

- What is the NCBI taxon ID for Homo sapiens?
- What organisms are classified under the genus Streptococcus?
- What is the complete taxonomic lineage of E. coli K-12?

**Example use cases:** Filtering BLAST searches by taxonomic group; Cross-referencing organism information across databases; Building phylogenetic analyses with consistent taxonomic labels

**Input Data Accepted:** Organism names, taxon IDs, taxonomic ranks

**Output Data Provided:** Taxonomic records, lineages, cross-references

**Strengths:** Universal standard for organism identification in bioinformatics; Covers all organisms in NCBI databases; Freely accessible; Excellent API access; Cross-referenced by hundreds of databases

**Limitations:** Taxonomy may not always reflect current phylogenomic understanding; Some taxonomic groups are inconsistently classified; Synonyms and name changes can cause confusion; Not a comprehensive catalog of all known species

**Common beginner mistakes:** Using organism names instead of taxon IDs for programmatic access; Not checking for taxonomic synonyms when searching; Assuming NCBI taxonomy is always current

**When to Use It:** Use NCBI Taxonomy for organism identification, taxonomic filtering of database searches, and as a cross-reference for integrating data across biological databases.

**When NOT to Use It:** Do not use NCBI Taxonomy as the sole reference for phylogenetic analyses; use GTDB for bacteria/archaea or specialized phylogenetic databases.

**Related databases / alternatives:** UniProt Taxonomy: Protein-focused taxonomy; Catalogue of Life: Comprehensive species catalog; GTDB: Genome-based bacterial/archaeal taxonomy

**How It Connects to Other Resources:** NCBI Taxonomy taxon IDs are used as cross-references in GenBank, RefSeq, UniProt, STRING, and hundreds of other databases.

**API / FTP / programmatic access:** E-utilities API at <https://eutils.ncbi.nlm.nih.gov/>; Entrez taxonomy database. FTP downloads at <https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/>. Python package ete3 available for taxonomy manipulation.

**Evidence/curation level:** Manually curated; regularly reviewed.

**Data Update Status:** Continuously updated; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; public domain

**Citation / Recommended Reference:** Schoch CL et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database, 2020:baaa062. <https://doi.org/10.1093/database/baaa062>

**Beginner-Friendly Explanation:** NCBI Taxonomy is a database that assigns a unique number (called a taxon ID) to every organism that has genetic data in NCBI's databases. This number is used as a universal identifier for organisms across thousands of biological databases and tools. For example, Homo sapiens has taxon ID 9606, and this number is used consistently across GenBank, UniProt, STRING, and many other databases. NCBI Taxonomy also provides the complete taxonomic classification (kingdom, phylum, class, order, family, genus, species) for each organism, making it easy to find all organisms in a particular group.

**Advanced Technical Explanation:** NCBI Taxonomy implements a hierarchical tree structure with nodes representing taxonomic units at various ranks (superkingdom, kingdom, phylum, class, order, family, genus, species, and many intermediate ranks). Each node has a unique taxon ID, scientific name, and rank. The taxonomy is maintained by NCBI curators and updated based on published taxonomic revisions. The E-utilities API provides programmatic access to taxonomy data through the efetch, esearch, and elink utilities. The taxonomy dump files (available via FTP) include nodes.dmp (taxonomic structure), names.dmp (scientific and common names), and merged.dmp (merged taxon IDs).

### One practical workflow example:

Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/taxonomy> and search for your organism.

Step 2: Note the taxon ID for use in downstream analyses.

Step 3: Use the taxon ID to filter BLAST searches: add "txid9606[Organism]" to your BLAST query.

Step 4: Use the E-utilities API to retrieve taxonomy data programmatically: <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=taxonomy&id=9606&rettype=xml>

Step 5: Use the ete3 Python package for taxonomy manipulation and visualization.

Step 6: Cross-reference with UniProt Taxonomy for protein-specific queries.

## U2 – UniProt Taxonomy

**Official Website URL:** <https://www.uniprot.org/taxonomy>

**Resource Type:** Database

**Main Biological Domain:** Proteins / Omics

**What It Is Used For:** UniProt Taxonomy is used to access taxonomic information for organisms represented in the UniProt protein database, providing a protein-centric view of organism taxonomy. It is used for filtering protein searches by organisms, understanding the taxonomic distribution of protein families, and cross-referencing with NCBI Taxonomy.

**What Data It Contains:** UniProt Taxonomy contains taxonomic records for all organisms with proteins in UniProt, with cross-references to NCBI Taxonomy, scientific names, common names, and lineages. The database is synchronized with NCBI Taxonomy.

**Main question it helps answer:** What proteins are available in UniProt for this organism or taxonomic group?

**Typical user:** Bioinformatician / Researcher

**Example scientific questions:**

- What proteins are available in UniProt for *Arabidopsis thaliana*?
- What is the taxonomic distribution of this protein family?
- How many reviewed (Swiss-Prot) proteins are available for this organism?

**Example use cases:**

- Filtering UniProt searches by organism or taxonomic group
- Understanding the taxonomic coverage of a protein family
- Cross-referencing organism information with NCBI Taxonomy

**Input Data Accepted:** Organism names, NCBI taxon IDs, UniProt taxonomy IDs.

**Output Data Provided:** Taxonomic records, protein counts, cross-references.

**Strengths:** Integrated with UniProt protein database; Synchronized with NCBI Taxonomy; Freely accessible; Useful for protein-centric taxonomy queries

**Limitations:** Only covers organisms with proteins in UniProt; Less comprehensive than NCBI Taxonomy for all organisms; Taxonomy follows NCBI Taxonomy (same limitations apply)

**Common beginner mistakes:** Not using taxonomy filters when searching UniProt for organism-specific proteins; Confusing UniProt taxonomy IDs with NCBI taxon IDs (they are the same)

**When to Use It:** Use UniProt Taxonomy when filtering UniProt protein searches by organism or when understanding the taxonomic distribution of protein families.

**When NOT Use It:** For comprehensive taxonomy queries, use NCBI Taxonomy directly.

**Related databases / alternatives:** NCBI Taxonomy: Comprehensive taxonomy reference; Catalogue of Life: Comprehensive species catalog

**How It Connects to Other Resources:** UniProt Taxonomy is synchronized with NCBI Taxonomy. Taxonomy IDs are used as cross-references in UniProt protein records.

**API / FTP / programmatic access:** UniProt REST API at <https://rest.uniprot.org/taxonomy/>; returns JSON or TSV. Python package uniprot available.

**Evidence/curation level:** Synchronized with NCBI Taxonomy; manually reviewed.

**Data Update Status:** Synchronized with NCBI Taxonomy updates; actively maintained.

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** The UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1):D523–D531. <https://doi.org/10.1093/nar/gkac1052>

**Beginner-Friendly Explanation:** UniProt Taxonomy is the taxonomy system used within the UniProt protein database. It is closely synchronized with NCBI Taxonomy, using the same taxon IDs, but provides a protein-centric view—showing how many proteins are available for each organism and allowing you to filter protein searches by organism or taxonomic group. For example, you can search UniProt for all reviewed proteins from mammals or find all proteins from a specific bacterial species. UniProt Taxonomy is most useful when you are working with protein data and want to understand the taxonomic distribution of proteins.

**Advanced Technical Explanation:** UniProt Taxonomy is maintained in synchrony with NCBI Taxonomy, using the same taxon IDs and hierarchical structure. The UniProt REST API provides programmatic access to taxonomy data, including protein counts for each taxon (reviewed/Swiss-Prot and unreviewed/TrEMBL). Taxonomy filters in UniProt queries use the "taxonomy\_id" field, which accepts NCBI taxon IDs. The taxonomy hierarchy is used for "lineage" queries that retrieve all proteins from a taxonomic group and all its descendants.

**One practical workflow example:**

- Step 1: Navigate to <https://www.uniprot.org/taxonomy> and search for your organism.
- Step 2: Note the taxon ID and the number of reviewed/unreviewed proteins.
- Step 3: Click "Proteins" to see all UniProt proteins for this organism.
- Step 4: Use the taxon ID in a UniProt search: taxonomy\_id:9606 AND reviewed:true.
- Step 5: Use the REST API to retrieve taxonomy data: <https://rest.uniprot.org/taxonomy/9606>.
- Step 6: Cross-reference with NCBI Taxonomy for additional organism information.

## U3 – Catalogue of Life

**Official Website URL:** <https://www.catalogueoflife.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Omics / Systems biology

**What It Is Used For:** The Catalogue of Life is used to access a comprehensive catalog of all known species on Earth, providing standardized species names, taxonomic classifications, and distribution information. It is used for biodiversity research, species identification, and as a reference for comprehensive species catalogs. The Catalogue of Life aims to catalog all ~8.7 million estimated species on Earth.

**What Data It Contains:** The Catalogue of Life contains records for over 4 million species from all kingdoms of life, with standardized scientific names, synonyms, taxonomic classifications, and distribution information. Data is compiled from over 170 specialist databases covering different taxonomic groups.

**Main question it helps answer:** Is this species name valid, and what is its accepted taxonomic classification?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- Does this species have a valid accepted name or a synonym?
- What is the complete taxonomic classification of this species?
- How many species are known in this genus?

**Example use cases:**

- Validating species names for biodiversity analyses
- Resolving taxonomic synonyms
- Accessing comprehensive species catalogs for ecological analyses

**Input Data Accepted:** Species names, taxonomic group names.

**Output Data Provided:** Species records, taxonomic classifications, synonyms, distribution data.

**Strengths:** Most comprehensive species catalog available; Covers all kingdoms of life; Resolves synonyms and provides accepted names; Freely accessible; Compiled from specialist databases

**Limitations:** Coverage is incomplete (not all species have been described); Some taxonomic groups are better covered than others; Taxonomy may lag primary literature; Less focused on molecular data than NCBI Taxonomy

**Common beginner mistakes:** Using the Catalogue of Life for molecular biology analyses (use NCBI Taxonomy instead); Not checking for synonyms when searching for species; Assuming all species are represented in the catalog.

**When to Use It:** Use the Catalogue of Life for biodiversity research, species name validation, and comprehensive species catalogs. Useful for ecological and macroevolutionary analyses.

**When NOT Use It:** Do not use the Catalogue of Life for molecular biology analyses; use NCBI Taxonomy instead. The Catalogue of Life is best for biodiversity and ecological research.

**Related databases / alternatives:** NCBI Taxonomy: Molecular biology taxonomy reference; GBIF: Global Biodiversity Information Facility; iNaturalist: Citizen science biodiversity data



**How It Connects to Other Resources:** The Catalogue of Life is linked to GBIF, ITIS, and other biodiversity databases. Species records include cross-references to specialist databases.

**API / FTP / programmatic access:** API at <https://api.catalogueoflife.org/>; returns JSON. Bulk downloads available.

**Evidence/curation level:** Compiled from specialist databases; quality varies by taxonomic group.

**Data Update Status:** Annual releases; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Bánki O et al. (2022). Catalogue of Life Checklist (Version 2022-11-14). <https://doi.org/10.48580/dfp8>

**Beginner-Friendly Explanation:** The Catalogue of Life is an ambitious project that aims to catalog every known species on Earth. It compiles information from over 170 specialist databases covering different groups of organisms—from bacteria to plants to insects to mammals—and provides a single, searchable resource for species names and classifications. The Catalogue of Life is particularly useful for checking whether a species name is valid (or whether it is a synonym for another name), and for finding the complete taxonomic classification of any known species. It is primarily used in biodiversity and ecological research rather than molecular biology.

**Advanced Technical Explanation:** The Catalogue of Life implements a species-level taxonomy compiled from specialist databases (Global Species Databases, GSDs) that cover specific taxonomic groups. The catalog uses a standardized data model that captures accepted names, synonyms, taxonomic status, and distribution information. The Catalogue of Life ChecklistBank provides programmatic access to the catalog data through a REST API. The catalog is used as the taxonomic backbone for GBIF (Global Biodiversity Information Facility) and other biodiversity databases.

#### **One practical workflow example:**

Step 1: Navigate to <https://www.catalogueoflife.org> and search for your species of interest.

Step 2: Check whether the name is accepted or a synonym.

Step 3: Note the accepted name and taxonomic classification.

Step 4: Use the API to retrieve species data programmatically:  
<https://api.catalogueoflife.org/dataset/3LR/taxon/search?q=Homo+sapiens>.

Step 5: Cross-reference with NCBI Taxonomy for molecular biology data.

Step 6: Use the Catalogue of Life as a reference for species name validation in biodiversity analyses.

## BEGINNER EXAMPLE (Category U):

---

A bioinformatics student wants to download all reviewed human proteins from UniProt. They navigate to UniProt Taxonomy, find the taxon ID for Homo sapiens (9606), and use it to filter their UniProt search: "taxonomy\_id:9606 AND reviewed:true." They find over 20,000 reviewed human proteins. They then check NCBI Taxonomy to confirm the taxon ID and find the complete taxonomic lineage.

## ADVANCED EXAMPLE (Category U):

---

A comparative genomicist is studying the evolution of a protein family across vertebrates. They use NCBI Taxonomy to retrieve all taxon IDs for vertebrates (taxon ID 7742 and all descendants), then use the E-utilities API to retrieve all RefSeq protein sequences for this family from vertebrate organisms. They use the ete3 Python package to build a species tree from NCBI Taxonomy and map the protein family distribution onto the tree.

## CONFUSION POINTS (Category U):

---

- NCBI Taxonomy and UniProt Taxonomy use the same taxon IDs (synchronized).
- The Catalogue of Life uses different IDs from NCBI Taxonomy.
- GTDB taxonomy for bacteria/archaea may differ significantly from NCBI Taxonomy.
- Taxonomic names change over time; always check for synonyms.
- "Species" in NCBI Taxonomy may include strains and subspecies.

## DECISION GUIDE (Category U):

---

Need taxon IDs for bioinformatics analysis? → NCBI Taxonomy

Need to filter UniProt protein searches by organism? → UniProt Taxonomy

Need comprehensive species catalog for biodiversity research? → Catalogue of Life

Need genome-based taxonomy for bacteria/archaea? → GTDB

Need to validate species names? → Catalogue of Life

## Category V: Antimicrobial Peptide and Peptide Databases

### OVERVIEW

Antimicrobial peptides (AMPs) are short peptides (typically 10–50 amino acids) that exhibit activity against bacteria, fungi, viruses, and parasites. They represent a promising class of therapeutic agents in the face of growing antibiotic resistance, and their study has expanded significantly over the past two decades. Dedicated AMP databases have been developed to catalog the growing number of characterized antimicrobial peptides, providing sequence, structure, activity, and mechanism information for researchers working in this field.

The major AMP databases—DRAMP, APD/APD6, dbAMP, CAMPR4, and DBAASP—differ in their scope, curation approach, and the types of information they provide. DRAMP (Data Repository of Antimicrobial Peptides) is one of the most comprehensive, covering natural, synthetic, and patented AMPs with detailed activity and structural information. APD/APD6 is the original Antimicrobial Peptide Database platform; APD3 is a historical version name, while APD6 is the current platform and includes natural, synthetic, and predicted AMPs. APD counts must be cited with source and date. DBAASP (Database of Antimicrobial Activity and Structure of Peptides) provides quantitative activity data with standardized assay conditions, making it particularly valuable for computational modeling.

A key challenge in AMP research is the diversity of experimental conditions used to measure antimicrobial activity. Minimum inhibitory concentration (MIC) values—the standard measure of antimicrobial potency—can vary by orders of magnitude depending on the bacterial strain, growth medium, inoculum size, and incubation conditions. This variability makes it difficult to compare activity data across studies and databases. Researchers must be aware of these limitations when using AMP databases for computational modeling or drug discovery. Additionally, the distinction between natural AMPs (derived from organisms) and synthetic AMPs (designed computationally or chemically) is important for understanding the biological context of the data.

## V1 – DRAMP (Data Repository of Antimicrobial Peptides)

**Official Website URL:** <https://dramp.cpu-bioinfor.org>

**Resource Type:** Database / Repository

**Main Biological Domain:** Antimicrobial peptides / Proteins

**What It Is Used For:** DRAMP is used to access a comprehensive repository of antimicrobial peptides, including natural, synthetic, and patented AMPs, with detailed sequence, structure, activity, and mechanism information. It is used for AMP discovery, computational modeling of AMP activity, and understanding the diversity of antimicrobial peptides. DRAMP is one of the most comprehensive AMP databases available.

**What Data It Contains:** DRAMP contains over 22,000 AMP entries, including natural AMPs from diverse organisms, synthetic AMPs, and patented peptides, with information on sequence, structure (where available), antimicrobial activity (MIC values), target organisms, mechanism of action, and physicochemical properties.

**Main question it helps answer:** What antimicrobial peptides are known, and what are their activities against specific pathogens?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What AMPs are active against MRSA?
- What are the physicochemical properties of known AMPs?
- What natural AMPs have been isolated from frogs?

**Example use cases:**

- Building machine learning models for AMP activity prediction
- Identifying AMPs with activity against specific pathogens
- Analyzing the physicochemical properties of AMPs

**Data Accepted Input:** Peptide sequences, organism names, target pathogen names.

**Output Data Provided:** AMP records with sequence, activity, and structural information.

**Strengths:** Comprehensive coverage of natural, synthetic, and patented AMPs; Detailed activity data; Physicochemical property calculations; Freely accessible; Regularly updated

**Limitations:** Activity data from different assays may not be directly comparable; Some entries may have incomplete information; Coverage of synthetic AMPs may be incomplete; MIC values vary with assay conditions

**Common beginner mistakes:** Comparing MIC values from different assays without considering conditions; Not filtering by activity type when searching; Using DRAMP as the sole source for AMP data.

**When to Use It:** Use DRAMP for comprehensive AMP data, particularly for building machine learning models or analyzing the diversity of antimicrobial peptides.

**When NOT to Use It:** Do not use DRAMP as the sole source for quantitative activity data; cross-reference with DBAASP for standardized activity measurements.

**Related databases / alternatives:** APD/APD6: current APD platform; natural, synthetic, and predicted AMPs; cite counts with source/date; DBAASP: Quantitative activity data; CAMPR4: alternative AMP database and prediction platform.

**How It Connects to Other Resources:** DRAMP cross-references UniProt, PubMed, and structural databases. Sequences are linked to NCBI protein records.

**API / FTP / programmatic access:** Data downloads available from the DRAMP website. Limited API access.

**Evidence/curation level:** Manually curated from primary literature; moderate quality.

**Data Update Status:** Regular updates; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Shi G et al. (2022). DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Research*, 50(D1):D488–D496. <https://doi.org/10.1093/nar/gkab651>

**Beginner-Friendly Explanation:** DRAMP is a database that collects information about antimicrobial peptides—short proteins that can kill bacteria, fungi, and other pathogens. It is one of the most comprehensive collections of antimicrobial peptides available, covering both natural peptides found in organisms (like defensins in human skin) and synthetic peptides designed in the laboratory. For each peptide, DRAMP provides the amino acid sequence, information about which pathogens it is active against, and its physical and chemical properties. This makes it a valuable resource for researchers developing new antibiotics or studying how antimicrobial peptides work.

**Advanced Technical Explanation:** DRAMP implements a comprehensive data model that captures peptide sequence, secondary structure (where available from NMR or X-ray crystallography), antimicrobial activity (MIC values against specific organisms), hemolytic activity (toxicity to red blood cells), mechanism of action (membrane disruption, intracellular targets, etc.), and physicochemical properties (charge, hydrophobicity, amphipathicity, isoelectric point). DRAMP's data is widely used for training machine learning models for AMP activity prediction, including deep learning models like AMPpred and iAMP-2L.

**One practical workflow example:**

- Step 1: Navigate to <https://dramp.cpu-bioinfor.org> and search for AMPs active against your target pathogen.
- Step 2: Filter by activity type (antibacterial, antifungal, antiviral) and target organism.
- Step 3: Download the AMP sequences and activity data in FASTA or CSV format.
- Step 4: Calculate physicochemical properties using the built-in tools or external tools (e.g., Peptides R package).
- Step 5: Use the data to train a machine learning model for AMP activity prediction.
- Step 6: Cross-reference with APD/APD6 and DBAASP for additional data.

## V2 – APD/APD6 (Antimicrobial Peptide Database; APD3 is historical, APD6 is the current platform)

**Official Website URL:** <https://aps.unmc.edu>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Antimicrobial peptides / Proteins

**What Is Used For:** APD/APD6 is used to access a curated antimicrobial peptide database including natural, synthetic, and predicted AMPs, with information on sequence, structure, activity, mechanism, physicochemical properties, and AMP classification. APD3 should be cited only when referring to the older 2016 APD3 publication/version; current use should refer to APD6.

**What Data It Contains:** APD6 counts must be cited with date and source. The APD6 NAR paper reported 5,188 peptide records as of 18 March 2025, while the official APD website reported 6,309 peptides as of 1 January 2026, including natural, synthetic, and AI-predicted AMPs. Do not mix these numbers without source/date.

**Main question it helps answer:** What natural antimicrobial peptides are known, and what are their mechanisms of action?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What natural AMPs have been characterized by insects?
- What is the mechanism of action of this AMP?
- What AMPs have known 3D structures?

**Example use cases:**

- Studying the diversity of natural AMPs
- Understanding AMP mechanisms of action
- Finding AMPs with known 3D structures for structural analysis

**Input Data Accepted:** Peptide sequences, organism names, AMP names.

**Output Data Provided:** AMP records with sequence, structure, activity, and mechanism information.

**Strengths:** Focus on natural, experimentally characterized AMPs; Detailed mechanism of action information; Includes 3D structure information; Freely accessible; Long-established database

**Limitations:** Smaller than DRAMP (focuses on natural AMPs only); Does not include synthetic or patented AMPs; Some entries may have incomplete information; Less comprehensive than DRAMP for overall AMP coverage

**Common beginner mistakes:** Using APD/APD6 as the sole source for AMP data (supplement with DRAMP); Not checking the mechanism of action information; Not using APD3's analysis tools for physicochemical property calculation.

**When to Use It:** Use APD3 when you need detailed information about natural AMPs, particularly their mechanisms of action and 3D structures.

**When NOT Use It:** Do not use APD/APD6 as the sole source for comprehensive AMP data; use DRAMP for broader coverage including synthetic AMPs.

**Related databases / alternatives:** DRAMP: Comprehensive AMP database; DBAASP: Quantitative activity data; CAMPR4: alternative AMP database and prediction platform

**How It Connects to Other Resources:** APD3 cross-references UniProt, PDB (for 3D structures), and PubMed. Sequences are linked to NCBI protein records.

**API / FTP / programmatic access:** Data downloads available from the APD3 website. Limited API access.

**Evidence/curation level:** Manually curated from primary literature; high quality for natural AMPs.

**Data Update Status:** Regular updates; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Wang G et al. (2016). APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44(D1):D1087–D1093. <https://doi.org/10.1093/nar/gkv1278>

**Beginner-Friendly Explanation:** APD/APD6 (Antimicrobial Peptide Database; APD3 is historical) is one of the original databases for antimicrobial peptides, focusing specifically on natural AMPs—peptides that organisms produce to defend themselves against pathogens. For each peptide, APD3 provides detailed information about its amino acid sequence, three-dimensional structure (when available), which pathogens it kills, and how it kills them (its mechanism of action). APD3 is particularly valuable for researchers who want to understand the biology of natural antimicrobial peptides and use this knowledge to design new antibiotics.

**Advanced Technical Explanation:** APD/APD6 implements a comprehensive data model for AMPs that captures sequence (with disulfide bond annotations), secondary structure (alpha-helix, beta-sheet, coil), 3D structure (PDB cross-references), antimicrobial activity (MIC values), hemolytic activity (HC50 values), mechanism of action (membrane disruption, DNA binding, enzyme inhibition, etc.), and physicochemical properties (net charge, hydrophobicity, amphipathicity). APD3 provides a web-based tool for calculating physicochemical properties of user-submitted peptide sequences and for predicting AMP activity using a statistical model trained on APD3 data.

#### **One practical workflow example:**

Step 1: Navigate to <https://aps.unmc.edu> and search for AMPs from your organism of interest.

Step 2: Review the mechanism of action information for AMPs of interest.

Step 3: Check for 3D structure availability (PDB cross-references).

Step 4: Download the AMP sequences for computational analysis.

Step 5: Use APD3's online tools to calculate physicochemical properties of your peptide.

Step 6: Cross-reference with DRAMP for additional AMP data.



## V3 – dbAMP

**Official Website URL:** <https://dbamp.kuicr.kyoto-u.ac.jp>

**Resource Type:** Database

**Main Biological Domain:** Antimicrobial peptides / Proteins

**What It Is Used For:** dbAMP is used to access a database of antimicrobial peptides with experimental activity data and physicochemical properties. NOTE: The current operational status of dbAMP should be verified before use, as the database may have limited maintenance. For current AMP data, DRAMP or APD/APD6 are recommended as primary resources.

**What Data It Contains:** dbAMP contains AMP entries with sequence, activity data, and physicochemical properties, compiled from primary literature.

**Main question helps answer:** What antimicrobial peptides are available in the dbAMP database?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:** What AMPs are available in dbAMP for comparison with other databases?

**Example use cases:**

- Cross-referencing AMP data across databases
- Historical benchmarking of AMP prediction methods

**Input Data Accepted:** Peptide sequences, AMP names.

**Output Data Provided:** AMP records with sequence and activity information.

**Strengths:** Historical AMP data; Freely accessible (when operational)

**Limitations:** Operational status uncertain; verify before use; May have limited maintenance; Smaller than DRAMP or APD/APD6; Not recommended as primary resource

**Common beginner mistakes:**

- Using dbAMP as a primary resource without verifying its status
- Not cross-referencing with DRAMP or APD/APD6

**When to Use It:** Use dbAMP only as a supplementary resource or for historical benchmarking. Verify operational status before use.

**When NOT Use It:** Do not use dbAMP as the primary AMP database; use DRAMP or APD/APD6 instead.

**Related databases / alternatives:**

- DRAMP: Comprehensive AMP database (recommended)
- APD/APD6: current APD platform; natural, synthetic, and predicted AMPs; cite counts with source/date (recommended)
- CAMPR4: alternative AMP database and prediction platform

**How It Connects to Other Resources:** dbAMP cross-references UniProt and PubMed.

**API / FTP / programmatic access:** Limited; verify status.

**Evidence/curation level:** Manually curated; limited recent updates.

**Data Update Status:** Uncertain; verify status.

**Licensing / access restrictions:** Freely available for academic use (when operational)

**Citation / Recommended Reference:** Jhong JH et al. (2019). dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Research*, 47(D1):D285–D297. <https://doi.org/10.1093/nar/gky1030>

**Beginner-Friendly Explanation:** dbAMP is a database of antimicrobial peptides that was developed at Kyoto University. It provides information about AMP sequences, activities, and physicochemical properties. However, the current operational status of dbAMP is uncertain, and researchers should verify that the database is accessible and up to date before using it. For current AMP research, DRAMP or APD/APD6 are more reliable primary resources.

**Advanced Technical Explanation:** dbAMP was designed to integrate AMP data with transcriptome and proteome data, providing a broader context for AMP expression and function. The database includes physicochemical property calculations and tools for AMP prediction. However, given the uncertain maintenance status, researchers should use DRAMP or APD/APD6 as primary resources and treat dbAMP data as supplementary.

**One practical workflow example:**

- Step 1: Verify that <https://dbamp.kuicr.kyoto-u.ac.jp> is accessible.
- Step 2: If accessible, search for your AMP of interest.
- Step 3: Cross-reference all dbAMP data with DRAMP and APD/APD6.
- Step 4: For current research, use DRAMP or APD/APD6 as the primary resource.
- Step 5: Report the database version and access date for reproducibility.

## V4 – CAMPR4 (Collection of Anti-Microbial Peptides Release 4; CAMP/CAMPR4 are historical names)

**Official Website URL:** <https://camp.bicnirrh.res.in>

**Resource Type:** Database / Tool

**Main Biological Domain:** Antimicrobial peptides / Proteins

**What It Is Used For:** CAMPR4 is used to access curated information on natural and synthetic antimicrobial peptides and to use AMP prediction and analysis tools. It includes AMP sequences, structures, patents, family signatures, source organisms, target organisms, modifications, and prediction algorithms.

**What Data It Contains:** CAMPR4 has been reported to contain 24,243 AMP sequences, 933 structures, 2,143 patents, and 263 AMP family signatures, along with prediction tools and curated AMP metadata. Cite the CAMPR4 paper and access date when using these numbers.

**Main question it helps answer:** What antimicrobial peptides are known, and can I predict whether my peptide has antimicrobial activity?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- Does my peptide sequence have predicted antimicrobial activity?
- What AMPs are active against gram-negative bacteria?
- What are the physicochemical properties of known AMPs?

**Example use cases:**

- Predicting AMP activity for novel peptide sequences
- Accessing AMP data for computational modeling
- Analyzing physicochemical properties of AMPs

**Input Data Accepted:** Peptide sequences, AMP names.

**Output Data Provided:** AMP records, activity predictions, physicochemical properties.

**Strengths:** Includes AMP prediction tools; Multiple machine learning classifiers; Freely accessible; Covers diverse AMP types

**Limitations:** Prediction accuracy varies for novel peptides; Some entries may have incomplete information; Less comprehensive than DRAMP; Website availability may be intermittent

**Common beginner mistakes:** Treating CAMPR4 predictions as definitive without experimental validation; Not cross-referencing with DRAMP for comprehensive data.

**When to Use It:** Use CAMP when you need AMP prediction tools for novel peptide sequences, or when accessing AMP data for computational modeling.

**When NOT Use It:** Do not use CAMP as the sole source for AMP data; use DRAMP for comprehensive coverage.

**Related databases / alternatives:** DRAMP: Comprehensive AMP database; APD/APD6: current APD platform; natural, synthetic, and predicted AMPs; cite counts with source/date; DBAASP: Quantitative activity data.

**How It Connects to Other Resources:** CAMPR4 cross-references UniProt and PubMed.

**API / FTP / programmatic access:** Data downloads available from the CAMPR4 website. Limited API access.

**Evidence/curation level:** Manually curated; moderate quality.

**Data Update Status:** Periodic updates; maintained as of 2024.

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Gawde U et al. (2023). CAMPR4: a database of natural and synthetic antimicrobial peptides. Nucleic Acids Research, 51(D1):D377-D383. <https://doi.org/10.1093/nar/gkac933>

**Beginner-Friendly Explanation:** CAMPR4 (Collection of Anti-Microbial Peptides Release 4) is a database and analysis platform for antimicrobial peptides. In addition to providing a collection of known AMPs with their sequences and activities, CAMP offers tools for predicting whether a new peptide sequence might have antimicrobial activity. These prediction tools use machine learning algorithms trained on known AMPs to evaluate new sequences. This makes CAMP particularly useful for researchers who have designed new peptides and want to assess their potential antimicrobial activity before conducting laboratory experiments.

**Advanced Technical Explanation:** CAMPR4 integrates curated AMP sequence data with structures, patents, family signatures, and prediction tools; earlier CAMP releases implemented multiple machine-learning classifiers for AMP prediction, including Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), and Discriminant Analysis (DA). Each classifier is trained on a curated set of AMPs and non-AMPs, with features derived from physicochemical properties (amino acid composition, dipeptide composition, physicochemical indices). CAMP also provides tools for calculating physicochemical properties (charge, hydrophobicity, amphipathicity) and for structural analysis of AMPs.

**One practical workflow example:**

Step 1: Navigate to <https://camp.bicnirrh.res.in> and access the prediction tools.

Step 2: Submit your peptide sequence for AMP activity prediction.

Step 3: Review the predictions from multiple classifiers (SVM, ANN, RF, DA).

Step 4: Check the physicochemical properties of your peptide.

Step 5: Cross-reference with DRAMP and APD/APD6 for similar known AMPs.

Step 6: Validate predictions experimentally using MIC assays.

## V5 – DBAASP (Database of Antimicrobial Activity and Structure of Peptides)

---

**Official Website URL:** <https://dbaasp.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Antimicrobial peptides / Proteins

**What It Is Used For:** DBAASP is used to access quantitative antimicrobial activity data for peptides, with standardized assay conditions and detailed structural information. It is particularly valuable for computational modeling of AMP activity because of its focus on quantitative, standardized data. DBAASP provides detailed information on peptide structure, activity, and toxicity.

**What Data It Contains:** DBAASP contains over 15,000 peptide entries with quantitative antimicrobial activity data (MIC values), hemolytic activity, cytotoxicity, and structural information (secondary structure, 3D structure where available). The database emphasizes standardized activity measurements and detailed assay conditions.

**Main question it helps answer:** What is the quantitative antimicrobial activity of this peptide under standardized conditions?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What is the MIC of this peptide against E. coli under standardized conditions?
- What peptides have both high antimicrobial activity and low hemolytic activity?
- What structural features correlate with antimicrobial activity?

**Example use cases:**

- Building quantitative QSAR models for AMP activity prediction
- Identifying AMPs with favorable therapeutic indices
- Analyzing structure-activity relationships for AMPs

**Input Data Accepted:** Peptide sequences, AMP names, target organism names.

**Output Data Provided:** Quantitative activity data, structural information, toxicity data.

**Strengths:** Focus on quantitative, standardized activity data; Detailed assay condition information; Includes toxicity data (hemolytic activity, cytotoxicity); Freely accessible; Valuable for computational modeling

**Limitations:** Smaller than DRAMP; Some assay conditions may still vary; Coverage of synthetic AMPs may be incomplete

**Common beginner mistakes:** Not checking assay conditions when comparing MIC values; Not considering hemolytic activity when evaluating AMPs; Using DBAASP as the sole source for AMP data.

**When to Use It:** Use DBAASP when you need quantitative, standardized activity data for computational modeling or when evaluating the therapeutic potential of AMPs (activity vs. toxicity).

**When NOT Use It:** Do not use DBAASP as the sole source for comprehensive AMP data; use DRAMP for broader coverage.

**Related databases / alternatives:**

- DRAMP: Comprehensive AMP database
- APD/APD6: current APD platform; natural, synthetic, and predicted AMPs; cite counts with source/date.
- CAMPR4: AMP prediction tools

**How It Connects to Other Resources:** DBAASP cross-references UniProt, PDB, and PubMed.

**API / FTP / programmatic access:** Data downloads available from <https://dbaasp.org/download>. Limited API access.

**Evidence/curation level:** Manually curated from primary literature; high quality for quantitative data.

**Data Update Status:** Regular updates; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available for academic use

**Citation / Recommended Reference:** Pirtsckhalava M et al. (2021). DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Research, 49(D1):D288–D297. <https://doi.org/10.1093/nar/gkaa991>

**Beginner-Friendly Explanation:** DBAASP (Database of Antimicrobial Activity and Structure of Peptides) is a database that focuses on providing high-quality, quantitative data about how well antimicrobial peptides work against pathogens. What makes DBAASP special is its emphasis on standardized measurements and detailed information about the experimental conditions used to measure activity. This is important because the same peptide can appear to have very different activity depending on how the experiment was done. DBAASP also includes information about toxicity (how harmful the peptide is to human cells), which is crucial for evaluating whether a peptide could be used as a therapeutic.

**Advanced Technical Explanation:** DBAASP implements a comprehensive data model for quantitative AMP activity that captures MIC values with units, target organism (with NCBI taxon ID), growth medium, inoculum size, incubation conditions, and assay method. This level of detail enables meta-analyses that account for assay condition variability. DBAASP also captures hemolytic activity (HC50 values against human red blood cells) and cytotoxicity (IC50 values against mammalian cell lines), enabling calculation of therapeutic index (MIC/HC50 or MIC/IC50). The database is widely used for training quantitative QSAR models for AMP activity prediction.

**One practical workflow example:**

- Step 1: Navigate to <https://dbaasp.org> and search for AMPs active against your target pathogen.
- Step 2: Filter by assay conditions (e.g., CLSI standard conditions) for consistent data.
- Step 3: Download the activity data (MIC values) and hemolytic activity data.
- Step 4: Calculate the therapeutic index (MIC/HC50) for each AMP.
- Step 5: Use the data to build a QSAR model for AMP activity prediction.
- Step 6: Cross-reference with DRAMP for additional AMP data.

## BEGINNER EXAMPLE (Category V):

---

A pharmacology student wants to find antimicrobial peptides active against MRSA (methicillin-resistant *Staphylococcus aureus*). They search DRAMP for AMPs with activity against *S. aureus* and find over 500 entries. They filter for AMPs with MIC < 10 µg/mL and download the sequences. They then use CAMP to predict the activity of a novel peptide they have designed, and check DBAASP for the hemolytic activity of similar peptides.

## ADVANCED EXAMPLE (Category V):

---

A computational chemist is building a machine learning model to predict AMP activity against gram-negative bacteria. They download all AMP sequences with MIC data against *E. coli* from DRAMP and DBAASP, filter for standardized assay conditions, and calculate physicochemical features (charge, hydrophobicity, amphipathicity, secondary structure propensity). They train a gradient boosting model and validate it on a held-out test set. They use the model to screen a library of novel peptide sequences and identify candidates for experimental validation.

## CONFUSION POINTS (Category V):

---

- MIC values from different assays are not directly comparable; always check assay conditions.
- Hemolytic activity (toxicity to red blood cells) is different from cytotoxicity (toxicity to mammalian cells).
- Natural AMPs and synthetic AMPs may have different mechanisms of action.
- dbAMP operational status is uncertain; verify before use.
- AMP prediction tools have limited accuracy for novel peptide sequences; experimental validation is essential.

## DECISION GUIDE (Category V):

---

Need comprehensive AMP data including synthetic and patented peptides? → DRAMP

Need detailed information about natural AMP mechanisms? → APD3

Need quantitative, standardized activity data for computational modeling? → DBAASP

Need AMP prediction tools for novel sequences? → CAMP

Need to cross-reference AMP data across databases? → Use multiple databases.



## Category W: Cancer Genomics Databases

### OVERVIEW

Cancer genomics databases have transformed our understanding of the molecular basis of cancer by cataloging the somatic mutations, copy number alterations, gene expression changes, and epigenomic modifications that drive tumor development and progression. The large-scale cancer genomics projects—TCGA (The Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium)—have generated comprehensive molecular profiles for thousands of tumors across dozens of cancer types, providing an unprecedented view of cancer genomic diversity. These datasets have enabled the identification of cancer driver genes, the classification of tumors into molecular subtypes, and the discovery of potential therapeutic targets.

The cancer genomics database landscape includes both primary data repositories and analysis portals. TCGA and ICGC provide access to raw and processed genomic data from large-scale cancer sequencing projects. cBioPortal provides an interactive platform for exploring and analyzing cancer genomics data, integrating data from TCGA, ICGC, and many other studies. COSMIC (Catalogue of Somatic Mutations in Cancer) provides a comprehensive catalog of somatic mutations in cancer, curated from primary literature and large-scale sequencing projects. OncoKB provides a precision oncology knowledge base that annotates cancer mutations with clinical significance and therapeutic implications.

A critical consideration when working with cancer genomics data is the distinction between somatic mutations (acquired during tumor development) and germline variants (inherited from parents). Cancer genomics databases focus on somatic mutations, which are the drivers of tumor development. However, germline variants can also affect cancer risk and treatment response, and researchers must be careful to distinguish between these two types of variation. Additionally, the interpretation of cancer mutations requires knowledge of the functional significance of each mutation—whether it is a driver mutation that contributes to tumor development or a passenger mutation that is simply present in the tumor without contributing to its growth.

## W1 – TCGA (The Cancer Genome Atlas)

**Official Website URL:** <https://www.cancer.gov/tcga>

**Resource Type:** Database / Repository / Dataset Collection

**Main Biological Domain:** Clinical genomics / Variants / Omics

**What It Is Used For:** TCGA is used to access comprehensive molecular profiles (genomic, transcriptomic, epigenomic, proteomic) for over 11,000 tumors across 33 cancer types, providing a foundational resource for cancer genomics research. It is used for identifying cancer driver genes, classifying tumors into molecular subtypes, discovering biomarkers, and understanding the molecular basis of cancer. TCGA data is accessible through the GDC (Genomic Data Commons) portal.

**What Data It Contains:** TCGA contains molecular profiles for 11,000+ tumors across 33 types of cancer, including whole-exome sequencing (somatic mutations), copy number variation, RNA-seq (gene expression), DNA methylation, miRNA expression, and RPPA (protein expression). Clinical data including survival, treatment, and pathology information is also available.

**Main question it helps answer:** What are the molecular alterations in this cancer type, and how do they relate to clinical outcomes?

**Typical user:** Researcher / Bioinformatician / Data analyst / Clinician

**Example scientific questions:**

- What are the most frequently mutated genes in breast cancer?
- What molecular subtypes of glioblastoma are defined by gene expression?
- What mutations are associated with poor prognosis in lung adenocarcinoma?

**Example use cases:** Identifying cancer driver genes through mutation frequency analysis; Classifying tumors into molecular subtypes; Discovering biomarkers for prognosis or treatment response.

**Input Data Accepted:** Cancer type names, gene names, sample IDs.

**Output Data Provided:** Molecular profiles, clinical data, processed analysis results.

**Strengths:** Comprehensive multi-omics data for 33 cancer types; Large sample sizes enabling statistical analyses; Freely accessible through GDC; Standardized data processing; Extensive clinical data

**Limitations:** Primarily bulk tumor data (not single-cell); Tumor purity varies across samples; Some cancer types have limited sample sizes; Data access requires GDC account for controlled-access data; Raw data requires significant computational resources

**Common beginner mistakes:** Not distinguishing between open-access and controlled-access data; Not accounting for tumor purity when analyzing mutation data; Not using the GDC portal for data access (using older TCGA portals) Not considering batch effects when integrating TCGA data with other datasets.

**When to Use It:** Use TCGA for comprehensive cancer genomics analyses, particularly for identifying driver genes, classifying tumors, and discovering biomarkers. TCGA is the primary reference for cancer genomics research.

**When NOT Use It:** Do not use TCGA for single-cell analyses; use single-cell cancer datasets instead. For clinical interpretation of specific mutations, use OncoKB or COSMIC.

**Related databases / alternatives:** ICGC: International cancer genomics data; cBioPortal: Interactive cancer genomics analysis; COSMIC: Somatic mutation catalog; GEO: Additional cancer genomics datasets

**How It Connects to Other Resources:** TCGA data is accessible through the GDC portal. Data is cross-referenced to cBioPortal, COSMIC, and clinical databases. Gene annotations use Ensembl and NCBI Gene.

**API / FTP / programmatic access:** GDC API at <https://api.gdc.cancer.gov/>; returns JSON. Python package gdc-client available. Data is also accessible through cBioPortal API.

**Evidence/curation level:** Experimentally generated with standardized protocols; quality-controlled.

**Data Update Status:** Data collection complete for original TCGA; new data through GDC from other projects

**Licensing / access restrictions:** Open-access data freely available; controlled-access data requires dbGaP authorization.

**Citation / Recommended Reference:** Cancer Genome Atlas Research Network et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120. <https://doi.org/10.1038/ng.2764>

**Beginner-Friendly Explanation:** TCGA (The Cancer Genome Atlas) is a landmark cancer research project that has comprehensively characterized the molecular changes in over 11,000 tumors from 33 different types of cancer. For each tumor, TCGA measured multiple types of molecular data—which genes are mutated, how many copies of each gene are present, which genes are turned on or off, and how DNA is chemically modified. All this data is freely available through the GDC (Genomic Data Commons) portal, making TCGA one of the most valuable resources in cancer research. Researchers use TCGA data to understand what drives different types of cancer and to identify potential targets for new treatments.

**Advanced Technical Explanation:** TCGA implemented standardized protocols for sample collection, nucleic acid extraction, and molecular profiling across all participating institutions. Somatic mutations were called using multiple algorithms (MuTect, VarScan, SomaticSniper) with consensus calling. Copy number alterations were detected using SNP arrays (Affymetrix SNP 6.0) and whole-exome sequencing. Gene expression was measured by RNA-seq (Illumina HiSeq) with RSEM quantification. DNA methylation was measured by Illumina 450K arrays. The TCGA Pan-Cancer Atlas integrated data across all 33 types of cancer to identify pan-cancer patterns of mutation, expression, and clinical outcomes.

#### **One practical workflow example:**

Step 1: Navigate to <https://portal.gdc.cancer.gov> and search for your cancer type of interest.

Step 2: Filter for open-access data (somatic mutations, gene expression) and download.

Step 3: Use the GDC API to retrieve data programmatically: <https://api.gdc.cancer.gov/files?filters=...>

Step 4: Load the mutation data (MAF format) in R using maftools for analysis.

Step 5: Identify frequently mutated genes and perform oncoprint visualization.

Step 6: Cross-reference with cBioPortal for interactive analysis and additional datasets.

## W2 – ICGC (International Cancer Genome Consortium)

**Official Website URL:** <https://icgc.org> / <https://dcc.icgc.org>

**Resource Type:** Database / Repository

**Main Biological Domain:** Clinical genomics / Variants

**What It Is Used For:** ICGC is used to access comprehensive cancer genomics data from international cancer research projects, complementing TCGA with data from non-US cancer cohorts and additional cancer types. The ICGC Data Coordination Centre (DCC) provides access to somatic mutation, copy number, and expression data from over 25,000 cancer genomes across 50+ cancer types. ICGC ARGO (Accelerating Research in Genomic Oncology) is the next-generation ICGC project.

**What Data It Contains:** ICGC contains genomic data from over 25,000 cancer genomes across 50+ cancer types from international cohorts, including whole-genome sequencing, RNA-seq, and clinical data. Data is accessible through the ICGC DCC portal and the ICGC ARGO platform.

**Main question it helps answer:** What somatic mutations and genomic alterations are present in this cancer type from international cohorts?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What are the mutational signatures in this cancer type from international cohorts?
- How do mutation frequencies differ between TCGA and ICGC cohorts?
- What cancer types are covered by ICGC but not TCGA?

**Example use cases:**

- Validating TCGA findings in independent international cohorts
- Accessing cancer types not covered by TCGA
- Studying mutational signatures across cancer types

**Input Data Accepted:** Cancer type names, gene names, project IDs.

**Output Data Provided:** Somatic mutation data, copy number data, expression data, clinical data.

**Strengths:** International cohorts complementing TCGA; Covers additional cancer types; Whole-genome sequencing data available; Freely accessible (open-access data)

**Limitations:** Data access requires registration for controlled-access data; Data formats may differ from TCGA; Some projects have limited sample sizes; Website and portal may be updated/migrated

**Common beginner mistakes:** Not registering for controlled-access data; Not checking data format compatibility with TCGA data; Not using the ICGC DCC portal for data access.

**When to Use It:** Use ICGC to complement TCGA data with international cohorts, to access cancer types not covered by TCGA, or to validate findings in independent cohorts.

**When NOT Use It:** Do not use ICGC as a substitute for TCGA; they are complementary resources.

**Related databases / alternatives:**

- TCGA: US cancer genomics data
- cBioPortal: Interactive cancer genomics analysis
- COSMIC: Somatic mutation catalog

**How It Connects to Other Resources:** ICGC data is cross-referenced to TCGA, cBioPortal, and COSMIC. Data is accessible through the ICGC DCC portal and ICGC ARGO.

**API / FTP / programmatic access:** ICGC DCC API at <https://dcc.icgc.org/api/v1/>; returns JSON. Data downloads are available from the DCC portal.

**Evidence/curation level:** Experimentally generated with standardized protocols; quality-controlled.

**Data Update Status:** ICGC ARGO is the active next-generation project; original ICGC data collection complete.

**Licensing / access restrictions:** Open-access data freely available; controlled-access data requires DACO authorization.

**Citation / Recommended Reference:** Zhang J et al. (2019). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database, 2019:baz026. <https://doi.org/10.1093/database/baz026>

**Beginner-Friendly Explanation:** ICGC (International Cancer Genome Consortium) is an international research project similar to TCGA, but with a focus on cancer cohorts from countries around the world. While TCGA primarily used samples from US patients, ICGC includes samples from Europe, Asia, Australia, and other regions, providing a more globally representative view of cancer genomics. ICGC covers over 50 cancer types and has generated genomic data from over 25,000 tumors. The data is accessible through the ICGC Data Coordination Centre portal, and some data requires registration to access.

**Advanced Technical Explanation:** ICGC implemented standardized data submission and quality control procedures across participating projects from over 20 countries. The ICGC DCC provides a unified data portal with standardized data formats (VCF for mutations, BED for copy number, TSV for expression) and a REST API for programmatic access. The ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) project analyzed whole-genome sequencing data from 2,658 cancer genomes across 38 cancer types, providing the most comprehensive analysis of cancer whole-genome data to date.

**One practical workflow example:**

- Step 1: Navigate to <https://dcc.icgc.org> and search for your cancer type of interest.
- Step 2: Browse available projects and check sample sizes and data types.
- Step 3: Register for a DACO account if controlled-access data is needed.
- Step 4: Download open-access mutation data (SSM: Simple Somatic Mutations) in TSV format.
- Step 5: Use the ICGC DCC API for programmatic access.
- Step 6: Cross-reference with TCGA data for validation.

## W3 – cBioPortal for Cancer Genomics

**Official Website URL:** <https://www.cbioportal.org>

**Resource Type:** Portal / Tool / Database

**Main Biological Domain:** Clinical genomics / Variants

**What It Is Used For:** cBioPortal is used to interactively explore, visualize, and analyze cancer genomics data from TCGA, ICGC, and many other studies, providing a user-friendly interface for cancer genomics analysis without requiring bioinformatics expertise. It is used for querying mutations, copy number alterations, and expression data across cancer types, for survival analysis, and for network analysis of cancer genes. cBioPortal is one of the most widely used tools in cancer genomics.

**What Data It Contains:** cBioPortal contains data from over 300 cancer genomics studies with over 100,000 samples, including TCGA, ICGC, and many published studies. Data types include somatic mutations, copy number alterations, gene expression, DNA methylation, protein expression, and clinical data.

**Main question it helps answer:** What is the frequency and clinical significance of alterations in my gene(s) of interest across cancer types?

**Typical user:** Researcher / Clinician / Bioinformatician / Data analyst

**Example scientific questions:**

- What is the mutation frequency of KRAS across different cancer types?
- What is the survival impact of TP53 mutations in lung cancer?
- What genes are co-altered with BRCA1 in breast cancer?

**Example use cases:**

- Querying mutation frequencies across cancer types
- Survival analysis based on molecular alterations
- Identifying co-occurring and mutually exclusive alterations

**Input Data Accepted:** Gene names, cancer type names, study IDs.

**Output Data Provided:** Mutation frequencies, OncoPrint visualizations, survival curves, network analyses.

**Strengths:** User-friendly interface for non-bioinformaticians; Integrates data from hundreds of studies; Excellent visualization tools; Freely accessible; Comprehensive API

**Limitations:** Data quality varies across studies; Some analyses require bioinformatics expertise; Not suitable for raw data analysis; Some features require registration

**Common beginner mistakes:** Not selecting the appropriate study for their cancer type; Not distinguishing between different alteration types (mutation vs. amplification); Not using the OncoPrint for visualizing alterations across samples; Not checking the sample size of each study.

**When to Use It:** Use cBioPortal for interactive exploration of cancer genomics data, for querying mutation frequencies, and for survival analysis. Ideal for researchers without bioinformatics expertise.

**When NOT to Use It:** Do not use cBioPortal for raw data analysis; download data from GDC or ICGC DCC for custom analyses.

**Related databases / alternatives:** TCGA: Raw cancer genomics data; COSMIC: Somatic mutation catalog; OncoKB: Clinical mutation annotation; UCSC Xena: Alternative cancer genomics browser



**How It Connects to Other Resources:** cBioPortal integrates TCGA, ICGC, and published study data. Mutations are annotated with OncoKB and COSMIC. Gene information is linked to UniProt and Ensembl.

**API / FTP / programmatic access:** REST API at <https://www.cbioportal.org/api/>; returns JSON. Python package cbio-py available. R package cgdscr available.

**Evidence/curation level:** Curated from published studies; quality varies by study.

**Data Update Status:** Regularly updated with new studies; actively maintained as of 2024.

**Licensing / access restrictions:** Freely available; some datasets may have access restrictions.

**Citation / Recommended Reference:** Cerami E et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, 2(5):401–404. <https://doi.org/10.1158/2159-8290.CD-12-0095>

**Beginner-Friendly Explanation:** cBioPortal is a website that makes it easy to explore cancer genomics data without needing to be a bioinformatics expert. It provides access to data from hundreds of cancer studies, including TCGA, and lets you ask questions like "How often is this gene mutated in breast cancer?" or "Do patients with this mutation have better or worse survival?" The website provides interactive visualizations including OncoPrint (a visual summary of which genes are altered in which patients are altered) and survival curves. cBioPortal is one of the most widely used tools in cancer research because it makes complex genomics data accessible to researchers with diverse backgrounds.

**Advanced Technical Explanation:** cBioPortal implements a comprehensive data model that captures somatic mutations (with functional impact annotations from OncoKB and COSMIC), copy number alterations (deep deletion, shallow deletion, diploid, gain, amplification), mRNA expression (z-scores relative to diploid samples), protein expression (RPPA), and DNA methylation. The portal provides tools for OncoPrint visualization, mutual exclusivity/co-occurrence analysis (Fisher's exact test), survival analysis (Kaplan-Meier with log-rank test), network analysis (using STRING and Reactome), and comparison of molecular profiles across groups. The cBioPortal API provides programmatic access to all data and analysis results.

**One practical workflow example:**

Step 1: Navigate to <https://www.cbioportal.org> and select your cancer type and study.

Step 2: Enter your gene(s) of interest in the query box.

Step 3: Review the OncoPrint to see the frequency and type of alterations.

Step 4: Click "Survival" to see the impact of alterations on patient survival.

Step 5: Use "Comparison" to compare molecular profiles between altered and unaltered groups.

Step 6: Download the data for custom analysis using the cBioPortal API.



## W4 – COSMIC (Catalogue of Somatic Mutations in Cancer)

**Official Website URL:** <https://cancer.sanger.ac.uk/cosmic>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** Clinical genomics / Variants

**What It Is Used For:** COSMIC is used to access a comprehensive catalog of somatic mutations in human cancer, providing curated mutation data from primary literature and large-scale sequencing projects. It is used for identifying cancer driver genes, understanding mutational signatures, and annotating cancer mutations with functional significance. COSMIC is the primary reference for somatic mutation data in cancer research.

**What Data It Contains:** COSMIC contains over 17 million coding mutations from over 1.5 million tumor samples, curated from primary literature and large-scale sequencing projects. Data includes point mutations, insertions/deletions, copy number alterations, gene fusions, and mutational signatures. The Cancer Gene Census (CGC) provides a curated list of cancer driver genes.

**Main question it helps answer:** What somatic mutations have been reported in cancer, and which genes are known cancer drivers?

**Typical user:** Researcher / Clinician / Bioinformatician

**Example scientific questions:**

- What mutations in KRAS have been reported in cancer?
- What is the COSMIC mutational signature of this tumor?
- Is this mutation in the Cancer Gene Census?

**Example use cases:** Annotating somatic mutations with COSMIC IDs and functional significance; Identifying cancer driver genes using the Cancer Gene Census; Analyzing mutational signatures in tumor genomes.

**Input Data Accepted:** Gene names, mutation coordinates, COSMIC IDs

**Output Data Provided:** Mutation records, Cancer Gene Census, mutational signatures.

**Strengths:** Most comprehensive somatic mutation catalog; Cancer Gene Census for driver gene identification; Mutational signatures database; Freely accessible (with registration); Regularly updated

**Limitations:** Registration required for full data access; Some features require commercial license; Data quality varies across studies; Coverage biased toward well-studied cancer types

**Common beginner mistakes:** Not registering for full data access; Confusing COSMIC IDs with other mutation identifiers; Not using the Cancer Gene Census for driver gene identification; Not considering the evidence level for each mutation.

**When to Use It:** Use COSMIC for annotating somatic mutations, identifying cancer driver genes, and analyzing mutational signatures. COSMIC is the primary reference for somatic mutation data.

**When NOT to Use It:** Do not use COSMIC for germline variant annotation; use ClinVar or gnomAD instead. COSMIC focuses on somatic mutations.

**Related databases / alternatives:** cBioPortal: Interactive cancer genomics analysis; OncoKB: Clinical mutation annotation; TCGA: Primary cancer genomics data; ClinVar: Germline variant annotation

**How It Connects to Other Resources:** COSMIC mutations are cross-referenced to Ensembl, UniProt, and PubMed. COSMIC IDs are used by cBioPortal and other cancer genomics tools.

**API / FTP / programmatic access:** REST API at <https://cancer.sanger.ac.uk/cosmic/download>; returns JSON or TSV. Registration required. Python package cosmic-py available.

**Evidence/curation level:** Manually curated from primary literature and large-scale sequencing; high quality.

**Data Update Status:** Regular releases; actively maintained as of 2024.

**Licensing / access restrictions:** Free registration required; some features require commercial license.

**Citation / Recommended Reference:** Tate JG et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research, 47(D1):D941–D947. <https://doi.org/10.1093/nar/gky1015>

**Beginner-Friendly Explanation:** COSMIC (Catalogue of Somatic Mutations in Cancer) is a comprehensive database of mutations that have been found in cancer cells. Unlike inherited mutations that are present in all cells of the body, somatic mutations are acquired during a person's lifetime and are found only in cancer cells. COSMIC collects information about these mutations from thousands of scientific studies and large-scale cancer sequencing projects, creating the most comprehensive catalog of cancer mutations available. COSMIC also maintains the Cancer Gene Census—a curated list of genes that are known to drive cancer development—and a database of mutational signatures that describe the patterns of mutations caused by different processes (like UV radiation or tobacco smoke).

**Advanced Technical Explanation:** COSMIC implements a comprehensive somatic mutation data model that captures mutation type (substitution, insertion, deletion, complex), genomic coordinates (GRCh38), protein change (HGVS notation), functional impact (FATHMM score), and evidence level (primary literature, large-scale studies). The Cancer Gene Census (CGC) classifies cancer genes as oncogenes, tumor suppressors, or fusion genes, with evidence levels (Tier 1: well-established, Tier 2: emerging evidence). COSMIC mutational signatures (v3.3) define 79 single base substitution (SBS), 11 doublet base substitution (DBS), and 17 insertion/deletion (ID) signatures, each associated with specific mutational processes.

**One practical workflow example:**

Step 1: Navigate to <https://cancer.sanger.ac.uk/cosmic> and register for a free account.

Step 2: Search for your gene of interest to see all reported mutations.

Step 3: Filter by cancer type and mutation type.

Step 4: Check the Cancer Gene Census to see if your gene is a known driver.

Step 5: Download the mutation data for your gene in TSV format.

Step 6: Use the COSMIC mutational signatures tool to analyze the signature of your tumor.

## W5 – OncoKB

**Official Website URL:** <https://www.oncokb.org>

**Resource Type:** Knowledgebase

**Main Biological Domain:** Clinical genomics / Variants / Drugs

**What It Is Used For:** OncoKB is used to access a precision oncology knowledge base that annotates cancer mutations with clinical significance and therapeutic implications. NOTE: OncoKB operates on a tiered access model; academic use is free with registration, but commercial use requires a license. It is used for clinical interpretation of cancer mutations, identifying actionable mutations, and understanding the therapeutic implications of specific alterations.

**What Data It Contains:** OncoKB contains clinical annotations for over 6,000 alterations in over 600 genes, with evidence levels for oncogenicity and therapeutic implications. Each alteration is classified by its oncogenic effect (oncogenic, likely oncogenic, neutral, etc.) and its therapeutic implications (FDA-approved therapies, clinical trials, etc.).

**Main question it helps answer:** Is this cancer mutation clinically actionable, and what therapies are available for patients with this mutation?

**Typical user:** Clinician / Researcher / Bioinformatician

**Example scientific questions:**

- Is this EGFR mutation actionable with approved therapies?
- What is the evidence level for this BRAF V600E mutation?
- What clinical trials are available for patients with this mutation?

**Example use cases:** Clinical interpretation of cancer mutations from tumor sequencing; Identifying actionable mutations in cancer genomics studies; Understanding the therapeutic landscape for specific cancer mutations.

**Input Data Accepted:** Gene names, mutation names (HGVS notation), cancer types.

**Output Data Provided:** Oncogenicity classifications, therapeutic implications, evidence levels.

**Strengths:** Expert-curated clinical annotations; Evidence-based classification system; Therapeutic implications with FDA approval status; Freely accessible for academic use; Regularly updated

**Limitations:** Tiered access model; commercial use requires license; Coverage limited to clinically relevant mutations; Not comprehensive for all cancer mutations; Annotations may lag behind primary literature

**Common beginner mistakes:** Not registering for academic access; Confusing OncoKB evidence levels with COSMIC or ClinVar classifications; Using OncoKB for non-cancer variant annotation; Not checking the evidence level for each annotation.

**When to Use It:** Use OncoKB for clinical interpretation of cancer mutations, particularly for identifying actionable mutations and understanding therapeutic implications. Essential for precision oncology.

**When NOT to Use It:** Do not use OncoKB for comprehensive somatic mutation data; use COSMIC instead. OncoKB focuses on clinically relevant mutations.

**Related databases / alternatives:**

- COSMIC: Comprehensive somatic mutation catalog



- ClinVar: Germline variant annotation
- cBioPortal: Interactive cancer genomics analysis
- JAX CKB: Alternative precision oncology knowledge base

**How It Connects to Other Resources:** OncoKB annotations are integrated into cBioPortal. Mutations are cross-referenced to COSMIC and ClinVar. Therapies are linked to FDA drug databases.

**API / FTP / programmatic access:** REST API at <https://www.oncokb.org/api/v1/>; returns JSON. Registration required. Python package oncokb-annotator available.

**Evidence/curation level:** Expert-curated; evidence-based classification; high quality

**Data Update Status:** Regular updates; actively maintained as of 2024.

**Licensing / access restrictions:** Free for academic use with registration; commercial license required.

**Citation / Recommended Reference:** Chakravarty D et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology, 1:1–16. <https://doi.org/10.1200/PO.17.00011>

**Beginner-Friendly Explanation:** OncoKB is a knowledge base that helps doctors and researchers understand the clinical significance of mutations found in cancer patients. When a patient's tumor is sequenced, hundreds of mutations may be found, but only a few are likely to be driving the cancer and potentially treatable with specific drugs. OncoKB provides expert-curated information about which mutations are clinically important, what drugs are approved to treat cancers with those mutations, and what clinical trials might be available. OncoKB is free for academic use with registration, making it an important resource for precision oncology research.

**Advanced Technical Explanation:** OncoKB implements a tiered evidence classification system for both oncogenicity (Oncogenic, Likely Oncogenic, Predicted Oncogenic, Neutral, Likely Neutral, Inconclusive, Unknown) and therapeutic implications (Level 1: FDA-approved biomarker; Level 2: Standard care biomarker; Level 3A: Compelling clinical evidence; Level 3B: Standard care or investigational in other tumor type; Level 4: Compelling biological evidence; Level R1: Standard care resistance; Level R2: Investigational resistance). This classification system is aligned with AMP/ASCO/CAP guidelines for somatic variant interpretation. OncoKB's API enables integration into clinical genomics pipelines for automated mutation annotation.

**One practical workflow example:**

Step 1: Navigate to <https://www.oncokb.org> and register for a free academic account.

Step 2: Search for your mutation of interest (e.g., EGFR L858R).

Step 3: Review the oncogenicity classification and therapeutic implications.

Step 4: Check the evidence level and support literature.

Step 5: Use the OncoKB API to annotate a list of mutations programmatically.

Step 6: Integrate OncoKB annotations into your clinical genomics pipeline.

## BEGINNER EXAMPLE (Category W):

---

A cancer biology student wants to understand the mutation landscape of colorectal cancer. They navigate to cBioPortal, select the TCGA colorectal adenocarcinoma study, and query the top 10 most frequently mutated genes. They find that APC (82%), TP53 (60%), KRAS (43%), and PIK3CA (18%) are the most commonly mutated. They then check OncoKB to see which of these mutations are clinically actionable and find that KRAS mutations are associated with resistance to anti-EGFR therapy.

## ADVANCED EXAMPLE (Category W):

---

A computational oncologist is analyzing whole-genome sequencing data from a cohort of pancreatic cancer patients. They use COSMIC mutational signatures to identify the dominant mutational processes in each tumor. They find that a subset of tumors shows a signature consistent with BRCA1/2 deficiency (Signature 3). They cross-reference with TCGA data to validate their findings and use OncoKB to identify PARP inhibitors as potential therapeutic options for patients with this signature.

## CONFUSION POINTS (Category W):

---

- TCGA and ICGC are data repositories; cBioPortal is an analysis portal that uses their data.
- COSMIC IDs (e.g., COSM476) are different from dbSNP IDs (rs numbers); COSMIC is for somatic mutations.
- OncoKB evidence levels are different from ClinVar pathogenicity classifications.
- Tumor purity affects mutation calling; low-purity tumors may have missed mutations.
- Somatic mutations (in COSMIC, TCGA) are different from germline variants (in ClinVar, gnomAD).

## DECISION GUIDE (Category W):

---

Need raw multi-omics cancer data? → TCGA (via GDC)

Need international cancer cohort data? → ICGC

Need interactive cancer genomics analysis? → cBioPortal

Need comprehensive somatic mutation catalog? → COSMIC

Need clinical interpretation of cancer mutations? → OncoKB

Need to identify cancer driver genes? → COSMIC Cancer Gene Census

## Category X: Model Organism Databases

### OVERVIEW

Model organism databases are specialized resources that integrate genomic, genetic, phenotypic, and functional data for specific organisms that are widely used in biological research. These organisms—including the fruit fly *Drosophila melanogaster*, the nematode *Caenorhabditis elegans*, the mouse *Mus musculus*, the zebrafish *Danio rerio*, the budding yeast *Saccharomyces cerevisiae*, the plant *Arabidopsis thaliana*, and the fission yeast *Schizosaccharomyces pombe*—have been chosen as model systems because of their experimental tractability, short generation times, and the availability of powerful genetic tools. Model organism databases provide curated, organism-specific information that goes far beyond what is available in general databases.

Each model organism database serves as the authoritative reference for its organism, providing gene annotations, mutant phenotypes, genetic interactions, expression data, and community resources. These databases are maintained by dedicated teams of curators who manually review the primary literature and integrate data from large-scale genomic and genetic studies. The depth of curation in model organism databases is unmatched by general databases—for example, WormBase contains phenotype data for thousands of *C. elegans* genes, and SGD provides comprehensive genetic interaction data for yeast. This depth of information makes model organism databases invaluable for understanding gene function and for translating findings from model organisms to human biology.

A key challenge in using model organism databases is the translation of findings from model organisms to human biology. While model organisms share many fundamental biological processes with humans, there are also important differences in gene function, regulation, and physiology. Ortholog databases (like OrthoFinder, InParanoid, and the DIOPT tool at FlyBase) help researchers identify human orthologs of model organism genes, enabling the translation of findings across species. Researchers must be aware of the limitations of ortholog prediction and the importance of experimental validation when translating findings from model organisms to human biology.

## X1 – FlyBase

**Official Website URL:** <https://flybase.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** FlyBase is used to access comprehensive genomic, genetic, and functional data for *Drosophila melanogaster* (fruit fly) and related species, providing the authoritative reference for *Drosophila* biology. It is used for gene function analysis, genetic screen data, phenotype information, and as a resource for translating *Drosophila* findings to human biology. FlyBase integrates data from the primary literature and large-scale genomic studies.

**What Data It Contains:** FlyBase contains genome annotations for *D. melanogaster* and 12 related species, with gene records including sequence, expression, mutant phenotypes, genetic interactions, protein interactions, GO annotations, and literature references. The database also contains stock information for *Drosophila* genetic resources.

**Main question it helps answer:** What is known about the function and phenotype of this *Drosophila* gene, and what is its human ortholog?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What phenotypes are associated with loss of function of this *Drosophila* gene?
- What is the human ortholog of this *Drosophila* gene?
- What genetic interactions have been reported for this gene?

**Example use cases:**

- Finding the human ortholog of a *Drosophila* gene identified in a genetic screen
- Accessing phenotype data for *Drosophila* mutants
- Identifying genetic interactions for network analysis

**Input Data Accepted:** Gene names, FlyBase IDs, CG numbers

**Output Data Provided:** Gene records, phenotype data, genetic interactions, expression data

**Strengths:** Comprehensive *Drosophila* gene information; Excellent ortholog prediction tools (DIOPT); Phenotype data from thousands of mutants; Freely accessible; Regularly updated

**Limitations:** Focused on *Drosophila*; limited for other organisms; Some gene functions may not be conserved in humans; Curation lag for very recent publications

**Common beginner mistakes:** Not using DIOPT for human ortholog identification; Not checking the evidence level for phenotype annotations; Confusing FlyBase gene IDs (FBgn) with CG numbers

**When to Use It:** Use FlyBase for *Drosophila* gene function analysis, phenotype data, and human ortholog identification. Essential for *Drosophila* researchers.

**When NOT Use It:** Do not use FlyBase for non-*Drosophila* organisms; use the appropriate model organism database.





**Related databases / alternatives:** WormBase: *C. elegans* database; MGI: Mouse database; ZFIN: Zebrafish database; DIOPT: Ortholog prediction tool

**How It Connects to Other Resources:** FlyBase cross-references UniProt, Ensembl, NCBI Gene, and GO. Human orthologs are identified using DIOPT.

**API / FTP / programmatic access:** REST API at <https://api.flybase.org/>; returns JSON. FTP downloads at <https://ftp.flybase.net/>. Python package flybase-py available.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Larkin A et al. (2021). FlyBase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Research*, 49(D1):D899–D907. <https://doi.org/10.1093/nar/gkaa1026>

**Beginner-Friendly Explanation:** FlyBase is the primary database for the fruit fly *Drosophila melanogaster*, one of the most important model organisms in biology. For each *Drosophila* gene, FlyBase provides information about its DNA sequence, when and where it is expressed, what happens when it is mutated, and how it interacts with other genes. FlyBase is particularly valuable because many *Drosophila* genes have human counterparts (orthologs), and studying the fly gene can provide insights into human biology and disease. FlyBase also provides tools for finding the human ortholog of any *Drosophila* gene, making it a bridge between fly biology and human medicine.

**Advanced Technical Explanation:** FlyBase implements a comprehensive gene model that captures gene structure (with alternative transcripts and isoforms), expression data (from RNA-seq, in situ hybridization, and reporter assays), mutant phenotypes (with controlled vocabulary from the *Drosophila* Phenotype Ontology), genetic interactions (enhancer/suppressor relationships), protein interactions (from BioGRID and IntAct), and GO annotations. The DIOPT (DRSC Integrative Ortholog Prediction Tool) integrates 13 ortholog prediction algorithms to provide consensus ortholog predictions between *Drosophila* and human genes. FlyBase uses the Chado database schema for data storage.

#### One practical workflow example:

Step 1: Navigate to <https://flybase.org> and search for your *Drosophila* gene of interest.

Step 2: Review the gene summary, including expression, phenotypes, and interactions.

Step 3: Use the DIOPT tool to find human orthologs: <https://www.flyrnai.org/diopt>.

Step 4: Check the phenotype data for loss-of-function and gain-of-function alleles.

Step 5: Download the gene data for computational analysis.

Step 6: Cross-reference with MGI or WormBase for ortholog function in other model organisms.

## X2 – WormBase

---

**Official Website URL:** <https://wormbase.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** WormBase is used to access comprehensive genomic, genetic, and functional data for *Caenorhabditis elegans* (nematode) and related nematode species, providing the authoritative reference for *C. elegans* biology. It is used for gene function analysis, RNAi phenotype data, cell lineage information, and as a resource for translating *C. elegans* findings to human biology.

**What Data It Contains:** WormBase contains genome annotations for *C. elegans* and related nematodes, with gene records including sequence, expression, RNAi phenotypes, genetic interactions, protein interactions, GO annotations, and cell lineage data. The complete cell lineage of *C. elegans* (959 somatic cells) is a unique resource.

**Main question it helps answer:** What is known about the function and phenotype of this *C. elegans* gene, and what is its human ortholog?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What phenotype results from RNAi knockdown of this *C. elegans* gene?
- What is the expression pattern of this gene during *C. elegans* development?
- What is the human ortholog of this *C. elegans* gene?

**Example use cases:**

- Accessing RNAi phenotype data for *C. elegans* genes
- Studying gene expression during *C. elegans* development
- Identifying human orthologs of *C. elegans* genes

**Input Data Accepted:** Gene names, WormBase IDs, sequence names

**Output Data Provided:** Gene records, phenotype data, expression data, cell lineage information

**Strengths:** Comprehensive *C. elegans* gene information; Complete cell lineage data; Extensive RNAi phenotype data; Freely accessible; Regularly updated

**Limitations:** Focused on *C. elegans*; limited for other organisms; Some gene functions may not be conserved in humans; Curation lag for very recent publications

**Common beginner mistakes:** Not using WormBase IDs (WBGene) for programmatic access; Not checking the evidence level for phenotype annotations; Not using the cell lineage viewer for developmental studies

**When to Use It:** Use WormBase for *C. elegans* gene function analysis, RNAi phenotype data, and human ortholog identification.

**When NOT to Use It:** Do not use WormBase for non-nematode organisms.

**Related databases / alternatives:** FlyBase: *Drosophila* database; MGI: Mouse database; ZFIN: Zebrafish database



**How It Connects to Other Resources:** WormBase cross-references UniProt, Ensembl, NCBI Gene, and GO. Human orthologs are identified using OrthoFinder and other tools.

**API / FTP / programmatic access:** REST API at <https://wormbase.org/rest/>; returns JSON. FTP downloads at <https://ftp.wormbase.org/>. Python package wormbase-py available.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Harris TW et al. (2020). WormBase: a modern Model Organism Information Resource. Nucleic Acids Research, 48(D1):D762–D767. <https://doi.org/10.1093/nar/gkz920>

**Beginner-Friendly Explanation:** WormBase is the primary database for the nematode worm *Caenorhabditis elegans*, a tiny transparent worm that is one of the most important model organisms in biology. *C. elegans* was the first multicellular organism to have its genome completely sequenced, and its complete cell lineage (the developmental history of every cell in the worm) has been mapped. WormBase provides comprehensive information about every *C. elegans* gene, including what happens when the gene is turned off using RNAi (RNA interference), where the gene is expressed during development, and what human genes it is related to. This makes WormBase an invaluable resource for studying fundamental biological processes.

**Advanced Technical Explanation:** WormBase implements a comprehensive data model that captures gene structure, expression (from RNA-seq, reporter assays, and in situ hybridization), RNAi phenotypes (from genome-wide RNAi screens), genetic interactions, protein interactions, GO annotations, and cell lineage data. The complete *C. elegans* cell lineage (959 somatic cells in hermaphrodites) is a unique resource that enables cell-specific gene expression analysis. WormBase uses the Chado database schema and provides a REST API for programmatic data access.

**One practical workflow example:**

Step 1: Navigate to <https://wormbase.org> and search for your *C. elegans* gene.

Step 2: Review the gene summary, including expression, phenotypes, and interactions.

Step 3: Check the RNAi phenotype data from genome-wide screens.

Step 4: Use the cell lineage viewer to see expression in specific cells.

Step 5: Find human orthologs using the ortholog section of the gene page.

Step 6: Download gene data for computational analysis.

## X3 – MGI (Mouse Genome Informatics)

**Official Website URL:** <https://www.informatics.jax.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics / Diseases

**What It Is Used For:** MGI is used to access comprehensive genomic, genetic, and phenotypic data for the laboratory mouse (*Mus musculus*), providing the authoritative reference for mouse genetics and genomics. It is used for gene function analysis, mouse mutant phenotypes, disease models, and as a resource for translating mouse findings to human biology. MGI integrates data from the Mouse Genome Database (MGD) and the Gene Expression Database (GXD).

**What Data It Contains:** MGI contains genome annotations for mouse, with gene records including sequence, expression (from GXD), mutant phenotypes (from MGD), allele information, disease models (using HPO and DO), GO annotations, and literature references. The database also contains information on mouse strains and genetic resources.

**Main question it helps answer:** What is known about the function and phenotype of this mouse gene, and what human diseases does it model?

**Typical user:** Researcher / Bioinformatician / Clinician

**Example scientific questions:**

- What phenotypes are associated with knockout of this mouse gene?
- What mouse models are available for this human disease?
- What is the expression pattern of this gene during mouse development?

**Example use cases:**

- Finding mouse models for human diseases
- Accessing phenotype data for mouse mutants
- Studying gene expression during mouse development

**Input Data Accepted:** Gene names, MGI IDs, human gene names

**Output Data Provided:** Gene records, phenotype data, disease models, expression data

**Strengths:** Comprehensive mouse gene information; Disease model annotations; Extensive phenotype data; Freely accessible; Regularly updated

**Limitations:** Focused on mouse; limited for other organisms; Some mouse phenotypes may not translate to humans; Curation lag for very recent publications

**Common beginner mistakes:** Not using MGI IDs for programmatic access; Not checking the evidence level for phenotype annotations; Not using the disease model annotations for translational research

**When to Use It:** Use MGI for mouse gene function analysis, phenotype data, and disease model identification. Essential for mouse researchers and translational biology.

**When NOT Use It:** Do not use MGI for non-mouse organisms.

**Related databases / alternatives:** FlyBase: *Drosophila* database; WormBase: *C. elegans* database; ZFIN: Zebrafish database; IMPC: International Mouse Phenotyping Consortium

**How It Connects to Other Resources:** MGI cross-references UniProt, Ensembl, NCBI Gene, GO, HPO, and DO. Human orthologs are identified using HomoloGene and other tools.

**API / FTP / programmatic access:** REST API at <https://www.informatics.jax.org/mgihome/other/mgijava.shtml>; FTP downloads at <https://www.informatics.jax.org/downloads/reports/>. Python package mgi-py available.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Bult CJ et al. (2019). Mouse Genome Database (MGD) 2019. Nucleic Acids Research, 47(D1):D801–D806. <https://doi.org/10.1093/nar/gky1056>

**Beginner-Friendly Explanation:** MGI (Mouse Genome Informatics) is the primary database for the laboratory mouse, which is the most widely used mammalian model organism in biomedical research. For each mouse gene, MGI provides information about its DNA sequence, where it is expressed during development, what happens when it is mutated, and what human diseases the mouse mutant models. This last feature is particularly valuable for translational research—if you are studying a human disease, MGI can help you find mouse models that mimic the disease, which can be used to test potential treatments. MGI is maintained by the Jackson Laboratory, one of the world's leading mouse genetics research centers.

**Advanced Technical Explanation:** MGI implements a comprehensive data model that captures gene structure, expression (from GXD, with data from RNA-seq, in situ hybridization, immunohistochemistry, and reporter assays), mutant phenotypes (using the Mammalian Phenotype Ontology, MP), allele information (knockout, knockin, conditional, transgenic), disease models (using HPO and DO for human disease associations), and GO annotations. MGI participates in the International Mouse Phenotyping Consortium (IMPC), which is systematically phenotyping knockout mice for all protein-coding genes. The MGI REST API provides programmatic access to all data types.

**One practical workflow example:**

- Step 1: Navigate to <https://www.informatics.jax.org> and search for your gene of interest.
- Step 2: Review the gene summary, including expression, phenotypes, and disease models.
- Step 3: Check the allele section for available mouse mutants.
- Step 4: Use disease model annotations to find mouse models for your disease of interest.
- Step 5: Download gene data for computational analysis.
- Step 6: Cross-reference with IMPC for systematic phenotyping data.

## X4 – ZFIN (Zebrafish Information Network)

**Official Website URL:** <https://zfin.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** ZFIN is used to access comprehensive genomic, genetic, and phenotypic data for the zebrafish (*Danio rerio*), providing the authoritative reference for zebrafish biology. It is used for gene function analysis, zebrafish mutant phenotypes, expression data, and as a resource for translating zebrafish findings to human biology. Zebrafish are particularly valuable for developmental biology and drug screening.

**What Data It Contains:** ZFIN contains genome annotations for zebrafish, with gene records including sequence, expression (from in situ hybridization and RNA-seq), mutant phenotypes, morpholino knockdown data, GO annotations, and literature references. The database also contains information on zebrafish strains and genetic resources.

**Main question it helps answer:** What is known about the function and phenotype of this zebrafish gene, and what is its human ortholog?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What phenotype results from morpholino knockdown of this zebrafish gene?
- What is the expression pattern of this gene during zebrafish development?
- What zebrafish models are available for this human disease?

**Example use cases:**

- Accessing phenotype data for zebrafish mutants and morphants
- Studying gene expression during zebrafish development
- Finding zebrafish models for human diseases

**Input Data Accepted:** Gene names, ZFIN IDs, human gene names

**Output Data Provided:** Gene records, phenotype data, expression data, disease models

**Strengths:** Comprehensive zebrafish gene information; Extensive expression data from in situ hybridization; Disease model annotations; Freely accessible; Regularly updated

**Limitations:** Focused on zebrafish; limited for other organisms; Morpholino data may have off-target effects; Some gene functions may not be conserved in humans

**Common beginner mistakes:** Not distinguishing between morpholino knockdown and genetic mutant data; Not checking the evidence level for phenotype annotations; Not using ZFIN IDs for programmatic access

**When to Use It:** Use ZFIN for zebrafish gene function analysis, phenotype data, and disease model identification.

**When NOT to Use It:** Do not use ZFIN for non-zebrafish organisms.

**Related databases / alternatives:** FlyBase: *Drosophila* database; WormBase: *C. elegans* database ;MGI: Mouse database



**How It Connects to Other Resources:** ZFIN cross-references UniProt, Ensembl, NCBI Gene, GO, HPO, and DO.

**API / FTP / programmatic access:** REST API at <https://zfin.org/action/api/>; FTP downloads at <https://zfin.org/downloads>.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Bradford YM et al. (2022). Zebrafish Information Network, the knowledgebase for Danio rerio research. Genetics, 220(4):iyac016. <https://doi.org/10.1093/genetics/iyac016>

**Beginner-Friendly Explanation:** ZFIN (Zebrafish Information Network) is the primary database for the zebrafish *Danio rerio*, a small tropical fish that has become an important model organism in biomedical research. Zebrafish are particularly useful because their embryos are transparent, making it easy to observe development in real time, and they can be used for large-scale drug screening. ZFIN provides comprehensive information about zebrafish genes, including expression patterns during development, phenotypes when genes are mutated or knocked down, and connections to human diseases. Many human disease genes have zebrafish counterparts, and zebrafish models are used to study diseases ranging from cancer to heart disease to neurological disorders.

**Advanced Technical Explanation:** ZFIN implements a comprehensive data model that captures gene structure, expression (from in situ hybridization, immunohistochemistry, and RNA-seq), mutant phenotypes (using the Zebrafish Phenotype Ontology, ZP), morpholino knockdown data, GO annotations, and disease model annotations. ZFIN participates in the Alliance of Genome Resources, which coordinates data sharing between model organism databases. The ZFIN REST API provides programmatic access to all data types.

**One practical workflow example:**

- Step 1: Navigate to <https://zfin.org> and search for your zebrafish gene.
- Step 2: Review the gene summary, including expression, phenotypes, and disease models.
- Step 3: Check the expression data for hybridization images in situ.
- Step 4: Review mutant and morphant phenotype data.
- Step 5: Find human orthologs using the ortholog section.
- Step 6: Download gene data for computational analysis.



## X5 – SGD (Saccharomyces Genome Database)

**Official Website URL:** <https://www.yeastgenome.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** SGD is used to access comprehensive genomic, genetic, and functional data for the budding yeast *Saccharomyces cerevisiae*, providing the authoritative reference for yeast biology. It is used for gene function analysis, genetic interaction data, protein interaction data, and as a resource for understanding fundamental eukaryotic biology. SGD is particularly valuable for its comprehensive genetic interaction data.

**What Data It Contains:** SGD contains genome annotations for *S. cerevisiae*, with gene records including sequence, expression, mutant phenotypes, genetic interactions (from the Boone/Andrews lab screens), protein interactions, GO annotations, and literature references. The database also contains information on yeast strains and genetic resources.

**Main question it helps answer:** What is known about the function and genetic interactions of this yeast gene?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What genetic interactions have been reported for this yeast gene?
- What is the phenotype of deletion of this yeast gene?
- What is the human ortholog of this yeast gene?

**Example use cases:**

- Accessing genetic interaction data for network analysis
- Studying gene function in yeast
- Finding human orthologs of yeast genes

**Input Data Accepted:** Gene names, SGD IDs, systematic names

**Output Data Provided:** Gene records, genetic interaction data, phenotype data, expression data

**Strengths:** Comprehensive yeast gene information; Most complete genetic interaction dataset for any organism; Freely accessible; Regularly updated; Excellent API

**Limitations:** Focused on *S. cerevisiae*; limited for other organisms; Some gene functions may not be conserved in humans; Curation lag for very recent publications

**Common beginner mistakes:** Not using systematic gene names (e.g., YAL001C) for programmatic access; Not using the genetic interaction data for network analysis; Confusing SGD with PomBase (different yeast species)

**When to Use It:** Use SGD for yeast gene function analysis, genetic interaction data, and human ortholog identification.

**When NOT Use It:** Do not use SGD for *S. pombe* (fission yeast); use PomBase instead.

**Related databases / alternatives:** PomBase: *S. pombe* database; FlyBase: *Drosophila* database; WormBase: *C. elegans* database



**How It Connects to Other Resources:** SGD cross-references UniProt, Ensembl, NCBI Gene, GO, and BioGRID (for genetic interactions).

**API / FTP / programmatic access:** REST API at <https://www.yeastgenome.org/backend/>; FTP downloads at <https://downloads.yeastgenome.org/>. Python package yeastgenome available.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Cherry JM et al. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research, 40(D1):D700–D705. <https://doi.org/10.1093/nar/gkr1029>

**Beginner-Friendly Explanation:** SGD (Saccharomyces Genome Database) is the primary database for the budding yeast *Saccharomyces cerevisiae*, one of the most important model organisms in biology. Yeast is used to study fundamental cellular processes like cell division, DNA repair, and protein folding because many of these processes are conserved between yeast and humans. SGD provides comprehensive information about every yeast gene, including what happens when it is deleted, how it interacts with other genes, and what human genes it is related to. SGD is particularly famous for its genetic interaction data—the most comprehensive dataset of gene-gene interactions for any organism—which has been used to map the functional organization of the yeast cell.

**Advanced Technical Explanation:** SGD implements a comprehensive data model that captures gene structure, expression (from RNA-seq and microarray data), mutant phenotypes (using the Ascomycete Phenotype Ontology, APO), genetic interactions (from the Boone/Andrews lab synthetic genetic array screens, covering ~5.4 million gene pairs), protein interactions (from BioGRID), GO annotations, and pathway information. The genetic interaction data from SGD represents the most comprehensive genetic interaction dataset for any organism, covering ~90% of all possible gene pairs in *S. cerevisiae*. SGD participates in the Alliance of Genome Resources for data sharing with other model organism databases.

**One practical workflow example:**

- Step 1: Navigate to <https://www.yeastgenome.org> and search for your yeast gene.
- Step 2: Review the gene summary, including function, phenotypes, and interactions.
- Step 3: Check the genetic interaction data for synthetic lethal and suppressor interactions.
- Step 4: Use the GO annotations for functional enrichment analysis.
- Step 5: Find human orthologs using the ortholog section.
- Step 6: Download genetic interaction data for network analysis.

## X6 – TAIR (The Arabidopsis Information Resource)

**Official Website URL:** <https://www.arabidopsis.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** TAIR is used to access comprehensive genomic, genetic, and functional data for the model *Arabidopsis thaliana*, providing the authoritative reference for Arabidopsis biology. It is used for gene function analysis, mutant phenotypes, expression data, and as a resource for plant biology research. TAIR is the primary reference for plant genomics.

**What Data It Contains:** TAIR contains genome annotations for *A. thaliana*, with gene records including sequence, expression, mutant phenotypes, protein interactions, GO annotations, and literature references. The database also contains information on Arabidopsis ecotypes and genetic resources.

**Main question it helps answer:** What is known about the function and phenotype of this Arabidopsis gene?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What phenotype results from T-DNA insertion in this Arabidopsis gene?
- What is the expression pattern of this gene in different tissues?
- What is the function of this Arabidopsis gene based on mutant analysis?

**Example use cases:**

- Accessing phenotype data for Arabidopsis mutants
- Studying gene expression in different tissues and conditions
- Finding T-DNA insertion lines for gene function studies

**Input Data Accepted:** Gene names, TAIR IDs (AT numbers), AGI codes

**Output Data Provided:** Gene records, phenotype data, expression data, T-DNA insertion information

**Strengths:** Comprehensive Arabidopsis gene information; T-DNA insertion line information; Extensive expression data; Freely accessible; Regularly updated

**Limitations:** Focused on Arabidopsis; limited for other plants; Some features require registration; Curation lag for very recent publications

**Common beginner mistakes:** Not using AGI codes (AT1G01010) for programmatic access; Not checking T-DNA insertion line availability; Not using TAIR for gene expression data

**When to Use It:** Use TAIR for Arabidopsis gene function analysis, phenotype data, and T-DNA insertion line identification.

**When NOT to Use It:** Do not use TAIR for non-Arabidopsis plants; use Phytozome or other plant databases.

**Related databases / alternatives:** Phytozome: Multi-plant genome database; PLAZA: Plant comparative genomics; FlyBase: Drosophila database (for comparison)

**How It Connects to Other Resources:** TAIR cross-references UniProt, Ensembl Plants, NCBI Gene, and GO.

**API / FTP / programmatic access:** FTP downloads at <https://www.arabidopsis.org/download/>. Limited API access. Python package tair-py available.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; some features require registration

**Citation / Recommended Reference:** Berardini TZ et al. (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, 53(8):474–485. <https://doi.org/10.1002/dvg.22877>

**Beginner-Friendly Explanation:** TAIR (The Arabidopsis Information Resource) is the primary database for *Arabidopsis thaliana*, a small flowering plant that is the most important model organism in plant biology. Arabidopsis was the first plant to have its genome completely sequenced, and it has been used to study fundamental plant processes like photosynthesis, flowering time, and responses to stress. TAIR provides comprehensive information about every Arabidopsis gene, including what happens when it is mutated, where it is expressed in the plant, and what its function is. TAIR is also a resource for finding T-DNA insertion lines—plants with a specific gene disrupted—which are widely used for gene function studies.

**Advanced Technical Explanation:** TAIR implements a comprehensive data model that captures gene structure (with alternative splicing), expression (from RNA-seq, microarray, and in situ hybridization data), mutant phenotypes (using the Plant Ontology, PO), T-DNA insertion line information (from SALK, SAIL, and other collections), protein interactions, GO annotations, and metabolic pathway information. TAIR uses the AGI (Arabidopsis Gene Identifier) system for gene naming (e.g., AT1G01010 for the first gene on chromosome 1). TAIR participates in the Alliance of Genome Resources for data sharing with other model organism databases.

**One practical workflow example:**

Step 1: Navigate to <https://www.arabidopsis.org> and search for your gene of interest.

Step 2: Review the gene summary, including function, expression, and phenotypes.

Step 3: Check the T-DNA insertion line section for available mutants.

Step 4: Order T-DNA insertion seeds from ABRC (Arabidopsis Biological Resource Center).

Step 5: Download gene expression data for computational analysis.

Step 6: Use GO annotations for functional enrichment analysis.

## X7 – PomBase

**Official Website URL:** <https://www.pombase.org>

**Resource Type:** Database / Knowledgebase

**Main Biological Domain:** DNA sequences / Omics

**What It Is Used For:** PomBase is used to access comprehensive genomic, genetic, and functional data for the fission yeast *Schizosaccharomyces pombe*, providing the authoritative reference for *S. pombe* biology. It is used for gene function analysis, mutant phenotypes, and as a resource for understanding fundamental eukaryotic biology. *S. pombe* is particularly valuable for studying cell cycle regulation and chromosome biology.

**What Data It Contains:** PomBase contains genome annotations for *S. pombe*, with gene records including sequence, expression, mutant phenotypes, genetic interactions, protein interactions, GO annotations, and literature references.

**Main question it helps answer:** What is known about the function and phenotype of this *S. pombe* gene?

**Typical user:** Researcher / Bioinformatician

**Example scientific questions:**

- What phenotype results from deletion of this *S. pombe* gene?
- What is the human ortholog of this *S. pombe* gene?
- What genetic interactions have been reported for this gene?

**Example use cases:**

- Accessing phenotype data for *S. pombe* mutants
- Studying cell cycle regulation in fission yeast
- Finding human orthologs of *S. pombe* genes

**Input Data Accepted:** Gene names, PomBase IDs, systematic names

**Output Data Provided:** Gene records, phenotype data, genetic interaction data

**Strengths:** Comprehensive *S. pombe* gene information; Excellent GO annotations; Freely accessible; Regularly updated

**Limitations:** Focused on *S. pombe*; limited for other organisms; Smaller genetic interaction dataset than SGD; Curation lag for very recent publications

**Common beginner mistakes:** Confusing PomBase (*S. pombe*) with SGD (*S. cerevisiae*); Not using systematic gene names for programmatic access

**When to Use It:** Use PomBase for *S. pombe* gene function analysis and phenotype data.

**When NOT to Use It:** Do not use PomBase for *S. cerevisiae*; use SGD instead.

**Related databases / alternatives:** SGD: *S. cerevisiae* database; FlyBase: *Drosophila* database; WormBase: *C. elegans* database

**How It Connects to Other Resources:** PomBase cross-references UniProt, Ensembl, NCBI Gene, and GO.



API / FTP / programmatic access: REST API at <https://www.pombase.org/api/>; FTP downloads at <https://www.pombase.org/data/>.

**Evidence/curation level:** Manually curated from primary literature; high quality

**Data Update Status:** Regular releases; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Lock A et al. (2019). PomBase 2018: user-driven reimplement of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Research*, 47(D1):D821–D827. <https://doi.org/10.1093/nar/gky961>

**Beginner-Friendly Explanation:** PomBase is the primary database for the fission yeast *Schizosaccharomyces pombe*, a yeast species that is particularly important for studying cell division and chromosome biology. Unlike budding yeast (*S. cerevisiae*, covered by SGD), fission yeast divides by splitting in the middle (like most cells), making it a better model for studying cell division in higher organisms. PomBase provides comprehensive information about every *S. pombe* gene, including what happens when it is deleted, how it interacts with other genes, and what human genes it is related to. Many genes involved in cell cycle regulation were first discovered in *S. pombe*.

**Advanced Technical Explanation:** PomBase implements a comprehensive data model similar to other model organism databases, with gene structure, expression, mutant phenotypes (using the Fission Yeast Phenotype Ontology, FYPO), genetic interactions, protein interactions, and GO annotations. PomBase is notable for its high-quality GO annotations, which are used as a benchmark for GO annotation quality. PomBase participates in the Alliance of Genome Resources for data sharing with other model organism databases. The PomBase REST API provides programmatic access to all data types.

**One practical workflow example:**

Step 1: Navigate to <https://www.pombase.org> and search for your *S. pombe* gene.

Step 2: Review the gene summary, including function, phenotypes, and interactions.

Step 3: Check the GO annotations for functional information.

Step 4: Find human orthologs using the ortholog section.

Step 5: Download gene data for computational analysis.

Step 6: Cross-reference with SGD for comparison with *S. cerevisiae*.

## X8 – RGD (Rat Genome Database)

---

**Resource Type:** Model Organism Database (Rat)

**Domain:** Rat genomics / Biomedical research

**Main Purpose:** Comprehensive genomic, genetic, and phenotypic data for the rat (*Rattus norvegicus*), including gene annotations, QTL data, disease annotations, and comparative genomics.

**Best Used For:** Rat genomics; cardiovascular and metabolic disease models; comparative genomics with human.

**Key Limitation:** Rat is less genetically tractable than mouse; fewer genetic tools available.

**Access / Licensing:** Open access; freely available at <https://rgd.mcw.edu>.

**Citation / Documentation:** Kaldunski M et al. (2022). The Rat Genome Database (RGD) facilitates genomic and phenotypic data integration across multiple species for biomedical research. *Nucleic Acids Research*, 50(D1):D619–D632. doi:10.1093/nar/gkab1007

## X9 – Xenbase

---

**Resource Type:** Model Organism Database (Xenopus)

**Domain:** Xenopus genomics / Developmental biology

**Main Purpose:** Genomic, genetic, and phenotypic data for *Xenopus laevis* and *Xenopus tropicalis*, including gene annotations, expression data, and developmental biology resources.

**Best Used For:** Xenopus developmental biology; gene expression during development; vertebrate developmental genetics.

**Key Limitation:** *X. laevis* is allotetraploid, complicating genomic analysis. *X. tropicalis* is the preferred genetic model.

**Access / Licensing:** Open access; freely available at <https://www.xenbase.org>.

**Citation / Documentation:** Karimi K et al. (2018). Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Research*, 46(D1):D861–D868. doi:10.1093/nar/gkx936

## X10 – Gramene

---

**Resource Type:** Database (Comparative Plant Genomics)

**Domain:** Plant genomics / Comparative genomics

**Main Purpose:** Comparative genomics database for plant species, integrating genomic, genetic, and functional data for rice, maize, wheat, sorghum, and other crop plants.

**Best Used For:** Comparative plant genomics; crop plant genomics; plant gene function; synteny analysis.

**Key Limitation:** Coverage varies by species; some crop plants are less well annotated than model plants.

**Access / Licensing:** Open access; freely available at <https://www.gramene.org>.

**Citation / Documentation:** Tello-Ruiz MK et al. (2023). Gramene 2023: a comparative genomics resource for crop and model plant species. *Nucleic Acids Research*, 51(D1):D1601–D1612. doi:10.1093/nar/gkac1017



## X11 – VEuPathDB (Eukaryotic Pathogen, Vector and Host Informatics Resource)

---

**Resource Type:** Database (Eukaryotic Pathogens)

**Domain:** Parasitology / Infectious disease / Genomics

**Main Purpose:** Integrated genomic and functional data for eukaryotic pathogens and their vectors, including Plasmodium, Toxoplasma, Leishmania, Trypanosoma, and other parasites.

**Best Used For:** Parasite genomics; infectious disease research; drug target identification in parasites.

**Key Limitation:** Specialized for eukaryotic pathogens; not suitable for bacterial or viral pathogens.

**Access / Licensing:** Open access; freely available at <https://veupathdb.org>.

**Citation / Documentation:** Amos B et al. (2022). VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. Nucleic Acids Research, 50(D1):D898–D911. doi:10.1093/nar/gkab929

## X12 – BV-BRC (Bacterial and Viral Bioinformatics Resource Center)

---

**Resource Type:** Database (Bacterial and Viral Pathogens)

**Domain:** Microbiology / Infectious disease / Genomics

**Main Purpose:** Comprehensive genomic and functional data for bacterial and viral pathogens, integrating data from PATRIC (bacterial) and IRD/ViPR (viral) resources.

**Best Used For:** Bacterial and viral pathogen genomics; antimicrobial resistance; comparative pathogen genomics.

**Key Limitation:** Specialized for pathogens; not suitable for non-pathogenic organisms.

**Related Resources:** VEuPathDB (eukaryotic pathogens), NCBI (general genomics), CARD (antimicrobial resistance)

**Access / Licensing:** Open access; freely available at <https://www.bv-brc.org>.

**Citation / Documentation:** Olson RD et al. (2023). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Research, 51(D1):D678–D689. doi:10.1093/nar/gkac1003

## BEGINNER EXAMPLE (Category X):

---

A genetics student has identified a novel gene in a *Drosophila* genetic screen that affects wing development. They search FlyBase for the gene and find that it has a human ortholog (NOTCH1) identified by DIOPT. They then check MGI to find mouse models with NOTCH1 mutations and find that Notch1 knockout mice have severe developmental defects. They use WormBase to check if the *C. elegans* ortholog (*lin-12*) has similar phenotypes.

## ADVANCED EXAMPLE (Category X):

---

A systems biologist is studying the conservation of genetic interaction networks across model organisms. They download the complete genetic interaction datasets from SGD (yeast) and WormBase (*C. elegans*), identify orthologous gene pairs using OrthoFinder, and compare the genetic interaction networks. They find that ~30% of genetic interactions are conserved between yeast and worm orthologs, and that conserved interactions are enriched for essential biological processes.

## CONFUSION POINTS (Category X):

---

SGD is for *S. cerevisiae* (budding yeast); PomBase is for *S. pombe* (fission yeast). They are different organisms.

Model organism gene names often differ from human gene names; always check orthologs.

Morpholino knockdown data in ZFIN may have off-target effects; genetic mutant data is more reliable.

Mouse phenotypes may not always translate to human phenotypes.

Alliance of Genome Resources (<https://www.alliancegenome.org>) provides cross-species data integration.

## DECISION GUIDE (Category X):

---

Studying *Drosophila*? → FlyBase; Studying *C. elegans*? → WormBase; Studying mouse? → MGI; Studying zebrafish? → ZFIN; Studying *S. cerevisiae* (budding yeast)? → SGD; Studying *Arabidopsis*? → TAIR; Studying *S. pombe* (fission yeast)? → PomBase; Need cross-species comparison? → Alliance of Genome Resources (<https://www.alliancegenome.org>)

## Category Y: Data Standards, Repositories, and FAIR Resources

### OVERVIEW

The FAIR principles—Findability, Accessibility, Interoperability, and Reusability—have become the guiding framework for data management in the life sciences. Published in 2016, the FAIR principles provide a set of guidelines for making scientific data more useful to both humans and machines. FAIR data is findable (with rich metadata and persistent identifiers), accessible (retrievable through standardized protocols), interoperable (using standardized formats and vocabularies), and reusable (with clear licensing and provenance information). The adoption of FAIR principles has driven the development of new databases, standards, and tools for data management and sharing.

The resources in this category serve different but complementary roles in the FAIR data ecosystem. FAIRsharing is a registry of data standards, databases, and policies that helps researchers find the appropriate standards and repositories for their data. BioSamples and BioProject provide standardized repositories for biological sample and project metadata, ensuring that experimental context is preserved alongside the data. BioStudies provides a repository for complete study datasets, including data that does not fit into specialized databases. Together, these resources form the infrastructure for FAIR data sharing in the life sciences.

A key challenge in implementing FAIR principles is the diversity of data types and research domains in the life sciences. Different communities have developed different standards and vocabularies, and achieving true interoperability requires agreement on common standards. FAIRsharing plays a crucial role in documenting and promoting these standards, helping researchers navigate the complex landscape of data standards and repositories. Researchers should consult FAIRsharing when planning data management strategies for new projects, to ensure that their data will be findable and reusable by others.

## Y1 – FAIRsharing

**Official Website URL:** <https://fairsharing.org>

**Resource Type:** Database / Registry / Portal

**Main Biological Domain:** Omics / Systems biology

**What It Is Used For:** FAIRsharing is used to discover, understand, and cite the data standards, databases, and data policies that are relevant to a specific research domain. It is used by researchers to find appropriate repositories for their data, by journals to recommend data deposition standards, and by funders to define data management requirements. FAIRsharing provides a curated registry of over 4,000 databases, standards, and policies.

**What Data It Contains:** FAIRsharing contains records for over 4,000 databases, over 1,000 data standards, and over 200 data policies, with metadata describing their scope, domain, status, and relationships. Records include information on data formats, ontologies, identifier schemes, and community standards.

**Main question it helps answer:** What databases and standards should I use for my research domain, and what are the data deposition requirements for my target journal?

**Typical user:** Researcher / Bioinformatician / Data analyst / Beginner student

**Example scientific questions:**

- What databases are recommended for depositing proteomics data?
- What data standards are used for genomics data?
- What are the data deposition requirements for Nature journals?

**Example use cases:**

- Finding appropriate repositories for data deposition
- Identifying relevant data standards for a research domain
- Checking journal data deposition requirements

**Input Data Accepted:** Research domain names, database names, standard names

**Output Data Provided:** Database records, standard records, policy records, relationships

**Strengths:** Comprehensive registry of databases and standards; Covers all life science domains; Freely accessible; Regularly updated; Used by major journals and funders

**Limitations:** Coverage may be incomplete for some domains; Record quality varies; Not a data repository itself; Some records may be outdated

**Common beginner mistakes:** Confusing FAIRsharing with a data repository (it is a registry, not a repository); Not checking FAIRsharing before choosing a data repository; Not using FAIRsharing to find relevant data standards

**When to Use It:** Use FAIRsharing when planning data management, choosing data repositories, or identifying relevant data standards. Consult FAIRsharing before submitting data to ensure compliance with journal and funder requirements.

**When NOT to Use It:** Do not use FAIRsharing to deposit data; use the appropriate repository instead.

**Related databases / alternatives:** re3data: Registry of research data repositories; BioSamples: Biological sample metadata; BioProject: Project metadata

**How It Connects to Other Resources:** FAIRsharing records link to databases, standards, and policies. FAIRsharing is used by journals (Nature, PLOS, etc.) and funders (NIH, Wellcome Trust) to define data requirements.

**API / FTP / programmatic access:** REST API at <https://api.fairsharing.org/>; returns JSON.

**Evidence/curation level:** Community-curated; quality-controlled by FAIRsharing team

**Data Update Status:** Continuously updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Sansone SA et al. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4):358–367. <https://doi.org/10.1038/s41587-019-0080-8>

**Beginner-Friendly Explanation:** FAIRsharing is a website that helps researchers find the right databases and standards for their research. When you finish a study and want to share your data, you need to know which database to deposit it in and what format to use. FAIRsharing provides a searchable registry of thousands of databases, data standards, and data policies, organized by research domain. You can search for databases that accept your type of data, find out what standards are used in your field, and check what your target journal requires for data deposition. FAIRsharing is particularly useful for early-career researchers who are new to data management.

**Advanced Technical Explanation:** FAIRsharing implements a comprehensive metadata schema for databases (covering scope, domain, taxonomy, data types, standards used, access conditions, and status), standards (covering type, domain, format, and relationships to databases), and policies (covering scope, requirements, and relationships to databases and standards). FAIRsharing uses persistent identifiers (DOIs) for all records and provides a REST API for programmatic access. FAIRsharing is integrated into the data management workflows of major journals (Nature, PLOS, eLife) and funders (NIH, Wellcome Trust, BBSRC), which use FAIRsharing records to define data deposition requirements.

#### **One practical workflow example:**

Step 1: Navigate to <https://fairsharing.org> and search for your research domain (e.g., "proteomics").

Step 2: Browse the recommended databases for your data type.

Step 3: Check the standards used by each database.

Step 4: Review the data policies of your target journal.

Step 5: Choose the appropriate repository and standard for your data.

Step 6: Cite the FAIRsharing record for the database in your methods section.

## Y2 – BioSamples

**Official Website URL:** <https://www.ebi.ac.uk/biosamples>

**Resource Type:** Repository / Database

**Main Biological Domain:** Omics / Clinical genomics

**What It Is Used For:** BioSamples is used to store and access standardized metadata for biological samples, providing persistent identifiers (BioSample accessions) for samples used in biological experiments. It is used to ensure that sample metadata is preserved alongside experimental data, enabling data reuse and integration. BioSamples accessions are required by many journals and databases for data submission.

**What Data It Contains:** BioSamples contains metadata for over 20 million biological samples, including organism, tissue, cell type, disease state, treatment, and other experimental conditions. Samples are linked to experimental data in ArrayExpress, ENA, and other EBI databases.

**Main question it helps answer:** What are the metadata and experimental conditions for this biological sample?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What are the metadata for this biological sample?
- What samples are available for this tissue and disease condition?
- What experimental data is linked to this sample?

**Example use cases:**

- Registering biological samples for data submission
- Accessing sample metadata for data reuse
- Finding samples with specific characteristics for meta-analyses

**Input Data Accepted:** Sample metadata (organism, tissue, disease, treatment, etc.)

**Output Data Provided:** BioSample accessions, metadata records, links to experimental data

**Strengths:** Standardized sample metadata; Persistent identifiers for samples; Links to experimental data; Freely accessible; Required by many journals

**Limitations:** Metadata quality varies across submissions; Some metadata fields may be incomplete; Not a data repository (stores metadata only); Controlled-access samples may have restricted metadata

**Common beginner mistakes:** Not registering samples before data submission; Not providing complete metadata for samples; Confusing BioSamples with BioProject (different purposes)

**When to Use It:** Use BioSamples to register biological samples before data submission and to access sample metadata for data reuse.

**When NOT to Use It:** Do not use BioSamples to deposit experimental data; use the appropriate data repository (GEO, ArrayExpress, ENA, etc.).

**Related databases / alternatives:** BioProject: Project-level metadata; NCBI BioSample: NCBI equivalent; SRA: Sequence data repository

**How It Connects to Other Resources:** BioSamples accessions are used in ArrayExpress, ENA, GEO, and other databases. BioSamples is linked to BioProject for project-level metadata.

**API / FTP / programmatic access:** REST API at <https://www.ebi.ac.uk/biosamples/samples>; returns JSON. Python package biosamples-client available.

**Evidence/curation level:** Submitter-provided metadata; some curation by EBI

**Data Update Status:** Continuously updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Courtot M et al. (2022). BioSamples database: FAIRer samples metadata to accelerate research and industry. *Nucleic Acids Research*, 50(D1):D1500–D1507. <https://doi.org/10.1093/nar/gkab1046>

**Beginner-Friendly Explanation:** BioSamples is a database that stores information about biological samples used in research experiments. When you perform an experiment—say, RNA-seq on human liver tissue—you need to record details about your sample: what organism it came from, what tissue it was, what disease state the donor had, and so on. BioSamples provides a standardized way to record this information and assigns a unique identifier (BioSample accession) to each sample. This identifier is then used when you deposit your experimental data in other databases, ensuring that the sample information is always linked to the data. Many journals and databases require BioSample accessions for data submission.

**Advanced Technical Explanation:** BioSamples implements a flexible metadata schema that captures sample attributes as key-value pairs, with controlled vocabulary terms from ontologies (EFO, UBERON, NCBITaxon, etc.) for standardized annotation. BioSamples accessions (SAMEA, SAMN, SAMD prefixes for EBI, NCBI, and DDBJ respectively) are persistent identifiers that are cross-referenced across databases. BioSamples participates in the INSDC (International Nucleotide Sequence Database Collaboration) for data sharing between EBI, NCBI, and DDBJ.

#### **One practical workflow example:**

Step 1: Navigate to <https://www.ebi.ac.uk/biosamples> and register for an account.

Step 2: Submit your sample metadata using the BioSamples submission portal.

Step 3: Receive your BioSample accession (e.g., SAMEA12345678).

Step 4: Use the BioSample accession when depositing your experimental data in ArrayExpress or ENA.

Step 5: Use the BioSamples API to retrieve sample metadata programmatically.

Step 6: Include the BioSample accession in your publication methods section.



## Y3 – BioProject

---

**Database Name:** BioProject

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/bioproject>

**Resource Type:** Repository / Database

**Main Biological Domain:** Omics / DNA sequences

**What It Is Used For:** BioProject is used to store and access project-level metadata for biological research projects, providing persistent identifiers (BioProject accessions) for research projects. It is used to organize related datasets under a single project umbrella, enabling data discovery and reuse. BioProject accessions are required by NCBI databases (SRA, GenBank, GEO) for data submission.

**What Data It Contains:** BioProject contains metadata for over 500,000 research projects, including project description, organism, data types, and links to associated datasets in SRA, GenBank, GEO, and other NCBI databases.

**Main question it helps answer:** What research project generated this dataset, and what other datasets are associated with this project?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What datasets are associated with this research project?
- What BioProject accession should I use for my data submission?
- What projects have generated data for this organism?

**Example use cases:**

- Registering research projects for data submission to NCBI databases
- Finding all datasets associated with a specific project
- Organizing related datasets under a single project

**Input Data Accepted:** Project metadata (title, description, organism, data types)

**Output Data Provided:** BioProject accessions, metadata records, links to associated datasets

**Strengths:** Standardized project metadata; Persistent identifiers for projects; Links to associated datasets; Required by NCBI databases; Freely accessible

**Limitations:** Metadata quality varies across submissions; Not a data repository (stores metadata only); Some projects may have incomplete metadata

**Common beginner mistakes:** Not registering a BioProject before data submission to NCBI; Confusing BioProject with BioSamples (different levels of metadata); Not linking all related datasets to the same BioProject

**When to Use It:** Use BioProject to register research projects before data submission to NCBI databases and to organize related datasets.

**When NOT to Use It:** Do not use BioProject to deposit experimental data; use SRA, GenBank, or GEO instead.



**Related databases / alternatives:** BioSamples: Sample-level metadata (EBI); SRA: Sequence data repository; GEO: Gene expression data repository

**How It Connects to Other Resources:** BioProject accessions are used in SRA, GenBank, GEO, and other NCBI databases. BioProject is linked to BioSamples for sample-level metadata.

**API / FTP / programmatic access:** E-utilities API at <https://eutils.ncbi.nlm.nih.gov/>; Entrez bioproject database. FTP downloads at <https://ftp.ncbi.nlm.nih.gov/bioproject/>.

**Evidence/curation level:** Submitter-provided metadata; some curation by NCBI

**Data Update Status:** Continuously updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; public domain

**Citation / Recommended Reference:** Barrett T et al. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*, 40(D1):D57–D63. <https://doi.org/10.1093/nar/gkr1163>

**Beginner-Friendly Explanation:** BioProject is a database maintained by NCBI that stores information about research projects. When you are conducting a research study that will generate multiple types of data—for example, genome sequencing, RNA-seq, and ChIP-seq from the same set of samples—BioProject provides a way to organize all this data under a single project identifier. This makes it easier for other researchers to find all the data from your study. BioProject accessions are required when submitting data to NCBI databases like SRA (for sequencing data) and GEO (for gene expression data).

**Advanced Technical Explanation:** BioProject implements a hierarchical data model that captures project-level metadata including project type (primary submission, umbrella project), organism, data types, and project description. BioProject accessions (PRJNA for NCBI, PRJEB for EBI, PRJDB for DDBJ) are persistent identifiers that are cross-referenced across INSDC databases. BioProject participates in the INSDC for data sharing between NCBI, EBI, and DDBJ. The E-utilities API provides programmatic access to BioProject metadata.

#### **One practical workflow example:**

Step 1: Navigate to <https://www.ncbi.nlm.nih.gov/bioproject> and register for an NCBI account.

Step 2: Submit your project metadata using the BioProject submission portal.

Step 3: Receive your BioProject accession (e.g., PRJNA123456).

Step 4: Use the BioProject accession when submitting data to SRA, GenBank, or GEO.

Step 5: Link all related datasets to the same BioProject.

Step 6: Include the BioProject accession in your publication methods section.

## Y4 – BioStudies

**Official Website URL:** <https://www.ebi.ac.uk/biostudies>

**Resource Type:** Repository / Database

**Main Biological Domain:** Omics / Systems biology

**What It Is Used For:** BioStudies is used to store and access complete study datasets, including data that does not fit into specialized databases. It is used to deposit supplementary data, complete study packages, and datasets from studies that span multiple data types. BioStudies provides a flexible repository for any type of biological data.

**What Data It Contains:** BioStudies contains over 5 million studies with associated data files, including supplementary data from publications, complete study packages, and datasets from diverse biological domains. Studies are linked to publications and other databases.

**Main question it helps answer:** Where can I find the complete dataset from this published study?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- Where is the complete dataset from this publication?
- What supplementary data is available for this study?
- What studies have deposited data for this biological domain?

**Example use cases:**

- Depositing complete study datasets for publication
- Accessing supplementary data from published studies
- Finding datasets that span multiple data types

**Input Data Accepted:** Any type of biological data files, study metadata

**Output Data Provided:** Study records, data files, links to publications

**Strengths:** Flexible repository for any data type; Links to publications; Freely accessible; Regularly updated; Accepts data that does not fit specialized databases

**Limitations:** Less specialized than domain-specific databases; Data quality varies across submissions; Some studies may have incomplete metadata

**Common beginner mistakes:** Not using BioStudies for data that does not fit specialized databases; Not linking BioStudies records to publications; Not providing complete metadata for studies

**When to Use It:** Use BioStudies to deposit complete study datasets, particularly for data that does not fit into specialized databases, or to provide a single access point for all data from a study.

**When NOT to Use It:** For specialized data types (genomics, proteomics, etc.), use the appropriate specialized database (SRA, PRIDE, etc.) and link to BioStudies.

**Related databases / alternatives:** Zenodo: General research data repository; Figshare: General research data repository; Dryad: Research data repository

**How It Connects to Other Resources:** BioStudies records link to publications (PubMed, Europe PMC) and other EBI databases (ArrayExpress, ENA, PRIDE).

**API / FTP / programmatic access:** REST API at <https://www.ebi.ac.uk/biostudies/api/v1/>; returns JSON. FTP downloads available.

**Evidence/curation level:** Submitter-provided; some curation by EBI

**Data Update Status:** Continuously updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0 for most studies

**Citation / Recommended Reference:** Sarkans U et al. (2018). The BioStudies database—one stop shop for all data from a biological study. *Nucleic Acids Research*, 46(D1):D1266–D1270. <https://doi.org/10.1093/nar/gkx965>

**Beginner-Friendly Explanation:** BioStudies is a database maintained by the European Bioinformatics Institute that provides a flexible repository for complete study datasets. When you publish a paper, you often have data that does not fit neatly into specialized databases—for example, custom analysis scripts, processed data files, or data from multiple different types of experiments. BioStudies provides a place to deposit all of this data together, linked to your publication, so that other researchers can access the complete dataset. BioStudies is particularly useful for studies that span multiple data types or that generate data that does not fit into existing specialized databases.

**Advanced Technical Explanation:** BioStudies implements a flexible data model that captures study-level metadata (title, description, organism, data types, publication links) and file-level metadata (file name, type, description). Studies are assigned persistent accession numbers (S-BSST prefix) and linked to publications through PubMed IDs and DOIs. BioStudies integrates with other EBI databases through cross-references, enabling discovery of related data across databases. The BioStudies REST API provides programmatic access to study metadata and file downloads.

**One practical workflow example:**

- Step 1: Navigate to <https://www.ebi.ac.uk/biostudies> and register for an account.
- Step 2: Prepare your study data files and metadata.
- Step 3: Submit your study using the BioStudies submission portal.
- Step 4: Receive your BioStudies accession (e.g., S-BSST123456).
- Step 5: Include the BioStudies accession in your publication.
- Step 6: Link your BioStudies record to specialized database accessions (SRA, ArrayExpress, etc.).

## BEGINNER EXAMPLE (Category Y):

---

A graduate student is preparing to submit their RNA-seq data to a public repository. They consult FAIRsharing to find the recommended repository for RNA-seq data and find that ArrayExpress and GEO are the primary options. They register their samples in BioSamples to get BioSample accessions, register their project in BioProject to get a BioProject accession, and then submit their data to ArrayExpress. They also deposit their complete analysis scripts and processed data in BioStudies.

## ADVANCED EXAMPLE (Category Y):

---

A data manager at a research institute is developing a data management plan for a large multi-omics study. They use FAIRsharing to identify the appropriate repositories and standards for each data type (genomics: SRA/ENA; proteomics: PRIDE; metabolomics: MetaboLights). They register all samples in BioSamples and the project in BioProject, ensuring consistent metadata across all submissions. They use BioStudies to provide a single access point for all data from the study.

## CONFUSION POINTS (Category Y):

---

- FAIRsharing is a registry, not a repository; it does not store data.
- BioSamples and BioProject store metadata, not experimental data.
- BioProject (NCBI) and BioSamples (EBI) are different systems, though they are synchronized through INSDC.
- FAIR principles are guidelines, not a specific database or tool.
- Data deposition requirements vary by journal and funder; always check FAIRsharing for current requirements.

## DECISION GUIDE (Category Y):

---

Need to find the right database for your data? → FAIRsharing

Need to register biological samples? → BioSamples (EBI) or NCBI BioSample

Need to register a research project for NCBI submission? → BioProject

Need to deposit a complete study dataset? → BioStudies

Need to check journal data requirements? → FAIRsharing

## Category Z: Database Directories and Resource Catalogs

### OVERVIEW

The bioinformatics database landscape is vast and constantly evolving, with thousands of databases covering every aspect of biological research. Navigating this landscape can be challenging, particularly for researchers who are new to a specific domain or who need to find resources for an unfamiliar data type. Database directories and resource catalogs provide curated, searchable registries of bioinformatics databases and tools, helping researchers find the resources they need. These meta-resources are essential for staying current with the rapidly changing database landscape.

The most authoritative source for bioinformatics database information is the annual database issue of Nucleic Acids Research (NAR), which has been published every January since 1993. Each year, the NAR database issue publishes papers describing new and updated databases, providing peer-reviewed descriptions of hundreds of resources. The NAR Molecular Biology Database Collection provides a searchable index of all databases described in NAR, making it the most comprehensive catalog of bioinformatics databases. bio.tools is a community-driven registry of bioinformatics tools and databases, providing standardized metadata for over 20,000 resources. Database Commons (NGDC) provides a comprehensive catalog of biological databases with a focus on Chinese and international resources.

A key challenge in using database directories is keeping up with the rapidly changing landscape. Databases are constantly being created, updated, and deprecated, and directory records may not always reflect the current status of a resource. Researchers should always verify the current status of a database before using it, particularly for resources that have not been updated recently. FAIRsharing (covered in Category Y) also serves as a database directory, with a focus on FAIR compliance and data standards. Together, these resources provide multiple perspectives on the bioinformatics database landscape.

## Z1 – NAR Molecular Biology Database Collection

**Official Website URL:** <https://www.oxfordjournals.org/nar/database/c> (also accessible at <https://www.nucleicacidsresearch.com/database-issue>)

**Resource Type:** Database / Portal / Literature Search Engine

**Main Biological Domain:** Omics / Systems biology

**What It Is Used For:** The NAR Molecular Biology Database Collection is used to discover and access peer-reviewed descriptions of bioinformatics databases published in the annual database issue of Nucleic Acids Research. It is used to find databases for specific biological domains, to access authoritative descriptions of database content and methods, and to cite databases in publications. The collection is the most comprehensive catalog of bioinformatics databases.

**What Data It Contains:** The NAR Database Collection contains records for over 1,800 databases described in NAR database issues from 1993 to the present, with links to the original publications and database websites. Records are organized by biological domain and include database descriptions, URLs, and citation information.

**Main question it helps answer:** What databases are available for this biological domain, and what are their peer-reviewed descriptions?

**Typical user:** Researcher / Bioinformatician / Beginner student

**Example scientific questions:**

- What databases are available for protein structure analysis?
- What is the peer-reviewed description of this database?
- What new databases were published in the latest NAR database issue?

**Example use cases:**

- Finding databases for a specific biological domain
- Accessing peer-reviewed database descriptions for citation
- Staying current with new database releases

**Input Data Accepted:** Biological domain names, database names, keywords

**Output Data Provided:** Database records, publication links, database URLs

**Strengths:** Peer-reviewed database descriptions; Comprehensive coverage of major databases; Annual updates; Freely accessible; Authoritative source for database citations

**Limitations:** Coverage limited to databases published in NAR; Some records may be outdated; Not all databases are published in NAR; Website interface may be limited

**Common beginner mistakes:** Not checking the NAR database issue for new databases; Not citing the NAR paper when using a database; Not verifying that the database is still active

**When to Use It:** Use the NAR Database Collection to find databases for a specific domain, to access peer-reviewed descriptions, and to find the correct citation for a database.

**When NOT to Use It:** For tools (not databases), use bio.tools instead. For FAIR compliance, use FAIRsharing.



**Related databases / alternatives:** bio.tools: Bioinformatics tool registry; FAIRsharing: FAIR database registry; Database Commons: Comprehensive database catalog

**How It Connects to Other Resources:** NAR database records link to database websites and publications. The NAR database issue is published annually in January.

**API / FTP / programmatic access:** Limited API access. Database records accessible through the NAR website.

**Evidence/curation level:** Peer-reviewed; high quality

**Data Update Status:** Annual updates (January database issue); actively maintained

**Licensing / access restrictions:** Freely accessible; Oxford University Press

**Citation / Recommended Reference:** Rigden DJ and Fernandez XM (2024). The 2024 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Research, 52(D1):D1–D9. <https://doi.org/10.1093/nar/gkad1173>

**Beginner-Friendly Explanation:** Every January, the journal Nucleic Acids Research publishes a special "database issue" that contains peer-reviewed descriptions of hundreds of bioinformatics databases. This has been happening since 1993, making it the most comprehensive and authoritative source of information about bioinformatics databases. The NAR Molecular Biology Database Collection provides a searchable index of all databases described in these annual issues, making it easy to find databases for any biological domain. When you use a database in your research, you should cite the NAR paper that describes it—the NAR Database Collection makes it easy to find the correct citation.

**Advanced Technical Explanation:** The NAR database issue is a peer-reviewed publication that provides standardized descriptions of bioinformatics databases, including their content, methods, and access information. Each database paper undergoes peer review by the NAR editorial board and external reviewers, ensuring a minimum quality standard. The NAR Database Collection provides a searchable index of all database papers, organized by biological domain (nucleotide sequences, protein sequences, structures, genomics, etc.). The collection is updated annually with new database papers from the January issue.

**One practical workflow example:**

Step 1: Navigate to <https://www.nucleicacidsresearch.com/database-issue>.

Step 2: Browse the database categories or search for your domain of interest.

Step 3: Find the database you need and click through to the NAR paper.

Step 4: Read the peer-reviewed description to understand the database content and methods.

Step 5: Note the citation information for use in your publication.

Step 6: Navigate to the database website using the URL provided in the paper.

## Z2 – bio.tools

**Official Website URL:** <https://bio.tools>

**Resource Type:** Database / Registry / Portal

**Main Biological Domain:** Omics / Systems biology

**What It Is Used For:** bio.tools is used to discover, understand, and cite bioinformatics tools and databases, providing a community-driven registry with standardized metadata for over 20,000 resources. It is used by researchers to find tools for specific tasks, by developers to register their tools, and by workflow systems to discover available tools. bio.tools uses the EDAM ontology for standardized tool annotation.

**What Data It Contains:** bio.tools contains records for over 20,000 bioinformatics tools and databases, with standardized metadata including tool name, description, function, input/output data types, operating system, programming language, and license. Records use the EDAM ontology for standardized annotation.

Main question it helps answer: What bioinformatics tools are available for this specific task, and what are their input/output requirements?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What tools are available for multiple sequence alignment?
- What tools accept FASTQ input and produce BAM output?
- What tools are available for protein structure prediction?

**Example use cases:**

- Finding tools for a specific bioinformatics task
- Discovering tools with specific input/output requirements
- Registering a new bioinformatics tool

**Input Data Accepted:** Tool names, biological domains, data types, functions

**Output Data Provided:** Tool records, metadata, links to tool websites and publications

**Strengths:** Comprehensive tool registry; Standardized metadata using EDAM ontology; Community-driven; Freely accessible; Integrated with workflow systems (Galaxy, CWL)

**Limitations:** Record quality varies across submissions; Some records may be outdated; Coverage may be incomplete for some domains; Not all tools are registered

**Common beginner mistakes:** Not checking bio.tools before developing a new tool (may already exist); Not registering new tools in bio.tools; Not using EDAM ontology terms for tool annotation

**When to Use It:** Use bio.tools to find bioinformatics tools for specific tasks, to discover tools with specific input/output requirements, and to register new tools.

**When NOT to Use It:** For databases (not tools), use the NAR Database Collection or FAIRsharing.

**Related databases / alternatives:** FAIRsharing: FAIR database registry; NAR Database Collection: Database catalog; Bioconductor: R bioinformatics packages; Galaxy ToolShed: Galaxy tool registry



**How It Connects to Other Resources:** bio.tools records link to tool websites, publications, and workflow systems. bio.tools uses the EDAM ontology for standardized annotation.

**API / FTP / programmatic access:** REST API at <https://bio.tools/api/>; returns JSON. Python package biotools-client available.

**Evidence/curation level:** Community-submitted; some curation by bio.tools team

**Data Update Status:** Continuously updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available; Creative Commons Attribution 4.0

**Citation / Recommended Reference:** Ison J et al. (2016). Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Research*, 44(D1):D38–D47. <https://doi.org/10.1093/nar/gkv1116>

**Beginner-Friendly Explanation:** bio.tools is a website that provides a searchable registry of bioinformatics tools and databases. With thousands of bioinformatics tools available, it can be difficult to know which tool to use for a specific task. bio.tools helps by providing standardized descriptions of tools, including what they do, what types of data they accept as input, and what types of data they produce as output. You can search for tools by function (e.g., "sequence alignment"), data type (e.g., "FASTQ"), or biological domain (e.g., "genomics"). bio.tools is particularly useful for finding tools that fit into a specific workflow.

**Advanced Technical Explanation:** bio.tools implements the EDAM (Editable Data, Algorithms, and Methods) ontology for standardized tool annotation, covering operations (what the tool does), topics (biological domain), data types (input/output), and formats (file formats). This standardized annotation enables semantic search and workflow integration. bio.tools is integrated with workflow systems including Galaxy, CWL (Common Workflow Language), and Nextflow, enabling automated tool discovery and workflow construction. The bio.tools REST API provides programmatic access to all tool records.

#### **One practical workflow example:**

Step 1: Navigate to <https://bio.tools> and search for your task of interest (e.g., "RNA-seq quantification").

Step 2: Browse the results and filter by operating system, programming language, or license.

Step 3: Click on a tool to see its full metadata, including input/output requirements.

Step 4: Follow the link to the tool website for installation and documentation.

Step 5: Use the bio.tools API to discover tools programmatically for workflow integration.

Step 6: Register your own tools in bio.tools to make them discoverable.

## Z3 – Database Commons

**Official Website URL:** <https://ngdc.cncb.ac.cn/databasecommons>

**Resource Type:** Database / Registry / Portal

**Main Biological Domain:** Omics / Systems biology

**What It Is Used For:** Database Commons is used to discover and access biological databases from around the world, providing a comprehensive catalog with standardized metadata. It is maintained by the National Genomics Data Center (NGDC) in China and provides a global perspective on the bioinformatics database landscape, including many databases from Chinese research institutions.

**What Data It Contains:** Database Commons contains records for over 5,000 biological databases, with metadata including database name, description, URL, biological domain, data types, and access information. Records cover databases from all biological domains and from research institutions worldwide.

**Main question it helps answer:** What biological databases are available for this domain, including resources from international institutions?

**Typical user:** Researcher / Bioinformatician / Data analyst

**Example scientific questions:**

- What databases are available for this biological domain?
- What databases have been developed by Chinese research institutions?
- What is the current status of this database?

**Example use cases:**

- Finding databases for a specific biological domain
- Discovering databases from international institutions
- Checking the current status of a database

**Input Data Accepted:** Biological domain names, database names, keywords

**Output Data Provided:** Database records, metadata, links to database websites

**Strengths:** Comprehensive coverage including international databases; Standardized metadata; Freely accessible; Regularly updated; Includes databases not covered by NAR

**Limitations:** Record quality varies; Some records may be outdated; Less peer-reviewed than NAR Database Collection; Interface may be less intuitive than bio.tools

**Common beginner mistakes:** Not checking Database Commons for domain-specific databases; Not verifying that the database is still active

**When to Use It:** Use Database Commons to find databases for a specific domain, particularly when looking for international resources or databases not covered by NAR.

**When NOT to Use It:** For peer-reviewed database descriptions, use the NAR Database Collection.

**Related databases / alternatives:** NAR Database Collection: Peer-reviewed database catalog; bio.tools: Tool registry; FAIRsharing: FAIR database registry

**How It Connects to Other Resources:** Database Commons records link to database websites and publications.

**API / FTP / programmatic access:** Limited API access. Database records accessible through the NGDC website.

**Evidence/curation level:** Community-submitted; curated by NGDC team

**Data Update Status:** Regularly updated; actively maintained as of 2024

**Licensing / access restrictions:** Freely available

**Citation / Recommended Reference:** Zhao W et al. (2020). Database Commons: a catalog of worldwide biological databases. *Nucleic Acids Research*, 48(D1):D1–D7. <https://doi.org/10.1093/nar/gkz1060>

**Beginner-Friendly Explanation:** Database Commons is a catalog of biological databases maintained by the National Genomics Data Center in China. It provides a comprehensive list of databases from research institutions around the world, including many databases from Chinese institutions that may not be covered by other catalogs. For each database, Database Commons provides a description, the URL, and information about what types of data it contains. Database Commons is a useful resource for finding databases that are not covered by the NAR Database Collection or bio.tools.

**Advanced Technical Explanation:** Database Commons implements a standardized metadata schema for biological databases, covering database name, description, URL, biological domain, data types, access conditions, and status. The catalog is maintained by the NGDC (National Genomics Data Center) at the Beijing Institute of Genomics, Chinese Academy of Sciences. Database Commons provides a global perspective on the bioinformatics database landscape, with particular coverage of databases from Chinese research institutions. The catalog is updated regularly with new database records.

**One practical workflow example:**

Step 1: Navigate to <https://ngdc.cncb.ac.cn/databasecommons>.

Step 2: Search for your biological domain of interest.

Step 3: Browse the results and filter by data type or organism.

Step 4: Click on a database to see its full metadata.

Step 5: Follow the link to the database website.

Step 6: Cross-reference with NAR Database Collection for peer-reviewed descriptions.

## BEGINNER EXAMPLE (Category Z):

---

A new graduate student wants to find databases for studying protein-protein interactions. They search the NAR Database Collection for "protein interaction" and find descriptions of STRING, BioGRID, IntAct, and other databases. They then check bio.tools for tools that can analyze protein interaction data. They use FAIRsharing to find the data standards used by these databases and to check the data deposition requirements for their target journal.

## ADVANCED EXAMPLE (Category Z):

---

A bioinformatics core facility manager is developing a resource guide for their institution. They use the NAR Database Collection to identify the most important databases for each biological domain, bio.tools to catalog the tools available for each analysis type, and FAIRsharing to document the data standards and policies relevant to their research community. They use Database Commons to find additional resources from international institutions.

## CONFUSION POINTS (Category Z):

---

The NAR Database Collection covers databases described in NAR; many important databases are not in NAR.

bio.tools covers tools AND databases; the NAR Database Collection covers only databases.

FAIRsharing covers databases, standards, AND policies; it is the most comprehensive for FAIR compliance.

Database Commons has the broadest coverage but is less peer-reviewed than NAR.

These directories are meta-resources; they do not contain biological data themselves.

## DECISION GUIDE (Category Z):

---

Need peer-reviewed database descriptions? → NAR Database Collection

Need to find bioinformatics tools? → bio.tools

Need to find international databases? → Database Commons

Need to check FAIR compliance? → FAIRsharing

Need to find data standards? → FAIRsharing

Need to cite a database? → NAR Database Collection (find the NAR paper)

## PART II — Expanded Omics, Clinical, Proteomics, Metabolomics, and Translational Data Resources

### Category AA: Controlled Human Genomics and Phenotype Repositories

#### Category Overview

Controlled human genomics and phenotype repositories are specialized archives designed to store, manage, and provide access to sensitive human genomic data linked to phenotypic and clinical information. Unlike public sequencing archives (SRA, ENA, DRA), which store data that can be freely downloaded by anyone, controlled-access archives require formal data access approval, data access committee (DAC) review, and compliance with participant consent restrictions. This distinction is not merely procedural — it reflects fundamental ethical and legal obligations to protect the privacy and autonomy of research participants who consented to data sharing under specific conditions.

The primary controlled-access archives are the European Genome-phenome Archive (EGA) at EMBL-EBI and the database of Genotypes and Phenotypes (dbGaP) at NCBI. EGA is the primary repository for European research cohorts and is required by many European funding agencies and journals. dbGaP is the primary repository for US-funded research, particularly NIH-funded studies. Both archives store individual-level genotype data (typically whole-genome or whole-exome sequencing, or SNP array data) linked to phenotypic and clinical metadata.

#### When to use controlled-access archives:

- When depositing human genomic data with individual-level phenotype or clinical information.
- When accessing individual-level human genomic data for research purposes.
- When participant consent specifies controlled data sharing.
- When regulatory requirements (GDPR, HIPAA, national data protection laws) mandate controlled access.

#### When NOT to use controlled-access archives:

- For non-human genomic data — use public repositories (SRA, ENA).
- For human genomic data that has been fully anonymized and for which consent permits open sharing — use public repositories.
- For GWAS summary statistics (not individual-level data) — these can often be shared openly via the GWAS Catalog.



### Common beginner mistakes:

- Depositing sensitive human genomic data in public repositories (SRA, GEO) — this is a serious ethical and potentially legal violation.
- Assuming that removing names and dates of birth is sufficient anonymization — re-identification from genomic data is possible even without direct identifiers.
- Attempting to access controlled-access data without an approved data access request.
- Confusing GWAS summary statistics (often publicly available) with individual-level genotype data (always controlled-access).

**WARNING: Human genomic data linked to phenotype information is sensitive personal data under GDPR (EU), HIPAA (US), and equivalent national laws. Unauthorized access, sharing, or deposition of such data can result in legal liability, loss of research privileges, and harm to research participants. Always consult your institutional data protection officer and ethics committee before handling human genomic data.**

## AA1 — EGA (European Genome-phenome Archive)

**Official Website URL:** <https://ega-archive.org>

**Resource Type:** Controlled-Access Archive

**Main Biological Domain:** Human genomics / Clinical genomics / Phenomics

**Short Definition:** EGA is the primary European controlled-access archive for human genomic and phenotypic data, providing secure storage and managed access to sensitive research data from human subjects.

**What It Is Used For:** EGA is used to deposit and access individual-level human genomic data (whole-genome sequencing, whole-exome sequencing, SNP arrays, RNA-seq, epigenomics) linked to phenotypic and clinical metadata, under controlled access governed by data access committees. It is the standard deposition archive for European research cohorts and is required by many European funding agencies (Wellcome Trust, MRC, EU Horizon) and journals.

**What Data It Contains:** EGA contains individual-level human genomic data from thousands of studies, including raw sequencing reads (FASTQ, BAM, CRAM), processed variant calls (VCF), SNP array data, and associated phenotypic and clinical metadata. Data is organized into studies, datasets, and data access policies. As of 2024, EGA hosts data from over 4,000 studies with over 1 million samples. Data types include germline and somatic variants, expression data, methylation data, and clinical records.

**Main Scientific Question It Helps Answer:** How can I securely deposit or access individual-level human genomic data with phenotypic information, in compliance with participant consent and data protection regulations?

**Typical Users:** Researchers depositing human genomic data; researchers requesting access to controlled-access human genomic datasets; data access committees; institutional data protection officers.

### Example Scientific Questions:

- Where should I deposit whole-genome sequencing data from a European patient cohort?
- How do I access individual-level genotype data from a published GWAS study?
- What is the accession number for the EGA dataset from this published study?
- How do I set up a Data Access Committee for my study?

### Example Use Cases:

- Depositing WGS data from a European rare disease cohort for controlled sharing.
- Requesting access to a published GWAS dataset for replication analysis.
- Accessing RNA-seq data from a cancer cohort for differential expression analysis.
- Federating data access across multiple European cohorts using EGA's federated model.

**Input Data Accepted:** Raw sequencing files (FASTQ, BAM, CRAM), processed files (VCF, BED), SNP array data, phenotypic metadata (in standardized formats), data access policies, consent documentation.

**Output Data Provided:** Controlled-access datasets with EGA accession numbers (EGAS for studies, EGAD for datasets, EGAF for files, EGAX for experiments, EGAR for runs, EGAP for policies, EGAC for DACs). Data available for download after access approval.

**Strengths:**

- The primary European controlled-access archive; required by many European funders and journals.
- Supports federated data access, allowing data to remain at national nodes while being discoverable through EGA.
- Provides infrastructure for Data Access Committees (DACs) to manage access requests.
- Integrates with EMBL-EBI resources (Ensembl, EVA, GWAS Catalog) for cross-referencing.
- Supports multiple data types: genomics, transcriptomics, epigenomics, proteomics, metabolomics.
- Provides stable, citable accession numbers for all deposited datasets.
- GDPR-compliant infrastructure for European human subject data.

**Limitations:**

- Access approval can take weeks to months depending on the DAC.
- Data access is restricted to approved research purposes; re-use for other purposes requires new approval.
- Federated EGA nodes (national nodes) may have different access procedures and timelines.
- Not suitable for non-human data or data that can be shared openly.
- Submission process requires significant metadata preparation and can be technically complex.
- Some older datasets may have incomplete metadata.

**Common Beginner Mistakes:**

- Attempting to access EGA data without submitting a formal data access request.
- Confusing EGA (controlled-access) with ENA (public sequencing archive) — they are different resources.
- Depositing data without first establishing a Data Access Committee and data access policy.
- Assuming all EGA data is accessible — each dataset has its own access policy and DAC.
- Not recording the EGA accession numbers and access approval details for reproducibility.

**When to Use It:** Use EGA when depositing or accessing individual-level human genomic data with phenotypic information from European research cohorts, or when required by European funders or journals. Use EGA when participant consent specifies controlled data sharing.

**When NOT to Use It:** Do not use EGA for non-human data (use SRA/ENA), for data that can be shared openly (use ENA/GEO), or for GWAS summary statistics without individual-level data (use GWAS Catalog). Do not use EGA for US-funded studies where dbGaP is required.

**Related Databases or Alternatives:** dbGaP (US equivalent); JGA (Japanese equivalent); AnVIL (NIH cloud platform); GWAS Catalog (open-access GWAS summary statistics); ENA (public sequencing archive); BioStudies (study metadata).

**How It Connects to Other Resources:** EGA integrates with EMBL-EBI resources including Ensembl (gene annotation), EVA (variant archive), GWAS Catalog (GWAS results), and BioStudies (study metadata). EGA accession numbers are cited in publications and cross-referenced in GWAS Catalog entries.

**API / FTP / Bulk Download / Programmatic Access:** EGA REST API at <https://ega-archive.org/metadata/v2/> for metadata queries. EGA download client (ega-download-client, Python-based) for data download after access approval. Pyega3 Python package for programmatic access. Federated EGA nodes may have separate APIs.

**Evidence or Curation Level:** Community-submitted; data quality depends on submitter; EGA performs format validation but not biological curation.

**Update Status:** Continuously updated with new submissions; actively maintained by EMBL-EBI as of 2025.

**Licensing or Access Restrictions:** Controlled access; each dataset has its own data access policy specifying permitted uses, consent restrictions, and access conditions. Access requires formal application to the relevant DAC. GDPR-compliant.

**Citation / Recommended Reference:** Freeberg MA et al. (2022). The European Genome-phenome Archive in 2021. *Nucleic Acids Research*, 50(D1):D980–D987. doi:10.1093/nar/gkab1059

**Beginner-Friendly Explanation:** EGA is a secure archive for human genomic data that is too sensitive to share publicly. When researchers collect DNA samples from patients or volunteers, the participants consent to their data being shared — but only with other researchers who have a legitimate scientific purpose and agree to protect the data. EGA provides the infrastructure for this controlled sharing: researchers deposit their data in EGA, set up a committee to review access requests, and other researchers can apply for access. EGA is like a secure library where you need to show your credentials and explain why you need the book before you can borrow it.

**Advanced Technical Explanation:** EGA implements a federated architecture with a central EGA at EMBL-EBI and national EGA nodes (e.g., EGA Sweden, EGA Finland, EGA Spain) that allow data to remain within national jurisdictions while being discoverable through the central EGA portal. The metadata model uses a hierarchical structure: Study > Dataset > Experiment > Run > File, with separate objects for Data Access Policies (EGAP) and Data Access Committees (EGAC). Data is encrypted at rest and in transit using AES-256 encryption. The EGA download client uses the Crypt4GH standard for file encryption, enabling secure data transfer. EGA supports the GA4GH Data Use Ontology (DUO) for standardized consent term representation.

#### **Practical Workflow Example:**

Step 1: Register at <https://ega-archive.org> and create a submitter account.

Step 2: Prepare metadata (study, samples, experiments, runs) in EGA XML format or using the EGA submission portal.

Step 3: Encrypt data files using the EGA encryption tool.

Step 4: Upload encrypted files via FTP or Aspera.

Step 5: Submit metadata and link to uploaded files.

Step 6: Establish a Data Access Committee and define a data access policy.

Step 7: Receive EGA accession numbers (EGAS, EGAD) for citation in publications.

**Reproducibility Notes:** Record the EGA study accession (EGAS), dataset accession (EGAD), and data access approval reference number in your methods section. Record the access date and the version of the data downloaded. Note any data processing steps applied after download. For federated EGA nodes, record the specific node accessed.

**Quality-Control Notes:** EGA performs format validation but not biological quality control. Assess data quality using standard QC tools (FastQC, MultiQC) after download. Check sample metadata completeness before analysis. Verify that the data access policy permits your intended use.

## AA2 — dbGaP (database of Genotypes and Phenotypes)

---

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/gap>

**Resource Type:** Controlled-Access Archive

**Main Biological Domain:** Human genomics / Clinical genomics / Phenomics

**Short Definition:** dbGaP is the primary US controlled-access archive for human genotype and phenotype data, hosted by NCBI, providing managed access to individual-level genomic data from NIH-funded and other studies.

**What It Is Used For:** dbGaP is used to deposit and access individual-level human genomic data (GWAS, WGS, WES, SNP arrays) linked to phenotypic and clinical metadata, under controlled access governed by NIH data access committees. It is required for most NIH-funded studies involving human genomic data and is the standard US repository for GWAS datasets.

**What Data It Contains:** dbGaP contains individual-level genotype data (SNP arrays, WGS, WES), phenotypic data (clinical measurements, disease status, demographics), and pedigree information from thousands of studies. Data types include PLINK format genotype files, VCF files, phenotype tables, and study documentation. As of 2024, dbGaP hosts data from over 1,000 studies with millions of participants. Open-access summary statistics are available for many studies without access approval.

**Main Scientific Question It Helps Answer:** How can I deposit or access individual-level human genotype-phenotype data from US-funded research studies?

**Typical Users:** Researchers depositing NIH-funded human genomic data; researchers requesting access to GWAS and sequencing datasets; NIH program officers; institutional signing officials.

### Example Scientific Questions:

- Where should I deposit GWAS data from an NIH-funded study?
- How do I access individual-level genotype data from the UK Biobank or other large cohorts deposited in dbGaP?
- What phenotypic data is available for the ARIC study in dbGaP?
- How do I apply for controlled access to a dbGaP dataset?

### Example Use Cases:

- Depositing GWAS data from an NIH-funded cardiovascular disease study.
- Accessing individual-level genotype data for a polygenic risk score analysis.
- Downloading phenotypic data from a longitudinal cohort study for secondary analysis.
- Performing a meta-analysis using multiple dbGaP datasets.

**Input Data Accepted:** Genotype files (PLINK, VCF), phenotype files (tab-delimited), pedigree files, study documentation, consent documentation, IRB approval.

**Output Data Provided:** Controlled-access datasets with dbGaP accession numbers (phs for studies, pht for phenotype tables, phv for variables). Open-access summary statistics available without approval. Controlled-access data available after NIH approval.

**Strengths:**

- The primary US repository for human genotype-phenotype data; required by NIH.
- Hosts many landmark GWAS datasets (ARIC, FHS, WHI, UK Biobank subset, etc.).
- Provides open-access summary statistics for many studies without requiring access approval.
- Integrates with NCBI resources (dbSNP, ClinVar, PubMed) for cross-referencing.
- Standardized data access request process through NIH eRA Commons.
- Provides phenotypic data dictionaries and study documentation.
- Supports cloud-based analysis through NIH-approved cloud platforms.

**Limitations:**

- Access approval process can take months; requires institutional signing official approval.
- Data access is restricted to approved research purposes; re-use requires new approval.
- Phenotypic data quality and completeness varies across studies.
- Some older datasets use legacy formats (PLINK binary) that require format conversion.
- Not suitable for non-US studies where EGA or JGA may be more appropriate.
- Cloud-based analysis is preferred for large datasets; local download may be impractical.

**Common Beginner Mistakes:**

- Attempting to download controlled-access data without an approved data access request.
- Confusing open-access summary statistics with individual-level controlled-access data.
- Not obtaining institutional signing official approval before submitting a data access request.
- Assuming all dbGaP data uses the same format — data formats vary by study.
- Not recording the dbGaP accession number and access approval details for reproducibility.

**When to Use It:** Use dbGaP when depositing NIH-funded human genomic data, when accessing US-based GWAS or sequencing datasets, or when NIH data sharing policies require dbGaP deposition. Use dbGaP for accessing open-access summary statistics from published GWAS studies.

**When NOT to Use It:** Do not use dbGaP for non-human data (use SRA), for European cohorts where EGA is required, or for data that can be shared openly. Do not use dbGaP for GWAS summary statistics that are already available in the GWAS Catalog.

**Related Databases or Alternatives:** EGA (European equivalent); JGA (Japanese equivalent); AnVIL (NIH cloud platform for dbGaP data); GWAS Catalog (open-access GWAS summary statistics); SRA (public sequencing archive); ClinVar (clinical variant interpretation).

**How It Connects to Other Resources:** dbGaP integrates with NCBI resources including dbSNP (variant identifiers), ClinVar (clinical significance), PubMed (publications), and SRA (raw sequencing data). dbGaP accession numbers are cited in publications and cross-referenced in GWAS Catalog entries. NIH-approved cloud platforms (Terra, DNAnexus, Seven Bridges) provide cloud-based access to dbGaP data.

**API / FTP / Bulk Download / Programmatic Access:** dbGaP REST API for metadata queries at <https://api.ncbi.nlm.nih.gov/datasets/v2/>. SRA Toolkit for downloading raw sequencing data associated with





dbGaP studies. Cloud-based access through NIH-approved platforms (Terra/AnVIL, DNAnexus, Seven Bridges) after access approval.

**Evidence or Curation Level:** Community-submitted; NIH performs administrative review but not biological curation. Data quality depends on the submitting study.

**Update Status:** Continuously updated with new submissions; actively maintained by NCBI as of 2025.

**Licensing or Access Restrictions:** Controlled access for individual-level data; requires NIH data access approval through eRA Commons. Open access for summary statistics. Data use agreements specify permitted uses and consent restrictions. HIPAA-compliant infrastructure.

**Citation / Recommended Reference:** Mailman MD et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10):1181–1186. doi:10.1038/ng1007-1181

**Beginner-Friendly Explanation:** dbGaP is the US government's secure archive for human genetic data linked to health information. When NIH funds a study that collects DNA from participants, the researchers are required to deposit their data in dbGaP so other scientists can use it for future research. However, because this data is sensitive, you cannot just download it — you need to apply for access, explain your research purpose, and get approval from NIH. Once approved, you can download the data and use it for your approved research. dbGaP also provides summary statistics (like which genetic variants are associated with a disease) that anyone can access without approval.

**Advanced Technical Explanation:** dbGaP implements a tiered access model: open-access data (summary statistics, study documentation) is available without registration; controlled-access data requires an approved Data Access Request (DAR) submitted through NIH eRA Commons. The DAR process involves institutional signing official approval, research use statement, and review by the relevant Data Access Committee (DAC). Approved access is typically granted for 1 year with renewal options. dbGaP uses the NCBI SRA infrastructure for raw sequencing data storage and the dbGaP phenotype data model for clinical variables. Cloud-based analysis is supported through NIH-approved cloud platforms that provide secure compute environments without requiring local data download.

### Practical Workflow Example:

Step 1: Register at <https://www.ncbi.nlm.nih.gov/gap> and link to your eRA Commons account.

Step 2: Search for the study of interest using the dbGaP study browser.

Step 3: Review the study documentation and data access policy.

Step 4: Submit a Data Access Request through eRA Commons, including research use statement and institutional signing official approval.

Step 5: After approval (typically 4–8 weeks), download data using the SRA Toolkit or access through an NIH-approved cloud platform.

Step 6: Record the dbGaP accession number (phs) and approval reference for your methods section.



## Short Index Entries — Category AA

### JGA (Japanese Genotype-phenotype Archive)

**Resource Type:** Controlled-Access Archive

**Domain:** Human genomics / Clinical genomics

**Main Purpose:** Japanese national controlled-access archive for human genotype-phenotype data, operated by DDBJ/NIG. Analogous to EGA (Europe) and dbGaP (US) for Japanese research cohorts.

**Best Used For:** Depositing or accessing individual-level human genomic data from Japanese research studies. Required by many Japanese funders and journals.

**Key Limitation:** Access requires application to the JGA Data Access Committee; process and timelines differ from EGA and dbGaP. Primarily serves Japanese research community.

**Related Resources:** EGA (European equivalent), dbGaP (US equivalent), DDBJ (public sequencing archive)

**Access / Licensing:** Controlled access; requires data access application. Japanese institutional affiliation typically required for access.

**Citation / Documentation:** Kodama Y et al. (2015). The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data. *Nucleic Acids Research*, 43(D1):D18–D22. doi:10.1093/nar/gku1120

### AnVIL (NHGRI Analysis, Visualization, and Informatics Lab-space)

**Resource Type:** Controlled-Access Cloud Platform / Repository

**Domain:** Human genomics / Cloud computing

**Main Purpose:** NIH/NHGRI cloud-based platform for storing, analyzing, and sharing large-scale human genomic data. Provides cloud compute environments (Terra) co-located with controlled-access data from dbGaP and other sources.

**Best Used For:** Large-scale analysis of controlled-access human genomic data without local data download; cloud-based GWAS, WGS, and multi-omics analyses; accessing NHGRI-funded datasets.

**Key Limitation:** Requires NIH data access approval for controlled-access datasets; cloud compute costs may apply; primarily designed for large-scale analyses.

**Related Resources:** dbGaP (data source), Terra (compute platform), NHGRI (funder), BioData Catalyst (NHLBI equivalent)

**Access / Licensing:** Controlled access for most datasets; requires dbGaP approval. Cloud compute costs may apply.

**Citation / Documentation:** Schatz MC et al. (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genomics*, 2(1):100085. doi:10.1016/j.xgen.2021.100085

### DUOS (Data Use Oversight System)

**Resource Type:** Data Access Management Tool / Registry



**Domain:** Human genomics / Data governance

**Main Purpose:** GA4GH-compliant tool for managing data access requests and consent-based data use restrictions. Uses the Data Use Ontology (DUO) to encode consent terms in machine-readable format, enabling automated data access decisions.

**Best Used For:** Encoding consent restrictions in machine-readable format; automating data access decisions; integrating with EGA, AnVIL, and other controlled-access platforms.

**Key Limitation:** Primarily a governance tool, not a data repository. Requires integration with a data repository for actual data storage and access.

**Related Resources:** EGA (integrates DUOS), AnVIL (integrates DUOS), GA4GH Data Use Ontology (DUO)

**Access / Licensing:** Open-source tool; freely available. Institutional deployment required.

**Citation / Documentation:** Woolley JP et al. (2018). Responsible sharing of biomedical data and biospecimens via the 'Automatable Discovery and Access Matrix' (ADA-M). npj Genomic Medicine, 3:17. doi:10.1038/s41525-018-0057-4

## Category AB: Proteomics Repositories and Mass Spectrometry Resources

### Category Overview

Proteomics repositories and mass spectrometry resources form the data infrastructure for experimental proteomics — the large-scale identification and quantification of proteins from biological samples using mass spectrometry (MS) and related technologies. This category is distinct from protein sequence databases (UniProt, RefSeq) and protein structure databases (PDB, AlphaFold DB): proteomics repositories store experimental MS data from specific biological samples, while protein sequence databases store curated protein sequences and annotations.

The ProteomeXchange Consortium is the central coordinating body for proteomics data repositories, providing a unified framework for data submission, accession number assignment, and data dissemination. ProteomeXchange member repositories include PRIDE (EMBL-EBI), MassIVE (UCSD), PeptideAtlas (ISB), jPOST (Japan), iProX (China), and Panorama Public (University of Washington). All ProteomeXchange members assign PXD accession numbers, enabling cross-repository discovery.

### Key distinctions in this category:

- **Protein sequence databases (UniProt, RefSeq):** Store protein sequences and functional annotations. Do not contain experimental MS data.
- **Proteomics repositories (PRIDE, MassIVE):** Store raw MS data files, processed identification results, and quantification data from specific experiments.
- **Protein expression databases (Human Protein Atlas, ProteomicsDB):** Provide pre-computed protein expression profiles derived from proteomics data.

## AB1 — ProteomeXchange Consortium

---

**Official Website URL:** <https://www.proteomexchange.org>

**Resource Type:** Registry / Consortium / Coordinating Body

**Main Biological Domain:** Proteomics / Mass spectrometry

**Short Definition:** ProteomeXchange is an international consortium of proteomics data repositories that provides a unified framework for proteomics data submission, accession number assignment, and data dissemination across member repositories.

**What It Is Used For:** ProteomeXchange is used to submit proteomics datasets to member repositories (PRIDE, MassIVE, PeptideAtlas, jPOST, iProX, Panorama Public) and receive a PXD accession number for citation in publications. It provides a central discovery portal for proteomics datasets across all member repositories.

**What Data It Contains:** ProteomeXchange itself is a coordinating body, not a data repository. It maintains a central metadata registry linking PXD accession numbers to datasets stored in member repositories. The ProteomeXchange website provides a unified search interface for discovering datasets across all member repositories.

**Main Scientific Question It Helps Answer:** Where should I submit my proteomics dataset, and how do I find public proteomics datasets for reanalysis?

**Typical Users:** Proteomics researchers submitting data for publication; researchers searching for public proteomics datasets; journal editors requiring data deposition.

**Example Scientific Questions:**

- Which ProteomeXchange member repository should I use for my DDA proteomics data?
- How do I find all public proteomics datasets for a specific protein or disease?
- What is the PXD accession number for the dataset from this publication?

**Example Use Cases:** Submitting a proteomics dataset to PRIDE and receiving a PXD accession number for a journal submission. Searching for public proteomics datasets related to Alzheimer's disease for meta-analysis. Verifying that a published dataset is publicly available before peer review.

**Input Data Accepted:** Metadata queries; dataset accession numbers (PXD).

**Output Data Provided:** Dataset metadata, links to member repositories, PXD accession numbers.

**Strengths:** Provides a unified discovery interface for proteomics datasets across all major repositories. PXD accession numbers are universally recognized and required by most proteomics journals. Supports multiple data types: DDA, DIA, SRM/MRM, top-down proteomics, cross-linking MS. Integrates with PRIDE, MassIVE, PeptideAtlas, jPOST, iProX, and Panorama Public.

**Limitations:** ProteomeXchange is a coordinating body, not a data repository; actual data is stored in member repositories. Dataset quality and completeness varies across member repositories and submitters. Not all proteomics data types are equally well supported across all member repositories.

**Common Beginner Mistakes:** Trying to download data directly from ProteomeXchange — data is stored in member repositories. Assuming all PXD datasets contain raw data — some datasets contain only processed results. Not checking which member repository hosts a specific PXD dataset before attempting download.

**When to Use It:** Use ProteomeXchange to discover proteomics datasets across all major repositories, to find the correct repository for data submission, and to obtain PXD accession numbers for publication.

**When NOT to Use It:** Do not use ProteomeXchange to download data — go directly to the member repository (PRIDE, MassIVE, etc.).

**Related Databases or Alternatives:** PRIDE (primary member), MassIVE (US member), PeptideAtlas (spectral library member), jPOST (Japanese member), iProX (Chinese member), Panorama Public (DIA member).

**How It Connects to Other Resources:** ProteomeXchange connects to all member repositories and to UniProt (protein identifiers), PubMed (publications), and PRIDE (primary data storage).

**API / FTP / Bulk Download / Programmatic Access:** ProteomeXchange REST API at <https://proteomecentral.proteomexchange.org/api/>. Returns JSON metadata for datasets. Python package proteomexchange-client available.

**Evidence or Curation Level:** Community-submitted; metadata quality depends on submitter; ProteomeXchange performs format validation.

**Update Status:** Continuously updated; actively maintained as of 2025.

**Licensing or Access Restrictions:** Open access for most datasets; some datasets may have embargo periods.

**Citation / Recommended Reference:** Deutsch EW et al. (2023). The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Research, 51(D1):D1539–D1548. doi:10.1093/nar/gkac1040

**Beginner-Friendly Explanation:** ProteomeXchange is like a central registry for proteomics data. When scientists publish a proteomics study, they are required to deposit their data in a public repository and get a PXD accession number (like PXD012345). ProteomeXchange coordinates this process across multiple repositories worldwide, so you can search for proteomics datasets from any of these repositories in one place. Think of it as the INSDC of proteomics — it doesn't store the data itself, but it coordinates the repositories that do.

**Advanced Technical Explanation:** ProteomeXchange implements a distributed data model where member repositories maintain their own data storage and access infrastructure while sharing metadata through the ProteomeXchange central registry. The consortium defines minimum information standards for proteomics data submission (MIAPE — Minimum Information About a Proteomics Experiment) and supports multiple data formats including mzML (raw data), mzIdentML (identification results), mzTab (quantification results), and SDRF-Proteomics (sample metadata). PXD accession numbers follow the format PXD#####.

**Practical Workflow Example:** Step 1: Prepare your proteomics data in standard formats (mzML for raw data, mzIdentML for results). Step 2: Choose a member repository (PRIDE for most DDA/DIA data; Panorama for DIA with Skyline; PASSEL for SRM). Step 3: Submit to the chosen repository and receive a PXD accession number. Step 4: Include the PXD accession number in your manuscript. Step 5: After publication, make the dataset public.

## AB2 — PRIDE (PRoteomics IDentifications Database)

**Official Website URL:** <https://www.ebi.ac.uk/pride>

**Resource Type:** Repository (Proteomics)

**Main Biological Domain:** Proteomics / Mass spectrometry

**Short Definition:** PRIDE is the primary European proteomics data repository, hosted by EMBL-EBI, and the largest ProteomeXchange member repository, storing raw and processed mass spectrometry proteomics data from thousands of published studies.

**What It Is Used For:** PRIDE is used to deposit and access raw mass spectrometry data (vendor formats, mzML), processed identification results (mzIdentML, mzTab), and associated metadata from proteomics experiments. It is the most widely used proteomics repository and is required by most major proteomics journals.

**What Data It Contains:** PRIDE contains raw MS data files (Thermo RAW, Bruker .d, Waters .raw, mzML), processed peptide/protein identification results (mzIdentML, mzTab), quantification data, and sample metadata. As of 2024, PRIDE hosts over 30,000 datasets with over 1 billion spectra. Data types include DDA (data-dependent acquisition), DIA (data-independent acquisition), SRM/MRM, top-down proteomics, cross-linking MS, and phosphoproteomics.

**Main Scientific Question It Helps Answer:** Where can I deposit or access raw mass spectrometry proteomics data from published studies?

**Typical Users:** Proteomics researchers depositing data for publication; researchers reanalyzing public proteomics datasets; bioinformaticians developing proteomics tools.

### Example Scientific Questions:

- Where should I deposit my DDA proteomics data for journal submission?
- How do I download the raw MS data from a published proteomics study?
- What proteomics datasets are available for human liver tissue?
- How do I find all PRIDE datasets for a specific protein?

### 10. Example Use Cases:

- Depositing raw MS data and search results from a phosphoproteomics study.
- Downloading raw data from a published study for reanalysis with a different search engine.
- Finding all public proteomics datasets for a specific cancer type.
- Accessing processed identification results for a meta-analysis of protein expression.

**Input Data Accepted:** Raw MS files (vendor formats, mzML), identification results (mzIdentML, mzTab), quantification data, sample metadata (SDRF-Proteomics format).

**Output Data Provided:** Raw MS files, processed results, metadata, PXD accession numbers, PRIDE Inspector visualizations.

### Strengths:

- Largest proteomics repository; over 30,000 datasets as of 2024.
- Supports all major MS data types and instrument vendors.

- Provides PRIDE Inspector tool for data visualization without downloading.
- Integrates with ProteomeXchange for unified accession numbers.
- Provides PRIDE Converter tools for format conversion.
- Free and open access for most datasets.
- Actively maintained by EMBL-EBI with long-term stability.

**Limitations:**

- Data quality and completeness varies across datasets; not all datasets include raw data.
- Large datasets can be slow to download; Aspera transfer recommended for large files.
- Metadata completeness varies; some older datasets have minimal metadata.
- Not all data types are equally well supported (e.g., some top-down proteomics formats).
- Reanalysis requires careful attention to instrument type, fragmentation method, and search parameters.

**Common Beginner Mistakes:**

- Searching for protein sequences in PRIDE — use UniProt instead.
- Assuming all PRIDE datasets contain raw data — some contain only processed results.
- Reanalyzing data without checking the original publication for search parameters.
- Not accounting for instrument type when reanalyzing data with a different search engine.
- Confusing PRIDE (raw data repository) with ProteomicsDB (expression profiles).

**When to Use It:** Use PRIDE when depositing proteomics data for publication, when accessing raw MS data for reanalysis, or when searching for public proteomics datasets.

**When NOT to Use It:** Do not use PRIDE for protein sequence data (use UniProt), protein structure data (use PDB), or pre-computed protein expression profiles (use Human Protein Atlas or ProteomicsDB).

**Related Databases or Alternatives:** ProteomeXchange (consortium), MassIVE (US equivalent), PeptideAtlas (spectral library), jPOST (Japanese member), iProX (Chinese member), Panorama Public (DIA), UniProt (protein sequences), Human Protein Atlas (protein expression).

**How It Connects to Other Resources:** PRIDE integrates with ProteomeXchange (PXD accession numbers), UniProt (protein identifiers), PubMed (publications), and Ensembl (gene annotations). PRIDE data is reprocessed by PeptideAtlas and ProteomicsDB.

**API / FTP / Bulk Download / Programmatic Access:** PRIDE REST API at <https://www.ebi.ac.uk/pride/ws/archive/v2/>. Returns JSON metadata. PRIDE-R Bioconductor package for R access. Aspera client for large file downloads. FTP at <ftp://ftp.pride.ebi.ac.uk/pride/data/archive/>.

**Evidence or Curation Level:** Community-submitted; PRIDE performs format validation and basic metadata checks but not biological curation.

**Update Status:** Continuously updated; actively maintained by EMBL-EBI as of 2025.

**Licensing or Access Restrictions:** Open access for most datasets; some datasets may have embargo periods (typically 6–12 months after submission).



**Citation / Recommended Reference:** Perez-Riverol Y et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50(D1):D543–D552. doi:10.1093/nar/gkab1038

**Beginner-Friendly Explanation:** PRIDE is the main public archive for proteomics data — the data generated when scientists use mass spectrometry to identify and measure proteins in biological samples. When a proteomics study is published, the raw data files and analysis results are deposited in PRIDE so other researchers can verify the results or reanalyze the data. PRIDE is like GEO for proteomics: it stores the experimental data from published studies. You can search PRIDE for datasets related to your protein or disease of interest and download the data for your own analysis.

**Advanced Technical Explanation:** PRIDE implements the MIAPE (Minimum Information About a Proteomics Experiment) standard for metadata and supports the PSI (Proteomics Standards Initiative) data formats: mzML for raw data, mzIdentML for identification results, and mzTab for quantification results. The PRIDE Converter tool suite enables conversion between vendor formats and standard formats. PRIDE Inspector provides a Java-based GUI for data visualization. The PRIDE REST API supports programmatic access to metadata and file listings. Large datasets are best accessed via Aspera (IBM Aspera Connect) for high-speed transfer. PRIDE data is reprocessed by the PRIDE Reanalysis pipeline and integrated into PeptideAtlas and ProteomicsDB.

#### **Practical Workflow Example:**

Step 1: Prepare raw data files in mzML format (or keep vendor format).

Step 2: Run database search (MaxQuant, Mascot, Sequest) and export results in mzIdentML or mzTab format.

Step 3: Prepare sample metadata in SDRF-Proteomics format.

Step 4: Submit to PRIDE using the PRIDE Submission Tool.

Step 5: Receive PXD accession number.

Step 6: For reanalysis, search PRIDE for datasets of interest, download raw files via FTP or Aspera, and reanalyze with your preferred search engine.

**Reproducibility Notes:** Record the PXD accession number, PRIDE dataset version, and download date. Record the search engine, protein database (UniProt release), search parameters, and FDR threshold used for identification. Note the instrument type and fragmentation method.

**Quality-Control Notes:** Assess raw data quality using PTXQC or iMonDB before reanalysis. Check that the protein database version matches the original study if comparing results. Verify FDR thresholds and decoy strategy used in the original analysis.

## AB3 — PeptideAtlas

**Official Website URL:** <http://www.peptideatlas.org>

**Resource Type:** Database / Spectral Library / Repository

**Main Biological Domain:** Proteomics / Mass spectrometry

**Short Definition:** PeptideAtlas is a multi-organism compendium of peptides identified in tandem mass spectrometry (MS/MS) experiments, providing a uniformly reanalyzed resource of peptide-level evidence for protein existence and expression.

**What It Is Used For:** PeptideAtlas is used to assess the experimental evidence for protein existence at the peptide level, to find reference MS/MS spectra for targeted proteomics assay development, and to access the PASSEL repository for SRM/MRM experiments. It provides a uniformly reprocessed view of public proteomics data.

**What Data It Contains:** PeptideAtlas contains peptide identifications from thousands of reanalyzed public proteomics datasets, organized by organism (human, mouse, yeast, etc.). The Human PeptideAtlas covers a large fraction of the human proteome with peptide-level evidence. PeptideAtlas also hosts PASSEL (PeptideAtlas SRM Experiment Library) for targeted proteomics data and SRMATlas for reference SRM assays.

**Main Scientific Question It Helps Answer:** What is the experimental peptide-level evidence for the existence of this protein, and what are the best peptides for targeted proteomics assays?

**Typical Users:** Proteomics researchers developing targeted assays; researchers assessing protein existence evidence; bioinformaticians integrating proteomics evidence.

### Example Scientific Questions:

- Has this protein been detected by mass spectrometry in human plasma?
- What are the best proteotypic peptides for a targeted SRM assay for this protein?
- What is the peptide-level evidence for this predicted protein?

### Example Use Cases:

- Selecting proteotypic peptides for SRM/MRM assay development.
- Assessing whether a predicted protein has been experimentally detected.
- Accessing reference SRM transitions from SRMATlas.
- Depositing SRM/MRM data in PASSEL.

**Input Data Accepted:** Protein/gene names, peptide sequences, organism names.

**Output Data Provided:** Peptide identifications, spectral evidence, SRM transitions, protein detection status.

### Strengths:

- Uniformly reprocessed data enables fair comparison across datasets.
- Provides peptide-level evidence for protein existence.
- SRMATlas provides reference SRM transitions for targeted proteomics.
- PASSEL provides a repository for SRM/MRM experimental data.
- Integrates with ProteomeXchange for accession numbers.

### Limitations:

- Coverage depends on available public datasets; some proteins may not be detected.
- Reanalysis uses a single search pipeline; results may differ from original publications.
- Not suitable for accessing raw MS data — use PRIDE or MassIVE.
- SRMATlas coverage is primarily for human and a few model organisms.

#### Common Beginner Mistakes:

- Confusing PeptideAtlas (peptide evidence) with PRIDE (raw data repository).
- Assuming absence from PeptideAtlas means a protein does not exist — it may simply not have been detected in available datasets.
- Using PeptideAtlas peptide counts as a measure of protein abundance.

**When to Use It:** Use PeptideAtlas when assessing peptide-level evidence for protein existence, when selecting proteotypic peptides for targeted assays, or when accessing SRM reference transitions.

**When NOT to Use It:** Do not use PeptideAtlas for raw MS data (use PRIDE), protein sequences (use UniProt), or protein expression profiles (use Human Protein Atlas).

**Related Databases or Alternatives:** PRIDE (raw data), MassIVE (raw data), ProteomeXchange (consortium), SRMATlas (SRM transitions), PASSEL (SRM data), UniProt (protein sequences).

**How It Connects to Other Resources:** PeptideAtlas integrates with UniProt (protein identifiers), ProteomeXchange (PXD accessions), and neXtProt (human protein knowledgebase).

**API / FTP / Bulk Download / Programmatic Access:** PeptideAtlas REST API at <http://www.peptideatlas.org/api/>. PASSEL API for SRM data. FTP access for bulk downloads.

**Evidence or Curation Level:** Uniformly reprocessed from public datasets; peptide identifications at 1% FDR.

**Update Status:** Regularly updated with new builds; actively maintained by ISB as of 2025.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Farrah T et al. (2014). State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *Journal of Proteome Research*, 13(1):60–75. doi:10.1021/pr4010037

**Beginner-Friendly Explanation:** PeptideAtlas is a database that collects and reanalyzes proteomics data from many published studies to provide a comprehensive view of which proteins have been experimentally detected by mass spectrometry. For each protein, PeptideAtlas shows which peptides have been identified and in which tissues or conditions. It is particularly useful for targeted proteomics: if you want to measure a specific protein using SRM/MRM mass spectrometry, PeptideAtlas can tell you which peptides are the best targets.

**Advanced Technical Explanation:** PeptideAtlas uses a uniform reanalysis pipeline (ProteinProphet/PeptideProphet) to reprocess public proteomics datasets at a consistent 1% FDR threshold. The resulting peptide identifications are integrated into organism-specific builds that are updated periodically. The canonical protein list is derived from UniProt/Swiss-Prot. SRMATlas provides computationally predicted and experimentally validated SRM transitions for proteotypic peptides, enabling rapid assay development for targeted proteomics.

#### Practical Workflow Example:



Step 1: Navigate to <http://www.peptideatlas.org> and search for your protein of interest.

Step 2: Review the peptide evidence — number of observations, tissues detected, spectral evidence.

Step 3: For targeted assay development, access SRMATlas for reference transitions.

Step 4: For SRM data deposition, use PASSEL.

**Reproducibility Notes:** Record the PeptideAtlas build version and organism used. Note the FDR threshold and search pipeline. For SRMATlas transitions, record the transition set version.

**Quality-Control Notes:** PeptideAtlas uses a uniform 1% FDR threshold, but this does not guarantee all identifications are correct. Check spectral evidence for critical identifications. Absence from PeptideAtlas does not mean a protein is absent — it may simply not have been detected in available datasets.

## AB4 — MassIVE (Mass Spectrometry Interactive Virtual Environment)

---

**Official Website URL:** <https://massive.ucsd.edu>

**Resource Type:** Repository (Proteomics)

**Main Biological Domain:** Proteomics / Mass spectrometry

**Short Definition:** MassIVE is a US-based ProteomeXchange member repository for mass spectrometry data, hosted by UCSD, providing open access to raw MS data and supporting reanalysis through the GNPS platform for metabolomics and natural products.

**What It Is Used For:** MassIVE is used to deposit and access raw mass spectrometry data from proteomics and metabolomics experiments. It is the primary US-based alternative to PRIDE for proteomics data deposition and is the host platform for GNPS (Global Natural Products Social Molecular Networking) for metabolomics.

**What Data It Contains:** MassIVE contains raw MS data files, processed identification results, and metadata from proteomics and metabolomics experiments. It hosts the GNPS spectral library and molecular networking platform for metabolomics. As of 2024, MassIVE hosts tens of thousands of datasets.

**Main Scientific Question It Helps Answer:** Where can I deposit or access US-based proteomics or metabolomics MS data?

**Typical Users:** Proteomics and metabolomics researchers depositing data; researchers reanalyzing public MS datasets; natural products chemists using GNPS.

**Example Scientific Questions:**

- Where should I deposit my proteomics data if I am a US-based researcher?
- How do I access the GNPS molecular networking platform for metabolomics?
- What proteomics datasets are available for human brain tissue in MassIVE?

**Example Use Cases:** Depositing DDA proteomics data from a US-funded study. Accessing raw MS data for reanalysis. Using GNPS for metabolomics spectral matching and molecular networking. Depositing metabolomics data for publication.

**Input Data Accepted:** Raw MS files (vendor formats, mzML), processed results, metadata.

**Output Data Provided:** Raw MS files, processed results, PXD accession numbers, GNPS molecular networks.

**Strengths:** US-based repository; preferred by many US researchers and journals. Hosts GNPS platform for metabolomics molecular networking. Supports both proteomics and metabolomics data. Integrates with ProteomeXchange for PXD accession numbers. Provides ReDU (Repository-scale Data Unification) for cross-dataset analysis.

**Limitations:** Smaller than PRIDE in terms of proteomics dataset count. Interface less polished than PRIDE for proteomics-specific workflows. GNPS is primarily for metabolomics; proteomics researchers may prefer PRIDE.

**Common Beginner Mistakes:** Confusing MassIVE (proteomics/metabolomics repository) with GNPS (metabolomics analysis platform) — GNPS is hosted on MassIVE. Assuming MassIVE and PRIDE contain the same datasets — they are separate repositories.



**When to Use It:** Use MassIVE when depositing proteomics data as a US-based researcher, when using GNPS for metabolomics analysis, or when accessing US-based proteomics datasets.

**When NOT to Use It:** Do not use MassIVE for protein sequences (use UniProt) or protein expression profiles (use Human Protein Atlas). For European proteomics data, PRIDE is more comprehensive.

**Related Databases or Alternatives:** PRIDE (European equivalent), ProteomeXchange (consortium), GNPS (metabolomics platform), PeptideAtlas (spectral library).

**How It Connects to Other Resources:** MassIVE integrates with ProteomeXchange (PXD accessions), GNPS (metabolomics), and UniProt (protein identifiers).

**API / FTP / Bulk Download / Programmatic Access:** MassIVE REST API at <https://massive.ucsd.edu/ProteoSAFe/>. GNPS API for metabolomics. FTP access for bulk downloads.

**Evidence or Curation Level:** Community-submitted; format validation performed.

**Update Status:** Continuously updated; actively maintained by UCSD as of 2025.

**Licensing or Access Restrictions:** Open access for most datasets.

**Citation / Recommended Reference:** Wang M et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, 34(8):828–837. doi:10.1038/nbt.3597

**Beginner-Friendly Explanation:** MassIVE is a public archive for mass spectrometry data, similar to PRIDE but based in the US. It is particularly notable for hosting GNPS, a platform for metabolomics data analysis that allows researchers to identify metabolites by comparing their mass spectra to a reference library. For proteomics, MassIVE serves the same purpose as PRIDE: storing raw data from published studies so other researchers can verify or reanalyze the results.

**Advanced Technical Explanation:** MassIVE implements the ProteomeXchange data model for proteomics submissions and provides PXD accession numbers through the ProteomeXchange consortium. The GNPS platform on MassIVE provides molecular networking (spectral similarity-based clustering), library search, and dereplication tools for metabolomics. MassIVE ReDU (Repository-scale Data Unification) enables cross-dataset analysis of public metabolomics data using standardized metadata.

#### **Practical Workflow Example:**

Step 1: Register at <https://massive.ucsd.edu>.

Step 2: Upload raw MS files and metadata.

Step 3: Submit dataset and receive MSV and PXD accession numbers.

Step 4: For GNPS analysis, upload data to GNPS and run molecular networking.

**Reproducibility Notes:** Record the MassIVE accession number (MSV) and PXD accession number. Record the download date and data version.

**Quality-Control Notes:** Assess raw data quality before reanalysis. Check metadata completeness. For GNPS, verify spectral library version used for annotation.

## Short Index Entries — Category AB

### jPOST (Japan ProteOme STandard Repository/Database)

**Resource Type:** Repository (Proteomics)

**Domain:** Proteomics / Mass spectrometry

**Main Purpose:** Japanese ProteomeXchange member repository for proteomics data, operated by DDBJ/NIG. Provides data deposition and access for Japanese proteomics studies.

**Best Used For:** Depositing proteomics data from Japanese research groups; accessing Japanese proteomics datasets.

**Key Limitation:** Smaller dataset collection than PRIDE; primarily serves Japanese research community.

**Related Resources:** PRIDE (European equivalent), MassIVE (US equivalent), ProteomeXchange (consortium)

**Access / Licensing:** Open access; freely available.

**Citation / Documentation:** Okuda S et al. (2017). jPOSTrepo: an international standard data repository for proteomes. Nucleic Acids Research, 45(D1):D1107–D1111. doi:10.1093/nar/gkw1080

### iProX (Integrated Proteome Resources)

**Resource Type:** Repository (Proteomics)

**Domain:** Proteomics / Mass spectrometry

**Main Purpose:** Chinese ProteomeXchange member repository for proteomics data, operated by the National Center for Protein Sciences Beijing. Provides data deposition and access for Chinese proteomics studies.

**Best Used For:** Depositing proteomics data from Chinese research groups; accessing Chinese proteomics datasets.

**Key Limitation:** Smaller dataset collection than PRIDE; primarily serves Chinese research community.

**Related Resources:** PRIDE (European equivalent), MassIVE (US equivalent), ProteomeXchange (consortium)

**Access / Licensing:** Open access; freely available.

**Citation / Documentation:** Ma J et al. (2019). iProX: an integrated proteome resource. Nucleic Acids Research, 47(D1):D1211–D1217. doi:10.1093/nar/gky869

### Panorama Public

**Resource Type:** Repository (Proteomics — DIA/Targeted)

**Domain:** Proteomics / Mass spectrometry

**Main Purpose:** ProteomeXchange member repository for targeted proteomics (SRM/MRM/PRM) & DIA proteomics data analyzed with Skyline. Provides a repository for Skyline documents & associated raw data.

**Best Used For:** Depositing and accessing targeted proteomics data analyzed with Skyline; DIA proteomics data; SRM/MRM assay libraries.

**Key Limitation:** Primarily designed for Skyline-based workflows; less suitable for DDA proteomics.



**Related Resources:** PASSEL (SRM data), PRIDE (DDA proteomics), ProteomeXchange (consortium), Skyline (analysis tool)

**Access / Licensing:** Open access; freely available.

**Citation / Documentation:** Sharma V et al. (2018). Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Molecular & Cellular Proteomics*, 17(6):1239–1244. doi:10.1074/mcp.RA117.000543

## PASSEL (PeptideAtlas SRM Experiment Library)

---

**Resource Type:** Repository (Proteomics — SRM/MRM)

**Domain:** Proteomics / Targeted mass spectrometry

**Main Purpose:** Repository for SRM (Selected Reaction Monitoring) and MRM (Multiple Reaction Monitoring) proteomics experiments, hosted by PeptideAtlas/ISB. Provides a curated collection of SRM assays and experimental data.

**Best Used For:** Depositing and accessing SRM/MRM proteomics data; finding validated SRM assays for targeted proteomics.

**Key Limitation:** Focused on SRM/MRM data; not suitable for DDA or DIA proteomics.

**Related Resources:** PeptideAtlas (parent resource), SRMATlas (reference transitions), Panorama Public (DIA/targeted), ProteomeXchange (consortium)

**Access / Licensing:** Open access; freely available.

**Citation / Documentation:** Farrah T et al. (2012). PASSEL: the PeptideAtlas SRM experiment library. *Proteomics*, 12(8):1170–1175. doi:10.1002/pmic.201100515

## ProteomicsDB

---

**Resource Type:** Database (Protein Expression)

**Domain:** Proteomics / Human proteome

**Main Purpose:** Human proteome expression database providing protein expression profiles across tissues, cell lines, and body fluids, derived from reanalysis of public proteomics datasets. Provides a resource for exploring the human proteome at the protein level.

**Best Used For:** Exploring protein expression across human tissues and cell lines; finding proteomics evidence for protein expression; drug target analysis.

**Key Limitation:** Derived from reanalyzed public data; expression values are not directly comparable across datasets. Not a primary data repository.

**Related Resources:** Human Protein Atlas (protein expression), PRIDE (raw data), PeptideAtlas (peptide evidence), UniProt (protein sequences)

**Access / Licensing:** Freely accessible for academic use; some features require registration.

**Citation / Documentation:** Wilhelm M et al. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587. doi:10.1038/nature13319

## CPTAC (Clinical Proteomic Tumor Analysis Consortium)

---

**Resource Type:** Dataset Collection / Repository

**Domain:** Cancer proteomics / Clinical genomics

**Main Purpose:** NCI-funded consortium providing comprehensive proteogenomic characterization of human tumors, integrating proteomics, genomics, and clinical data for cancer research.

**Best Used For:** Accessing cancer proteogenomics data; integrating protein expression with genomic alterations in cancer; identifying cancer biomarkers.

**Key Limitation:** Focused on cancer; limited to specific cancer types covered by CPTAC studies. Data access may require registration.

**Related Resources:** TCGA/GDC (genomics), PRIDE (proteomics data), PDC (Proteomic Data Commons), cBioPortal (cancer genomics)

**Access / Licensing:** Open access for most data through the Proteomic Data Commons (PDC, <https://pdc.cancer.gov>); some data requires registration.

**Citation / Documentation:** Gillette MA et al. (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell*, 182(1):200–225. doi:10.1016/j.cell.2020.06.013

## Category AC: Metabolomics and Small-Molecule Omics Resources

### Category Overview

Metabolomics resources encompass a diverse ecosystem of databases, repositories, and spectral libraries that support the identification, annotation, and analysis of small molecules (metabolites, lipids, natural products) in biological systems. This category is distinct from chemical databases (PubChem, ChEMBL) in that it focuses specifically on metabolomics experiments and the biological context of metabolites, rather than chemical synthesis or drug activity.

### Key distinctions in this category:

- Metabolomics repositories (MetaboLights, Metabolomics Workbench): Store raw and processed metabolomics experimental data from specific studies, analogous to GEO for transcriptomics or PRIDE for proteomics.
- Metabolite knowledgebases (HMDB): Provide comprehensive information about metabolites including chemical properties, biological roles, disease associations, and spectral data.
- Spectral libraries (GNPS, MassBank, MoNA): Provide reference MS/MS spectra for metabolite identification by spectral matching.
- Lipid databases (LipidMaps): Provide comprehensive lipid classification, nomenclature, and structural information.
- Chemical ontologies (ChEBI): Provide ontological classification of chemical entities of biological interest.

### Metabolite annotation confidence levels (Metabolomics Standards Initiative):

- Level 1 (Confirmed): Identified by comparison with an authentic reference standard using at least two independent properties (e.g., retention time + MS/MS spectrum).
- Level 2 (Putative annotation): Identified by spectral matching to a library spectrum without a reference standard.
- Level 3 (Putative compound class): Identified as a member of a compound class based on spectral features.
- Level 4 (Unknown): Detected but not identified.

## AC1 — MetaboLights

---

**Official Website URL:** <https://www.ebi.ac.uk/metabolights>

**Resource Type:** Repository (Metabolomics)

**Main Biological Domain:** Metabolomics / Small-molecule omics

**Short Definition:** MetaboLights is the primary European metabolomics data repository, hosted by EMBL-EBI, providing open access to metabolomics experimental data including raw spectra, processed data, and associated metadata.

**What It Is Used For:** MetaboLights is used to deposit and access metabolomics experimental data from LC-MS, GC-MS, NMR, and other metabolomics platforms. It is required by many European journals and funding agencies for metabolomics data deposition.

**What Data It Contains:** MetaboLights contains raw metabolomics data files (mzML, nmrML, vendor formats), processed data (peak tables, metabolite annotations), and metadata following the ISA (Investigation/Study/Assay) framework. As of 2024, MetaboLights hosts over 7,000 studies covering diverse organisms, tissues, and experimental conditions.

**Main Scientific Question It Helps Answer:** Where can I deposit or access metabolomics experimental data from published studies?

**Typical Users:** Metabolomics researchers depositing data for publication; researchers reanalyzing public metabolomics datasets; bioinformaticians developing metabolomics tools.

**Example Scientific Questions:**

- Where should I deposit my LC-MS metabolomics data for journal submission?
- What metabolomics datasets are available for human plasma?
- How do I access the raw data from a published metabolomics study?

**Example Use Cases:**

- Depositing untargeted LC-MS metabolomics data from a disease study.
- Accessing raw data for reanalysis with a different metabolite identification pipeline.
- Finding metabolomics datasets for a specific organism or tissue.

**Input Data Accepted:** Raw MS files (mzML, vendor formats), NMR data (nmrML), processed peak tables, metadata (ISA format).

**Output Data Provided:** Raw data files, processed data, metadata, MTBLS accession numbers.

**Strengths:**

- Primary European metabolomics repository; required by many European journals.
- Supports multiple metabolomics platforms: LC-MS, GC-MS, NMR, CE-MS.
- Uses ISA framework for standardized metadata.
- Integrates with EMBL-EBI resources (ChEBI, Reactome).
- Provides MTBLS accession numbers for citation.
- Actively maintained by EMBL-EBI with long-term stability.

### Limitations:

- Data quality and completeness varies across studies.
- Metadata completeness varies; some older studies have minimal metadata.
- Not all metabolomics platforms are equally well supported.
- Reanalysis requires careful attention to instrument type and data processing parameters.

### Common Beginner Mistakes:

- Confusing MetaboLights (experimental data repository) with HMDB (metabolite knowledgebase).
- Assuming all MetaboLights studies contain raw data — some contain only processed results.
- Not reporting annotation confidence levels for metabolite identifications.

**When to Use It:** Use MetaboLights when depositing metabolomics data for publication (especially for European journals), when accessing raw metabolomics data for reanalysis, or when searching for public metabolomics datasets.

**When NOT to Use It:** Do not use MetaboLights for metabolite chemical information (use HMDB or ChEBI), spectral library matching (use GNPS or MassBank), or lipid classification (use LipidMaps).

**Related Databases or Alternatives:** Metabolomics Workbench (US equivalent), GNPS (spectral library/networking), HMDB (metabolite knowledgebase), MassBank (spectral library), ChEBI (chemical ontology).

**How It Connects to Other Resources:** MetaboLights integrates with ChEBI (metabolite identifiers), Reactome (pathway context), and PubChem (chemical structures). MTBLS accession numbers are cited in publications.

**API / FTP / Bulk Download / Programmatic Access:** MetaboLights REST API at <https://www.ebi.ac.uk/metabolights/ws/>. Returns JSON metadata. FTP at <ftp://ftp.ebi.ac.uk/pub/databases/metabolights/studies/> for bulk downloads.

**Evidence or Curation Level:** Community-submitted; EMBL-EBI performs format validation and metadata checks.

**Update Status:** Continuously updated; actively maintained by EMBL-EBI as of 2025.

**Licensing or Access Restrictions:** Open access for most studies; Creative Commons licenses.

**Citation / Recommended Reference:** Yurekten O et al. (2024). MetaboLights: open data repository for metabolomics. Nucleic Acids Research, 52(D1):D640–D646. doi:10.1093/nar/gkad1045

**Beginner-Friendly Explanation:** MetaboLights is the main public archive for metabolomics data in Europe. Metabolomics is the study of small molecules (metabolites) in biological samples, typically measured by mass spectrometry or NMR. When a metabolomics study is published, the raw data and analysis results are deposited in MetaboLights so other researchers can verify or reanalyze the data. MetaboLights is like GEO for metabolomics: it stores the experimental data from published studies.

**Advanced Technical Explanation:** MetaboLights implements the ISA (Investigation/Study/Assay) framework for metadata, which provides a hierarchical structure for describing metabolomics experiments. The ISA-Tab format is used for metadata submission. MetaboLights supports mzML for LC-MS data and nmrML for NMR data. The MetaboLights REST API provides programmatic access to study metadata and file listings. MetaboLights is a

member of the Metabolomics Society's data standards initiative and follows the Metabolomics Standards Initiative (MSI) reporting guidelines.

**Practical Workflow Example:** Step 1: Prepare raw data files in mzML format (or keep vendor format). Step 2: Prepare metadata in ISA-Tab format using the MetaboLights submission tool. Step 3: Submit to MetaboLights and receive an MTBLS accession number. Step 4: Include the MTBLS accession number in your manuscript. Step 5: After publication, make the study public.

**Reproducibility Notes:** Record the MTBLS accession number and access date. Record the metabolomics platform, instrument type, and data processing parameters. Note the metabolite annotation confidence levels and the spectral library used for annotation.

**Quality-Control Notes:** Assess raw data quality before reanalysis. Check metadata completeness. Verify annotation confidence levels for metabolite identifications. Note batch effects and normalization methods used.

## AC2 — Metabolomics Workbench

---

**Official Website URL:** <https://www.metabolomicsworkbench.org>

**Resource Type:** Repository (Metabolomics)

**Main Biological Domain:** Metabolomics / Small-molecule omics

**Short Definition:** Metabolomics Workbench is the primary US metabolomics data repository, funded by NIH, providing open access to metabolomics experimental data, metabolite reference standards, and analysis tools.

**What It Is Used For:** Metabolomics Workbench is used to deposit and access metabolomics experimental data from NIH-funded and other studies. It provides a repository for raw and processed metabolomics data, a metabolite reference database (RefMet), and analysis tools.

**What Data It Contains:** Metabolomics Workbench contains raw metabolomics data, processed data, and metadata from thousands of studies. It also hosts RefMet (Reference list of Metabolite names), a standardized metabolite nomenclature resource, and provides access to metabolite reference standards.

**Main Scientific Question It Helps Answer:** Where can I deposit or access US-based metabolomics experimental data?

**Typical Users:** Metabolomics researchers depositing NIH-funded data; researchers accessing public metabolomics datasets; researchers using RefMet for metabolite name standardization.

**Example Scientific Questions:**

- Where should I deposit my metabolomics data from an NIH-funded study?
- What is the standardized name for this metabolite in RefMet?
- What metabolomics datasets are available for mouse liver?

**Example Use Cases:**

- Depositing untargeted metabolomics data from an NIH-funded study.
- Using RefMet to standardize metabolite names across datasets.
- Accessing public metabolomics datasets for meta-analysis.

**Input Data Accepted:** Raw MS files, processed data, metadata.

**Output Data Provided:** Raw data files, processed data, metadata, study accession numbers, RefMet standardized names.

**Strengths:** Primary US metabolomics repository; required by NIH for metabolomics data sharing.; Hosts RefMet for standardized metabolite nomenclature.; Provides analysis tools including statistical analysis and pathway analysis.; Integrates with HMDB and other metabolite databases.

**Limitations:** Smaller dataset collection than MetaboLights.; Interface less intuitive than MetaboLights for some workflows.; RefMet coverage is not exhaustive for all metabolites.

**Common Beginner Mistakes:** Confusing Metabolomics Workbench (experimental data repository) with HMDB (metabolite knowledgebase).; Not using RefMet for metabolite name standardization when comparing across datasets.



**When to Use It:** Use Metabolomics Workbench when depositing NIH-funded metabolomics data, when using RefMet for metabolite name standardization, or when accessing US-based metabolomics datasets.

**When NOT to Use It:** Do not use Metabolomics Workbench for metabolite chemical information (use HMDB), spectral library matching (use GNPS), or European metabolomics data (MetaboLights is more comprehensive).

**Related Databases or Alternatives:** MetaboLights (European equivalent), GNPS (spectral library), HMDB (metabolite knowledgebase), RefMet (metabolite nomenclature).

**How It Connects to Other Resources:** Metabolomics Workbench integrates with HMDB (metabolite identifiers), KEGG (pathway context), and PubChem (chemical structures).

**API / FTP / Bulk Download / Programmatic Access:** Metabolomics Workbench REST API at <https://www.metabolomicsworkbench.org/rest/>. Returns JSON metadata and data.

**Evidence or Curation Level:** Community-submitted; NIH performs administrative review.

**Update Status:** Continuously updated; actively maintained by NIH as of 2025.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Sud M et al. (2016). Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44(D1):D463–D470. doi:10.1093/nar/gkv1042

**Beginner-Friendly Explanation:** Metabolomics Workbench is the US government's public archive for metabolomics data. It is similar to MetaboLights but funded by NIH and primarily used by US researchers. In addition to storing experimental data, Metabolomics Workbench provides RefMet, a standardized list of metabolite names that helps researchers use consistent terminology when describing metabolites across different studies.

**Advanced Technical Explanation:** Metabolomics Workbench implements a study-centric data model with standardized metadata fields. The RefMet database provides a standardized nomenclature for metabolites, enabling cross-study comparison and integration. The Metabolomics Workbench REST API provides programmatic access to study metadata, metabolite data, and RefMet mappings.

#### **Practical Workflow Example:**

Step 1: Register at <https://www.metabolomicsworkbench.org>.

Step 2: Prepare raw data and metadata.

Step 3: Submit study and receive accession number.

Step 4: Use RefMet to standardize metabolite names in your analysis.

## AC3 — HMDB (Human Metabolome Database)

**Official Website URL:** <https://hmdb.ca>

**Resource Type:** Knowledgebase (Metabolomics)

**Main Biological Domain:** Metabolomics / Human metabolome

**Short Definition:** HMDB is the most comprehensive curated database of human metabolites, providing chemical, biological, clinical, and spectral information for thousands of metabolites found in the human body.

**What It Is Used For:** HMDB is used to identify metabolites detected in human biological samples, to find chemical and biological information about specific metabolites, to access reference MS/MS spectra for metabolite identification, and to explore disease associations of metabolites.

**What Data It Contains:** HMDB contains information on over 220,000 metabolite entries (as of 2024), including chemical structures (SMILES, InChI), physical properties, biological roles, tissue distribution, disease associations, metabolic pathways, and reference MS/MS and NMR spectra. Metabolites are classified by chemical class, biological origin, and disease association.

**Main Scientific Question It Helps Answer:** What is this metabolite, what are its biological roles, and what diseases is it associated with?

**Typical Users:** Metabolomics researchers; clinical biochemists; pharmacologists; bioinformaticians.

### Example Scientific Questions:

- What is the chemical structure and biological role of this metabolite?
- What diseases are associated with elevated levels of this metabolite?
- What is the reference MS/MS spectrum for this metabolite?
- What metabolites are found in human cerebrospinal fluid?

### Example Use Cases:

- Identifying a metabolite detected in an untargeted metabolomics study.
- Finding reference spectra for metabolite identification by spectral matching.
- Exploring disease associations of metabolites altered in a disease study.
- Accessing metabolite chemical properties for computational analysis.

**Input Data Accepted:** Metabolite names, HMDB IDs, chemical structures (SMILES, InChI), MS/MS spectra.

**Output Data Provided:** Chemical structures, biological information, disease associations, spectral data, pathway context.

### 13. Strengths:

- Most comprehensive human metabolite database; over 220,000 entries.
- Provides chemical, biological, clinical, and spectral information in one resource.
- Includes reference MS/MS and NMR spectra for metabolite identification.
- Classifies metabolites by chemical class, biological origin, and disease association.
- Integrates with KEGG, Reactome, and other pathway databases.
- Freely accessible with bulk download available.

### Limitations:

- Focused on human metabolites; limited coverage of non-human organisms.
- Not all entries have experimental spectral data; some rely on predicted spectra.
- Disease associations vary in evidence quality; some are based on limited studies.
- Coverage of lipids is less comprehensive than LipidMaps.

### Common Beginner Mistakes:

- Confusing HMDB (metabolite knowledgebase) with MetaboLights (experimental data repository).
- Assuming all HMDB spectral data is experimentally validated — some spectra are predicted.
- Using HMDB for non-human metabolomics — coverage is primarily human.
- Not checking the evidence quality for disease associations.

**When to Use It:** Use HMDB when identifying human metabolites, finding chemical and biological information about metabolites, accessing reference spectra for metabolite identification, or exploring disease associations.

**When NOT to Use It:** Do not use HMDB for experimental metabolomics data (use MetaboLights), non-human metabolomics (use other resources), lipid classification (use LipidMaps), or chemical synthesis information (use PubChem or ChEMBL).

**Related Databases or Alternatives:** MetaboLights (experimental data), Metabolomics Workbench (experimental data), GNPS (spectral library), LipidMaps (lipids), KEGG Compound (metabolic context), ChEBI (chemical ontology), PubChem (chemical structures).

**How It Connects to Other Resources:** HMDB integrates with KEGG (pathway context), Reactome (pathway context), PubChem (chemical structures), ChEBI (chemical ontology), and DrugBank.

**API / FTP / Bulk Download / Programmatic Access:** HMDB REST API at <https://hmdb.ca/metabolites.xml> for bulk download. Individual metabolite records accessible via <https://hmdb.ca/metabolites/HMDB0000001>. Python package hmdb available.

**Evidence or Curation Level:** Manually curated from primary literature; evidence quality varies by entry.

**Update Status:** Regularly updated; HMDB 5.0 released 2022; actively maintained by University of Alberta.

**Licensing or Access Restrictions:** Open access; freely available for academic use.

**Citation / Recommended Reference:** Wishart DS et al. (2022). HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Research, 50(D1):D622–D631. doi:10.1093/nar/gkab1062

**Beginner-Friendly Explanation:** HMDB is the most comprehensive database of metabolites found in the human body. For each metabolite, HMDB provides its chemical structure, where it is found in the body, what biological processes it is involved in, what diseases it is associated with, and reference mass spectra that can be used to identify it in metabolomics experiments. If you detect an unknown compound in a metabolomics experiment, HMDB is one of the first places to look for identification.

**Advanced Technical Explanation:** HMDB implements a comprehensive data model covering chemical properties (molecular formula, molecular weight, SMILES, InChI, InChIKey), spectral data (MS/MS spectra at multiple collision energies, NMR spectra), biological data (tissue distribution, biofluid concentrations, metabolic pathways), and clinical data (disease associations, normal concentration ranges). HMDB IDs (HMDB0000001 format) are

widely used as metabolite identifiers in metabolomics workflows. The HMDB spectral library is used by many metabolomics tools for metabolite identification.

**Practical Workflow Example:** Step 1: Detect a metabolite feature in your LC-MS data. Step 2: Search HMDB by exact mass or molecular formula to find candidate metabolites. Step 3: Compare the experimental MS/MS spectrum to HMDB reference spectra. Step 4: Check biological context (tissue distribution, disease associations) to assess plausibility. Step 5: Report the annotation confidence level (Level 1 if confirmed by reference standard, Level 2 if spectral match only).

**Reproducibility Notes:** Record the HMDB version used for metabolite identification. Record the annotation confidence level for each metabolite. Note whether identification was based on reference standard (Level 1) or spectral matching (Level 2).

**Quality-Control Notes:** Check whether HMDB spectral data is experimentally validated or predicted. Verify disease associations against primary literature. For lipids, cross-check with LipidMaps for classification.

## AC4 — GNPS (Global Natural Products Social Molecular Networking)

**Official Website URL:** <https://gnps.ucsd.edu>

**Resource Type:** Tool / Spectral Library / Repository

**Main Biological Domain:** Metabolomics / Natural products / Mass spectrometry

**Short Definition:** GNPS is a web-based platform for mass spectrometry-based metabolomics and natural products research, providing spectral library matching, molecular networking, and a community-curated spectral library for metabolite identification.

**What It Is Used For:** GNPS is used for metabolite identification by spectral library matching, molecular networking (grouping structurally related compounds by spectral similarity), and data deposition for metabolomics studies. It is particularly widely used in natural products chemistry and untargeted metabolomics.

**What Data It Contains:** GNPS contains a community-curated spectral library with over 700,000 reference MS/MS spectra (as of 2024), covering metabolites, natural products, lipids, and drugs. It also hosts public metabolomics datasets deposited through MassIVE.

**Main Scientific Question It Helps Answer:** What metabolites are present in my sample, and how are they structurally related to each other?

**Typical Users:** Metabolomics researchers; natural products chemists; microbiome researchers; pharmacologists.

### Example Scientific Questions:

- What metabolites can be identified in my LC-MS/MS data by spectral matching?
- What structurally related compounds are present in my sample?
- What natural products are produced by this microorganism?

### Example Use Cases:

- Identifying metabolites in an untargeted metabolomics study by spectral library matching.
- Molecular networking to group structurally related compounds in a complex mixture.
- Dereplication of natural products to identify known compounds.
- Depositing metabolomics data in MassIVE through GNPS.

**Input Data Accepted:** MS/MS spectra (mzML, mzXML, MGF formats).

**Output Data Provided:** Spectral library matches, molecular networks, metabolite annotations.

### Strengths:

- Large community-curated spectral library with over 700,000 spectra.
- Molecular networking enables discovery of structurally related compounds.
- Web-based platform; no local installation required.
- Integrates with MassIVE for data deposition.
- Widely used in natural products and microbiome metabolomics.
- Supports multiple analysis workflows (FBMN, IIMN, MolNetEnhancer).

### Limitations:

- Spectral library coverage is incomplete; many metabolites lack reference spectra.

- Spectral matching is sensitive to instrument type and collision energy.
- Molecular networking requires careful parameter selection.
- Not suitable for targeted metabolomics or quantification.
- Web-based platform may be slow for large datasets.

#### Common Beginner Mistakes:

- Assuming all spectral matches are correct — spectral library matching has false positive rates.
- Not reporting annotation confidence levels for GNPS identifications.
- Using GNPS for targeted quantification — it is designed for untargeted identification.
- Not checking the spectral library version used for annotation.

**When to Use It:** Use GNPS for untargeted metabolomics metabolite identification, molecular networking, natural products dereplication, and metabolomics data deposition.

**When NOT to Use It:** Do not use GNPS for targeted quantification, protein identification, or genomics analysis.

**Related Databases or Alternatives:** MassIVE (data repository), MetaboLights (European repository), HMDB (metabolite knowledgebase), MassBank (spectral library), LipidMaps (lipid database).

**How It Connects to Other Resources:** GNPS integrates with MassIVE (data storage), HMDB (metabolite identifiers), and various metabolomics tools (MZmine, XCMS, MS-DIAL).

**API / FTP / Bulk Download / Programmatic Access:** GNPS REST API at <https://gnps.ucsd.edu/ProteoSAFe/>. Python package pygnps available. Workflow submission via web interface.

**Evidence or Curation Level:** Community-curated spectral library; annotation confidence varies by spectral match quality.

**Update Status:** Continuously updated; actively maintained by UCSD as of 2025.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Wang M et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, 34(8):828–837. doi:10.1038/nbt.3597

**Beginner-Friendly Explanation:** GNPS is a web platform for identifying metabolites in mass spectrometry data. When you measure a biological sample by mass spectrometry, you get thousands of spectra — each representing a different molecule. GNPS helps you identify these molecules by comparing your spectra to a large library of reference spectra from known compounds. It also uses a technique called molecular networking to group similar compounds together, which helps you discover structurally related metabolites even if they are not in the reference library.

**Advanced Technical Explanation:** GNPS implements Feature-Based Molecular Networking (FBMN) and classical molecular networking for spectral similarity-based compound grouping. The GNPS spectral library is community-curated and continuously updated. Spectral matching uses cosine similarity with configurable thresholds. GNPS supports multiple downstream analysis tools including MolNetEnhancer (chemical class annotation), CANOPUS (compound class prediction), and SIRIUS (molecular formula prediction). The GNPS workflow system is built on the ProteoSAFe platform.



**Practical Workflow Example:** Step 1: Convert raw MS data to mzML or MGF format. Step 2: Upload to GNPS and run the FBMN workflow. Step 3: Review spectral library matches and molecular network. Step 4: Annotate compound classes using MolNetEnhancer or CANOPUS. Step 5: Report annotation confidence levels.

**Reproducibility Notes:** Record the GNPS workflow version, spectral library version, and analysis parameters. Note the cosine similarity threshold and minimum matched peaks used for spectral matching. Report annotation confidence levels.

**Quality-Control Notes:** Verify spectral matches against the original reference spectra. Check for false positives using decoy spectral libraries. Note instrument type and collision energy for spectral matching.



## Short Index Entries — Category AC

### MassBank

---

**Resource Type:** Spectral Library

**Domain:** Metabolomics / Mass spectrometry

**Main Purpose:** European reference spectral library for mass spectrometry, providing high-quality experimental MS/MS spectra for metabolite identification. Hosted by EMBL-EBI.

**Best Used For:** Metabolite identification by spectral matching; reference spectra for method development.

**Key Limitation:** Smaller spectral library than GNPS; primarily covers small molecules and metabolites.

**Related Resources:** GNPS, MoNA (US spectral library), HMDB (metabolite knowledgebase)

**Access / Licensing:** Open access; freely available.

**Citation / Documentation:** Horai H et al. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714. doi:10.1002/jms.1777

### MoNA (MassBank of North America)

---

**Resource Type:** Spectral Library

**Domain:** Metabolomics / Mass spectrometry

**Main Purpose:** US-based mass spectral library aggregating spectra from multiple sources including HMDB, GNPS, and other libraries. Provides a unified spectral library for metabolite identification.

**Best Used For:** Metabolite identification by spectral matching; accessing aggregated spectral data from multiple sources.

**Key Limitation:** Aggregated library may contain duplicate or inconsistent spectra from different sources.

**Related Resources:** MassBank (European equivalent), GNPS (spectral library), HMDB.

**Access / Licensing:** Open access; freely available.

**Citation / Documentation:** Horai H et al. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714. doi:10.1002/jms.1777 [MoNA is an extension of MassBank; verify current citation]

### LipidMaps

---

**Resource Type:** Database / Knowledgebase

**Domain:** Lipidomics / Metabolomics

**Main Purpose:** Comprehensive lipid classification, nomenclature, and structural database. Provides the LIPID MAPS classification system for lipids, reference structures, and MS/MS spectra for lipidomics.

**Best Used For:** Lipid classification and nomenclature; lipidomics metabolite identification; accessing lipid structural information.

**Key Limitation:** Focused on lipids; not suitable for non-lipid metabolomics.

**Related Resources:** HMDB (human metabolites), GNPS (spectral library), MetaboLights (experimental data)

**Access / Licensing:** Open access; freely available.

**Citation / Documentation:** Fahy E et al. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of Lipid Research*, 50(Suppl):S9–S14. doi:10.1194/jlr.R800095-JLR200

## KEGG Compound

---

**Resource Type:** Database (Chemical/Metabolic)

**Domain:** Metabolomics / Metabolic pathways

**Main Purpose:** Chemical compound database within KEGG, providing structures, properties, and metabolic pathway context for compounds involved in biological reactions. Part of the KEGG LIGAND database.

**Best Used For:** Finding metabolic pathway context for compounds; KEGG pathway analysis; compound-reaction-enzyme linkages.

**Key Limitation:** Focused on metabolically relevant compounds; not comprehensive for all metabolites. KEGG bulk access requires license.

**Related Resources:** KEGG, HMDB (human metabolites), ChEBI (chemical ontology), Reactome (pathways)

**Access / Licensing:** Free web access; bulk download requires KEGG FTP license.

**Citation / Documentation:** Kanehisa M et al. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D587–D592. doi:10.1093/nar/gkac963

## ChEBI (Chemical Entities of Biological Interest)

---

**Resource Type:** Ontology / Database

**Domain:** Chemical biology / Metabolomics / Ontology

**Main Purpose:** Freely available ontology and database of chemical entities of biological interest, providing standardized chemical identifiers, classifications, and relationships for small molecules relevant to biology.

**Best Used For:** Standardized chemical entity annotation; ontological classification of metabolites; cross-database integration using ChEBI IDs.

**Key Limitation:** Focused on biologically relevant chemicals; not comprehensive for all synthetic compounds.

**Related Resources:** HMDB, LipidMaps, PubChem (chemical structures), OBO Foundry (ontology ecosystem)

**Access / Licensing:** Open access; freely available under Creative Commons Attribution 4.0.

**Citation / Documentation:** Hastings J et al. (2016). ChEBI in 2016: Improved services and an enhanced role in the life sciences. *Nucleic Acids Research*, 44(D1):D1214–D1219. doi:10.1093/nar/gkv1031

## Category AD: Reference Gene Annotation and Genome Annotation Resources

### Category Overview

Reference gene annotation resources provide the authoritative gene models, transcript isoforms, and genomic feature annotations used as the foundation for RNA-seq analysis, variant annotation, genome browsing, and comparative genomics. Choosing the correct annotation resource and version is one of the most consequential decisions in a genomics workflow, as different annotation sources can produce substantially different results for the same data.

The major human gene annotation resources are GENCODE, MANE, Ensembl, and RefSeq. These resources are not simply mirrors of each other — they use different transcript models, different evidence sources, and different curation approaches, leading to real differences in gene counts, transcript boundaries, and UTR definitions.

### Key distinctions:

- **GENCODE:** Comprehensive annotation of all human and mouse genes, including all transcript isoforms. The reference annotation for ENCODE and many RNA-seq pipelines. Produced by the GENCODE consortium (Ensembl + Havana).
- **MANE (Matched Annotation from NCBI and EMBL-EBI):** A joint NCBI/Ensembl project providing a single representative transcript per gene (MANE Select) and clinically important transcripts (MANE Plus Clinical). The recommended transcript for clinical variant reporting.
- **Ensembl:** Comprehensive genome annotation for hundreds of species. The primary annotation for the Ensembl genome browser and VEP variant annotation tool.
- **RefSeq:** NCBI's curated reference sequence database. The primary annotation for NCBI-centric workflows and clinical variant annotation.

**WARNING: Do not mix genome annotation versions within the same analysis. Ensembl IDs, RefSeq IDs, and GENCODE annotations may differ for the same gene. Always record the annotation version used. When comparing results across studies, verify that the same annotation version was used.**

## AD1 — GENCODE

**Official Website URL:** <https://www.gencodegenes.org>

**Resource Type:** Database (Gene Annotation)

**Main Biological Domain:** Genomics / Gene annotation / Transcriptomics

**Short Definition:** GENCODE is a comprehensive, high-quality gene annotation project for the human and mouse genomes, produced by the GENCODE consortium (Ensembl + Havana), providing annotation of all protein-coding genes, non-coding RNA genes, and pseudogenes with all transcript isoforms.

**What It Is Used For:** GENCODE is used as the reference gene annotation for RNA-seq alignment and quantification, genome browser visualization, variant annotation, and comparative genomics. It is the standard annotation for ENCODE, GTEx, and many other large-scale genomics projects.

**What Data It Contains:** GENCODE provides GTF/GFF3 annotation files for the human (GRCh38/hg38 and GRCh37/hg19) and mouse (GRCm39/mm39) genomes, including protein-coding genes, lncRNA genes, miRNA genes, pseudogenes, and other non-coding RNA genes. Each gene entry includes all annotated transcript isoforms with exon boundaries, UTR definitions, and evidence codes.

**Main Scientific Question It Helps Answer:** What are the gene models, transcript isoforms, and genomic coordinates for human and mouse genes?

**Typical Users:** Bioinformaticians performing RNA-seq analysis; genome browser users; variant annotation specialists; comparative genomics researchers.

### Example Scientific Questions:

- What are the transcript isoforms of TP53 in the human genome?
- What is the GENCODE annotation for this genomic region?
- How many protein-coding genes are annotated in the human genome?
- What is the difference between GENCODE basic and comprehensive annotation?

### Example Use Cases:

- Using GENCODE GTF as the reference annotation for STAR/HISAT2 RNA-seq alignment.
- Annotating variants with GENCODE gene models using Ensembl VEP.
- Counting reads per gene using featureCounts with GENCODE annotation.
- Comparing lncRNA annotations between GENCODE releases.

**Input Data Accepted:** Genome coordinates (GRCh38, GRCh37, GRCm39), gene names, Ensembl IDs, GENCODE IDs.

**Output Data Provided:** GTF/GFF3 annotation files, FASTA sequences, gene/transcript metadata.

### Strengths:

- Most comprehensive human and mouse gene annotation; includes all transcript isoforms.
- Produced by the GENCODE consortium combining Ensembl automated annotation with Havana manual curation.
- Standard annotation for ENCODE, GTEx, and many large-scale genomics projects.

- Provides both comprehensive and basic (representative) annotation sets.
- Regularly updated with new releases (approximately 2 per year).
- Freely available in standard formats (GTF, GFF3, FASTA).

**Limitations:**

- Human and mouse only; for other species, use Ensembl.
- Comprehensive annotation includes many low-confidence transcripts; use basic set for most analyses.
- Annotation changes between releases can affect reproducibility.
- GENCODE IDs (ENSG, ENST) are the same as Ensembl IDs but may differ in transcript models.
- lncRNA annotation is rapidly evolving; many lncRNAs have uncertain functional significance.

**Common Beginner Mistakes:**

- Using the comprehensive annotation set for RNA-seq quantification — the basic set is recommended for most analyses.
- Mixing GENCODE releases within the same analysis.
- Assuming GENCODE and RefSeq annotations are equivalent — they use different transcript models.
- Not recording the GENCODE release number used in the analysis.
- Confusing GENCODE IDs (ENSG) with gene symbols — they are different identifiers.

**When to Use It:** Use GENCODE when performing RNA-seq analysis on human or mouse data, when using ENCODE or GTEx data, when you need comprehensive isoform annotation, or when the analysis pipeline requires GTF format annotation.

**When NOT to Use It:** Do not use GENCODE for non-human/non-mouse species (use Ensembl), for clinical variant reporting (use MANE), or for NCBI-centric workflows (use RefSeq).

**Related Databases or Alternatives:** Ensembl (parent annotation), MANE (representative transcripts), RefSeq (NCBI annotation), UCSC Table Browser (genome browser access), Ensembl BioMart (data retrieval).

**How It Connects to Other Resources:** GENCODE annotations are the basis for Ensembl gene models. GENCODE IDs are used in GTEx, ENCODE, and many RNA-seq tools. GENCODE integrates with Ensembl VEP for variant annotation.

**API / FTP / Bulk Download / Programmatic Access:** GENCODE FTP at <https://ftp.ebi.ac.uk/pub/databases/genocode/>. GTF and GFF3 files available for all releases. Ensembl REST API provides access to GENCODE annotations programmatically.

**Evidence or Curation Level:** Manually curated (Havana) + computationally predicted (Ensembl); evidence codes indicate curation level for each transcript.

**Update Status:** Approximately 2 releases per year; GENCODE 47 (human) and GENCODE M35 (mouse) as of 2024.

**Licensing or Access Restrictions:** Open access; freely available under Creative Commons Attribution 4.0.

**Citation / Recommended Reference:** Frankish A et al. (2023). GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*, 51(D1):D942–D949. doi:10.1093/nar/gkac1071

**Beginner-Friendly Explanation:** GENCODE is the most comprehensive map of human and mouse genes. It tells you exactly where each gene is located in the genome, what its exons and introns are, and what different versions (isoforms) of each gene exist. When you do RNA-seq analysis, you need a reference annotation to count how many reads map to each gene — GENCODE provides this annotation. It is produced by a team of expert annotators who manually review gene models and combine them with computational predictions.

**Advanced Technical Explanation:** GENCODE is produced by the GENCODE consortium, which combines Ensembl automated annotation (using ab initio gene prediction, protein homology, and RNA-seq evidence) with Havana manual annotation (expert review of gene models). The resulting annotation is classified into biotype categories (protein\_coding, lncRNA, miRNA, pseudogene, etc.) and confidence levels (basic vs. comprehensive). GENCODE IDs are identical to Ensembl IDs (ENSG for genes, ENST for transcripts, ENSP for proteins). The basic annotation set contains one representative transcript per gene and is recommended for most RNA-seq analyses.

#### **Practical Workflow Example:**

Step 1: Download the GENCODE GTF file for the appropriate genome build and release from <https://ftp.ebi.ac.uk/pub/databases/gencode/>.

Step 2: Use the GTF file as the reference annotation for STAR or HISAT2 alignment.

Step 3: Use featureCounts or HTSeq to count reads per gene using the GTF.

Step 4: Record the GENCODE release number in your methods section.

**Reproducibility Notes:** Always record the GENCODE release number (e.g., GENCODE v47 for human GRCh38). Record the genome build (GRCh38 or GRCh37). Note whether the basic or comprehensive annotation set was used. Do not mix GENCODE releases within the same analysis.

**Quality-Control Notes:** Use the basic annotation set for RNA-seq quantification to avoid counting reads to low-confidence transcripts. Check that the genome build matches the alignment reference. Verify that gene IDs are consistent across annotation versions if comparing results.

## AD2 — MANE (Matched Annotation from NCBI and EMBL-EBI)

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/refseq/MANE/>

**Resource Type:** Database (Gene Annotation / Reference Transcripts)

**Main Biological Domain:** Genomics / Gene annotation / Clinical genomics

**Short Definition:** MANE is a joint NCBI/EMBL-EBI project providing a single representative transcript per human protein-coding gene (MANE Select) and clinically important additional transcripts (MANE Plus Clinical), harmonizing Ensembl and RefSeq annotations.

**What It Is Used For:** MANE is used to select a single representative transcript per gene for clinical variant reporting, RNA-seq analysis, and cross-database integration. MANE Select transcripts are the recommended transcripts for clinical variant annotation and reporting.

**What Data It Contains:** MANE provides a list of MANE Select transcripts (one per protein-coding gene, jointly agreed by NCBI and Ensembl) and MANE Plus Clinical transcripts (additional transcripts with clinical importance). Each MANE transcript has both a RefSeq accession (NM\_) and an Ensembl transcript ID (ENST), enabling cross-database integration.

**Main Scientific Question It Helps Answer:** Which single transcript should I use to represent each human gene for clinical variant reporting and cross-database integration?

**Typical Users:** Clinical genomicists; variant annotation specialists; bioinformaticians performing RNA-seq analysis; database developers.

### Example Scientific Questions:

- Which transcript should I use for clinical variant reporting for BRCA1?
- What is the MANE Select transcript for TP53?
- How do I ensure my variant annotation uses the same transcript as clinical databases?

**Example Use Cases:** Selecting the MANE Select transcript for clinical variant annotation. Harmonizing Ensembl and RefSeq transcript IDs for cross-database integration. Using MANE transcripts as the reference for RNA-seq quantification.

**Input Data Accepted:** Gene names, Ensembl IDs, RefSeq accessions.

**Output Data Provided:** MANE Select and MANE Plus Clinical transcript lists with RefSeq and Ensembl IDs.

**Strengths:** Jointly agreed by NCBI and Ensembl; harmonizes two major annotation systems. Provides a single representative transcript per gene, simplifying variant reporting. MANE Plus Clinical adds clinically important transcripts not covered by MANE Select. Enables cross-database integration using both RefSeq and Ensembl IDs. Recommended by ClinGen and clinical variant interpretation guidelines.

**Limitations:** Covers only human protein-coding genes; not applicable to non-coding genes or other species. MANE Select may not be the most biologically relevant transcript for all research contexts. Coverage is not yet complete for all protein-coding genes (ongoing project).

**Common Beginner Mistakes:** Using MANE Select for all analyses without considering whether the selected transcript is appropriate for the specific research question. Confusing MANE Select (one transcript per gene) with



GENCODE comprehensive (all isoforms). Not checking whether MANE coverage is complete for the genes of interest.

**When to Use It:** Use MANE when selecting a representative transcript for clinical variant reporting, when harmonizing Ensembl and RefSeq annotations, or when a single transcript per gene is needed for analysis.

**When NOT to Use It:** Do not use MANE when comprehensive isoform analysis is needed (use GENCODE), for non-human species (use Ensembl), or for non-coding RNA genes.

**Related Databases or Alternatives:** GENCODE (comprehensive annotation), Ensembl (genome annotation), RefSeq (NCBI annotation), ClinGen (clinical validity), ClinVar (clinical variants).

**How It Connects to Other Resources:** MANE integrates Ensembl and RefSeq annotations. MANE transcripts are used in ClinVar, ClinGen, and clinical variant interpretation tools.

**API / FTP / Bulk Download / Programmatic Access:** MANE transcript list available at <https://ftp.ncbi.nlm.nih.gov/refseq/MANE/>. Ensembl REST API provides MANE status for transcripts. NCBI Datasets API provides MANE information.

**Evidence or Curation Level:** Jointly curated by NCBI and Ensembl; high confidence.

**Update Status:** Regularly updated; MANE v1.3 as of 2024; actively maintained.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Morales J et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604(7905):310–315. doi:10.1038/s41586-022-04558-8

**Beginner-Friendly Explanation:** MANE solves a common problem in genomics: Ensembl and RefSeq often have different transcript models for the same gene, making it confusing to know which one to use. MANE provides a single, jointly agreed transcript for each human gene that is recognized by both Ensembl and RefSeq. This is particularly important for clinical genomics, where variant reports need to use a consistent transcript. If you are reporting a variant in a clinical context, use the MANE Select transcript.

**Advanced Technical Explanation:** MANE Select transcripts are chosen based on criteria including: the transcript must be protein-coding, it must be present in both RefSeq and Ensembl with identical CDS, and it should be the most biologically relevant transcript based on expression evidence and literature. MANE Plus Clinical transcripts are added when a non-MANE-Select transcript is clinically important (e.g., used in established clinical databases or guidelines). MANE transcripts have both RefSeq (NM\_) and Ensembl (ENST) accessions, enabling seamless cross-database integration.

### Practical Workflow Example:

Step 1: Download the MANE transcript list from <https://ftp.ncbi.nlm.nih.gov/refseq/MANE/>.

Step 2: For each gene of interest, identify the MANE Select transcript.

Step 3: Use the MANE Select transcript for variant annotation and reporting.

Step 4: Check MANE Plus Clinical for additional clinically important transcripts.

## AD3 — NCBI Assembly

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/assembly>

**Resource Type:** Database (Genome Assembly)

**Main Biological Domain:** Genomics / Genome assembly

**Short Definition:** NCBI Assembly is the primary repository for genome assembly sequences and metadata, providing access to reference genomes, chromosome-level assemblies, and scaffold-level assemblies for thousands of organisms.

**What It Is Used For:** NCBI Assembly is used to access reference genome sequences, to find the current reference assembly for a given organism, and to download genome sequences for alignment, annotation, and comparative genomics.

**What Data It Contains:** NCBI Assembly contains genome assembly sequences (FASTA), assembly metadata (assembly level, N50, contig count), annotation files (GFF3, GTF), and assembly reports for thousands of organisms. Reference assemblies for major model organisms (human GRCh38, mouse GRCm39, etc.) are maintained here.

**Main Scientific Question It Helps Answer:** What is the current reference genome assembly for this organism, and how do I download it?

**Typical Users:** Bioinformaticians performing genome alignment; comparative genomicists; genome annotation specialists.

### Example Scientific Questions:

- What is the current human reference genome assembly?
- How do I download the mouse reference genome?
- What genome assemblies are available for this organism?

### Example Use Cases:

- Downloading the human reference genome (GRCh38) for RNA-seq alignment.
- Finding the current reference assembly for a non-model organism.
- Comparing assembly quality metrics across assemblies.

**Input Data Accepted:** Organism names, taxonomy IDs, assembly accession numbers (GCA\_, GCF\_).

**Output Data Provided:** Genome sequences (FASTA), annotation files (GFF3, GTF), assembly reports.

### Strengths:

- Comprehensive repository for genome assemblies across all organisms.
- Provides both GenBank (GCA\_) and RefSeq (GCF\_) assemblies.
- Includes assembly quality metrics (N50, contig count, assembly level).
- Integrates with NCBI Datasets for programmatic access.

### Limitations:

- Assembly quality varies widely across organisms.
- Not all assemblies have annotation files.



- Reference assembly versions change; always record the assembly accession.

**Common Beginner Mistakes:**

- Using an outdated genome assembly version.
- Not recording the assembly accession number for reproducibility.
- Confusing GCA\_ (GenBank) and GCF\_ (RefSeq) accessions.

**When to Use It:** Use NCBI Assembly when downloading reference genome sequences, when finding the current reference assembly for an organism, or when comparing assembly quality metrics.

**When NOT to Use It:** Do not use NCBI Assembly for gene annotation files — use GENCODE or Ensembl for human/mouse, or NCBI RefSeq for other organisms.

**Related Databases or Alternatives:** NCBI Datasets (programmatic access), Ensembl (genome annotation), GENCODE (human/mouse annotation), RefSeq (reference sequences).

**How It Connects to Other Resources:** NCBI Assembly integrates with NCBI Datasets, RefSeq, & NCBI Gene.

**API / FTP / Bulk Download / Programmatic Access:** NCBI Datasets API at <https://api.ncbi.nlm.nih.gov/datasets/v2/>. Command-line tool: datasets download genome. FTP at <https://ftp.ncbi.nlm.nih.gov/genomes/>.

**Evidence or Curation Level:** Community-submitted; NCBI performs quality checks and assigns assembly levels.

**Update Status:** Continuously updated; actively maintained by NCBI.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Kitts PA et al. (2016). Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Research, 44(D1):D73–D80. doi:10.1093/nar/gkv1226

**Beginner-Friendly Explanation:** NCBI Assembly is where you go to download genome sequences. If you want to align RNA-seq reads to the human genome, you need to download the reference genome sequence first — NCBI Assembly is where you find it. It provides genome sequences for thousands of organisms, from humans to bacteria.

**Advanced Technical Explanation:** NCBI Assembly provides both GenBank (GCA\_) and RefSeq (GCF\_) assemblies. RefSeq assemblies are curated by NCBI and include annotation files; GenBank assemblies are submitted by researchers and may not have annotation. Assembly levels range from contig (lowest quality) to chromosome (highest quality). The NCBI Datasets command-line tool provides programmatic access to genome sequences and annotation files.

**Practical Workflow Example:**

Step 1: Search NCBI Assembly for your organism.

Step 2: Select the reference assembly (GCF\_ accession for RefSeq).

Step 3: Download using NCBI Datasets: datasets download genome accession GCF\_000001405.40.

Step 4: Record the assembly accession and version.

**Reproducibility Notes:** Record the assembly accession number (GCA\_ or GCF\_) and assembly name (e.g., GRCh38.p14). Record the download date.

**Quality-Control Notes:** Check assembly level (chromosome > scaffold > contig). Verify that the assembly matches the annotation version used.

## AD4 — NCBI Datasets

**Official Website URL:** <https://www.ncbi.nlm.nih.gov/datasets>

**Resource Type:** Tool / Portal (Data Access)

**Main Biological Domain:** Genomics / Bioinformatics

**Short Definition:** NCBI Datasets is a modern data access tool and portal providing programmatic and web-based access to NCBI biological data including genome sequences, gene annotations, protein sequences, and taxonomy information.

**What It Is Used For:** NCBI Datasets is used to download genome sequences, gene annotations, protein sequences, and other NCBI data in a standardized, reproducible way. It replaces older NCBI download methods with a modern API and command-line tool.

**What Data It Contains:** NCBI Datasets provide access to genome assemblies, gene annotations, protein sequences, taxonomy information, and other NCBI data. It does not store new data but provides a modern interface to existing NCBI databases.

**Main Scientific Question It Helps Answer:** How do I programmatically download genome sequences, gene annotations, and other NCBI data?

**Typical Users:** Bioinformaticians; computational biologists; researchers building automated pipelines.

**Example Scientific Questions:**

- How do I download the human reference genome and annotation in one command?
- How do I download all RefSeq protein sequences for a given organism?
- How do I get gene information for a list of gene IDs?

**Example Use Cases:**

- Downloading genome + annotation for RNA-seq pipeline setup.
- Bulk downloading protein sequences for a comparative genomics analysis.
- Automating genome data retrieval in a bioinformatics pipeline.

**Input Data Accepted:** Assembly accessions, gene IDs, taxonomy IDs, organism names.

**Output Data Provided:** Genome sequences (FASTA), annotation files (GFF3, GTF), protein sequences, metadata (JSON).

**Strengths:** Modern, standardized API replacing older NCBI download methods; Command-line tool (datasets) enables reproducible, automated downloads; Provides data packages with genome + annotation in one download; Supports multiple output formats.

**Limitations:** Relatively new tool; some older workflows may use FTP directly; API may change; record the tool version for reproducibility.

**Common Beginner Mistakes:** Using older NCBI FTP paths instead of NCBI Datasets for new workflows; Not recording the datasets tool version for reproducibility.

**When to Use It:** Use NCBI Datasets for programmatic download of NCBI genome sequences, annotations, and protein sequences in new workflows.

**When NOT to Use It:** Do not use NCBI Datasets for EBI data (use Ensembl FTP or BioMart), for expression data (use GEO), or for variant data (use dbSNP/ClinVar APIs).

**Related Databases or Alternatives:** NCBI Assembly (genome assemblies), RefSeq (reference sequences), NCBI Gene (gene information).

**How It Connects to Other Resources:** NCBI Datasets provides access to NCBI Assembly, RefSeq, NCBI Gene, and other NCBI databases.

**API / FTP / Bulk Download / Programmatic Access:** REST API at <https://api.ncbi.nlm.nih.gov/datasets/v2/>.  
Command-line tool: `conda install -c conda-forge ncbi-datasets-cli`. Python package `ncbi-datasets-pylib`.

**Evidence or Curation Level:** Tool providing access to NCBI databases; data quality depends on source database.

**Update Status:** Actively maintained by NCBI; regularly updated.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1):D8–D13. doi:10.1093/nar/gkx1095

**Beginner-Friendly Explanation:** NCBI Datasets is a modern tool for downloading data from NCBI. Instead of navigating complex FTP directories, you can use a simple command like `'datasets download genome accession GCF_000001405.40'` to download the human reference genome. It makes it much easier to get the data you need in a reproducible way.

**Advanced Technical Explanation:** NCBI Datasets provides a REST API and command-line tool that wraps access to multiple NCBI databases. The datasets command-line tool supports downloading genome assemblies, gene annotations, protein sequences, and taxonomy data. The API returns JSON metadata and supports filtering by organism, assembly level, and other criteria. The tool generates data packages (ZIP files) containing the requested data in standardized formats.

### **Practical Workflow Example:**

Step 1: Install: `conda install -c conda-forge ncbi-datasets-cli`.

Step 2: Download human genome: `datasets download genome accession GCF_000001405.40 --include genome,gff3`.

Step 3: Unzip and use the FASTA and GFF3 files in your pipeline.

## Short Index Entries — Category AD

### RefSeq Select

**Resource Type:** Database (Reference Sequences)

**Domain:** Genomics / Gene annotation

**Main Purpose:** A subset of RefSeq providing a single representative transcript per gene for human and mice, selected based on expression evidence and conservation. Predecessor to MANE for humans; still used for mice.

**Best Used For:** Selecting a representative transcript for human or mouse genes when MANE is not available or applicable.

**Key Limitation:** Being superseded by MANE for human; coverage and selection criteria differ from MANE.

**Related Resources:** MANE (successor for human), RefSeq (parent database), GENCODE.

**Access / Licensing:** Open access; freely available through NCBI.

### UCSC Table Browser

**Resource Type:** Tool / Database Interface

**Domain:** Genomics / Genome annotation

**Main Purpose:** Web-based tool for querying and downloading UCSC Genome Browser annotation tracks, including gene models, conservation scores, regulatory elements, and custom tracks.

**Best Used For:** Downloading specific genomic annotation tracks; querying genomic regions for annotation data; accessing UCSC-specific tracks not available elsewhere.

**Key Limitation:** Interface can be complex for beginners; some tracks are UCSC-specific and may not match Ensembl/GENCODE annotations exactly.

**Related Resources:** UCSC Genome Browser (parent), Ensembl BioMart (alternative), GENCODE.

**Access / Licensing:** Free academic access; some tracks require registration.

### Ensembl BioMart

**Resource Type:** Tool / Data Retrieval Interface

**Domain:** Genomics / Gene annotation

**Main Purpose:** Web-based and programmatic data mining tool for Ensembl databases, enabling bulk retrieval of gene annotations, sequences, variants, and cross-references across species.

**Best Used For:** Bulk retrieval of gene annotations; ID mapping between Ensembl, RefSeq, UniProt, and other identifiers; cross-species annotation retrieval.

**Key Limitation:** Query complexity can be challenging for beginners; large queries may time out; results depend on Ensembl release version.

**Related Resources:** Ensembl, GENCODE (gene annotation), UCSC Table Browser, biomaRt (R package)

**Access / Licensing:** Free access; biomaRt R/Bioconductor package for programmatic access.

## Category AE: Expression and Proteome Atlases

### Category Overview

Expression and proteome atlases provide pre-computed, integrated views of gene and protein expression across tissues, cell types, developmental stages, and species. Unlike primary expression repositories (GEO, ArrayExpress) that store individual study data, atlases integrate and harmonize data from multiple sources to provide a comprehensive, queryable view of expression patterns.

### Key distinctions:

- **GTEx:** Human tissue-specific mRNA expression from healthy adult donors; eQTL data; the reference for normal human tissue expression.
- **Human Protein Atlas:** Protein-level expression using antibody-based immunohistochemistry and mass spectrometry; subcellular localization; cancer tissue expression.
- **Bgee:** Cross-species expression comparison using anatomical ontologies; integrates data from multiple organisms.
- **recount3/ARCHS4:** Large-scale reprocessed human RNA-seq data; enables meta-analysis across thousands of studies.

**WARNING: Do not compare expression values across atlases without careful consideration of measurement technology (mRNA vs. protein), tissue definitions, normalization methods, and donor characteristics. GTEx measures mRNA in healthy adults; Human Protein Atlas measures protein in diverse tissues including cancer; Bgee integrates data across species and developmental stages.**



## AE1 — Human Protein Atlas

---

**Official Website URL:** <https://www.proteinatlas.org>

**Resource Type:** Database / Knowledgebase (Protein Expression)

**Main Biological Domain:** Proteomics / Transcriptomics / Cell biology

**Short Definition:** The Human Protein Atlas is a comprehensive resource providing protein expression profiles across human tissues, cell types, and cancer types, using antibody-based immunohistochemistry, mass spectrometry, and transcriptomics.

**What It Is Used For:** The Human Protein Atlas is used to explore protein expression across human tissues and cell types, to assess subcellular protein localization, to compare expression in normal versus cancer tissue, and to find antibody-based evidence for protein expression.

**What Data It Contains:** The Human Protein Atlas contains protein expression data for nearly all human protein-coding genes across 44 normal tissue types, 17 cancer types, and multiple cell lines, using immunohistochemistry (IHC), immunofluorescence (IF), and mass spectrometry. It also provides RNA expression data (from GTEx and FANTOM5), single-cell expression data, and blood protein data. The atlas is organized into sections: Tissue Atlas, Cell Atlas, Pathology Atlas, Blood Atlas, Brain Atlas, and Single Cell Type Atlas.

**Main Scientific Question It Helps Answer:** Where is this protein expressed in the human body, and what is its subcellular localization?

**Typical Users:** Cell biologists; cancer researchers; clinical researchers; drug target researchers; bioinformaticians.

**Example Scientific Questions:**

- In which tissues is this protein expressed at the protein level?
- What is the subcellular localization of this protein?
- Is this protein overexpressed in cancer compared to normal tissue?
- What is the blood concentration of this protein?

**Example Use Cases:**

- Validating antibody specificity using HPA antibody data.
- Identifying tissue-specific proteins for biomarker discovery.
- Comparing protein expression in tumor versus normal tissue.
- Exploring subcellular localization for cell biology research.

**Input Data Accepted:** Gene names, protein names, Ensembl IDs, UniProt IDs.

**Output Data Provided:** IHC images, protein expression scores, RNA expression data, subcellular localization data, cancer expression data.

**Strengths:**

- Provides protein-level expression data (not just mRNA) using IHC and mass spectrometry.
- Covers nearly all human protein-coding genes.
- Includes subcellular localization data from immunofluorescence.
- Provides cancer expression data from the Pathology Atlas.



- Integrates RNA and protein expression data.
- Freely accessible with bulk download available.
- Regularly updated with new data and improved antibodies.

**Limitations:**

- IHC-based expression scores are semi-quantitative and antibody-dependent.
- Antibody specificity varies; some antibodies may cross-react with related proteins.
- Not all proteins have mass spectrometry evidence; many rely on IHC only.
- Tissue coverage is not exhaustive; some rare tissues are not included.
- mRNA and protein expression do not always correlate.
- Cancer expression data is from tissue microarrays, not individual patient samples.

**Common Beginner Mistakes:**

- Treating IHC expression scores as quantitative protein abundance measurements.
- Assuming antibody-based evidence is always specific — check antibody validation status.
- Confusing Human Protein Atlas (protein expression) with GTEx (mRNA expression).
- Not checking the antibody validation status before using HPA data.

**When to Use It:** Use Human Protein Atlas when you need protein-level expression data, subcellular localization information, cancer expression data, or antibody-based evidence for protein expression.

**When NOT to Use It:** Do not use Human Protein Atlas for quantitative protein abundance measurements (use mass spectrometry proteomics data), for non-human species (use Bgee or Expression Atlas), or for eQTL data (use GTEx).

**Related Databases or Alternatives:** GTEx (mRNA expression), Bgee (cross-species expression), Expression Atlas (multi-species expression), PRIDE (proteomics data), UniProt (protein sequences).

**How It Connects to Other Resources:** Human Protein Atlas integrates with Ensembl (gene annotations), UniProt (protein identifiers), GTEx (RNA expression), and FANTOM5 (RNA expression). Antibody data links to antibody validation databases.

**API / FTP / Bulk Download / Programmatic Access:** Human Protein Atlas REST API at <https://www.proteinatlas.org/api/>. Returns JSON data. Bulk download at <https://www.proteinatlas.org/about/download>. R package hpar available.

**Evidence or Curation Level:** Antibody-based (IHC, IF) and mass spectrometry evidence; antibody validation status varies by protein.

**Update Status:** Regularly updated; HPA v23 as of 2024; actively maintained by Karolinska Institute.

**Licensing or Access Restrictions:** Open access; Creative Commons Attribution-ShareAlike 3.0 Unported.

**Citation / Recommended Reference:** Uhlen M et al. (2015). Proteomics. Tissue-based map of the human proteome. Science, 347(6220):1260419. doi:10.1126/science.1260419

**Beginner-Friendly Explanation:** The Human Protein Atlas is a comprehensive map of where proteins are found in the human body. For each protein, it shows which tissues it is expressed in, what it looks like inside cells (subcellular localization), and whether it is overexpressed in cancer. The atlas uses antibodies to stain tissue sections

and visualize proteins, providing images that show exactly where each protein is located. It is one of the most useful resources for understanding where a protein is expressed and what it might be doing.

**Advanced Technical Explanation:** The Human Protein Atlas uses a systematic antibody-based approach to map protein expression across human tissues. Antibodies are generated against human proteins and used for IHC on tissue microarrays (TMAs) covering 44 normal tissue types and 17 cancer types. Expression is scored semi-quantitatively (not detected, low, medium, high). The Cell Atlas uses immunofluorescence to map subcellular localization. The Blood Atlas provides plasma protein concentrations from mass spectrometry. The Single Cell Type Atlas integrates single-cell RNA-seq data. All data is integrated with RNA expression data from GTEx and FANTOM5.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://www.proteinatlas.org> and search for your protein.

Step 2: Review the Tissue Atlas for tissue expression patterns.

Step 3: Check the Cell Atlas for subcellular localization.

Step 4: Review the Pathology Atlas for cancer expression.

Step 5: Check antibody validation status before using data in publications.

**Reproducibility Notes:** Record the HPA version used. Note the antibody ID and validation status for each protein. Record whether data is from IHC, IF, or mass spectrometry.

**Quality-Control Notes:** Always check antibody validation status (enhanced validation, supported, approved, uncertain). Proteins with uncertain antibody validation should be interpreted with caution. Cross-check IHC data with RNA expression data for consistency.

## AE2 — Bgee (Gene Expression Evolution Database)

---

**Official Website URL:** <https://www.bgee.org>

**Resource Type:** Database / Knowledgebase (Expression)

**Main Biological Domain:** Transcriptomics / Comparative genomics / Developmental biology

**Short Definition:** Bgee is a database for retrieval and comparison of gene expression patterns across multiple animal species, integrating expression data from RNA-seq, microarray, in situ hybridization, and EST data using anatomical and developmental ontologies for cross-species comparison.

**What It Is Used For:** Bgee is used for cross-species gene expression comparison, for identifying conserved expression patterns, for exploring expression across developmental stages, and for integrating expression data with anatomical ontologies.

**What Data It Contains:** Bgee integrates expression data from RNA-seq, microarray, in situ hybridization, and EST experiments across dozens of animal species, mapped to anatomical structures using Uberon and developmental stages using developmental ontologies. Expression calls (present/absent) are generated using a consistent pipeline across all data types.

**Main Scientific Question It Helps Answer:** Is this gene expressed in this tissue/cell type across species, and how is its expression conserved?

**Typical Users:** Evolutionary biologists; developmental biologists; comparative genomicists; bioinformaticians.

**Example Scientific Questions:**

- Is this gene expressed in the brain across vertebrate species?
- What is the expression pattern of this gene during development?
- Which genes are specifically expressed in this tissue across species?

**Example Use Cases:**

- Comparing expression of a gene of interest across human, mouse, zebrafish, and Drosophila.
- Identifying tissue-specific genes conserved across vertebrates.
- Exploring developmental expression patterns.

**Input Data Accepted:** Gene names, Ensembl IDs, anatomical terms (Uberon), species names.

**Output Data Provided:** Expression calls (present/absent), expression levels, anatomical expression maps, cross-species comparisons.

**Strengths:**

- Cross-species expression comparison using standardized anatomical ontologies.
- Integrates multiple data types (RNA-seq, microarray, ISH, EST).
- Uses Uberon for cross-species anatomical mapping.
- Provides expression calls with consistent quality thresholds.
- Freely accessible with bulk download available.

**Limitations:**

- Expression calls are binary (present/absent) or ranked; not quantitative abundance.

- Coverage varies by species; human and mouse are best covered.
- Cross-species comparison requires careful interpretation of anatomical homology.
- Not suitable for quantitative expression analysis — use GTEx or Expression Atlas.

**Common Beginner Mistakes:**

- Treating Bgee expression calls as quantitative abundance measurements.
- Assuming anatomical homology is straightforward across distantly related species.
- Confusing Bgee (cross-species comparison) with GTEx (human tissue expression).

**When to Use It:** Use Bgee when comparing gene expression across species, when exploring developmental expression patterns, or when integrating expression data with anatomical ontologies.

**When NOT to Use It:** Do not use Bgee for quantitative expression analysis (use GTEx or Expression Atlas), for human-only analysis (use GTEx or Human Protein Atlas), or for single-cell expression (use CellxGene or SCEA).

**Related Databases or Alternatives:** GTEx (human tissue expression), Human Protein Atlas (protein expression), Expression Atlas (multi-species expression), Uberon (anatomical ontology).

**How It Connects to Other Resources:** Bgee integrates with Ensembl (gene annotations), Uberon (anatomical ontology), and Gene Ontology (functional annotations).

**API / FTP / Bulk Download / Programmatic Access:** Bgee REST API at <https://www.bgee.org/api/>. R package BgeeDB available. Bulk download at <https://www.bgee.org/download/>.

**Evidence or Curation Level:** Integrated from multiple data types; expression calls generated using consistent quality thresholds.

**Update Status:** Regularly updated; Bgee 15 as of 2023; actively maintained by SIB Swiss Institute of Bioinformatics.

**Licensing or Access Restrictions:** Open access; Creative Commons Attribution 4.0.

**Citation / Recommended Reference:** Bastian FB et al. (2021). The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research*, 49(D1):D831–D847. doi:10.1093/nar/gkaa793

**Beginner-Friendly Explanation:** Bgee is a database that lets you compare where a gene is expressed across different animal species. For example, you can ask: is this gene expressed in the brain in humans, mice, zebrafish, and flies? Bgee uses standardized anatomical terms to make these comparisons meaningful, even when the anatomy of different species is quite different. It is particularly useful for evolutionary biology and for understanding whether a gene's expression pattern is conserved across species.

**Advanced Technical Explanation:** Bgee integrates expression data from RNA-seq, microarray, in situ hybridization, and EST experiments, mapping all data to anatomical structures using the Uberon cross-species anatomy ontology and developmental stages using species-specific developmental ontologies. Expression calls (present/absent) are generated using a consistent pipeline that accounts for data type-specific quality thresholds. The BgeeDB R package provides programmatic access to expression data and supports integration with other Bioconductor packages.

**Practical Workflow Example:**

Step 1: Navigate to <https://www.bgee.org> and search for your gene.

Step 2: Select the species of interest.

Step 3: Review expression calls across anatomical structures.

Step 4: Use the cross-species comparison tool to compare expression patterns.

Step 5: Download data using the BgeeDB R package for further analysis.

**Reproducibility Notes:** Record the Bgee version used. Note the species and anatomical terms used for comparison. Record the data types included in the analysis.

**Quality-Control Notes:** Check the evidence types supporting expression calls. Verify anatomical homology assumptions for cross-species comparisons. Note that expression calls are binary; for quantitative analysis, use the underlying expression data.

## AE3 — recount3

**Official Website URL:** <https://rna.recount.bio>

**Resource Type:** Dataset Collection / Tool

**Main Biological Domain:** Transcriptomics / RNA-seq

**Short Definition:** recount3 is a large-scale resource providing uniformly processed RNA-seq data from over 750,000 human and mouse samples from public repositories, enabling meta-analysis and cross-study comparison.

**What It Is Used For:** recount3 is used for large-scale meta-analysis of RNA-seq data, for accessing uniformly processed expression data from thousands of public studies, and for cross-study comparison of gene expression.

**What Data It Contains:** recount3 contains uniformly processed RNA-seq data (read counts, coverage) for over 750,000 human and mouse samples from SRA, GTEx, and TCGA, processed using a consistent pipeline (STAR alignment, GENCODE annotation). Data is available at the gene, exon, and junction levels.

**Main Scientific Question It Helps Answer:** How can I access and compare RNA-seq data from thousands of public studies in a uniform format?

**Typical Users:** Bioinformaticians performing meta-analysis; researchers exploring expression patterns across many studies; tool developers.

**Example Scientific Questions:**

- What is the expression of this gene across thousands of public RNA-seq studies?
- How do I perform a meta-analysis of RNA-seq data from multiple studies?
- What is the expression of this gene in GTEx and TCGA samples?

**Example Use Cases:**

- Meta-analysis of gene expression across hundreds of cancer studies.
- Identifying consistently expressed genes across diverse conditions.
- Accessing uniformly processed GTEx and TCGA expression data.

**Input Data Accepted:** Gene names, Ensembl IDs, SRA accessions, study IDs.

**Output Data Provided:** Read counts, coverage data, sample metadata.

**Strengths:** Uniformly processed data enables fair cross-study comparison; Covers over 750,000 samples from diverse studies; Includes GTEx and TCGA data in a consistent format; R/Bioconductor package (recount3) for programmatic access.

**Limitations:** Uniform processing may not be optimal for all studies; Large data volumes require significant storage and compute; Sample metadata quality varies across studies; Not suitable for studies requiring custom alignment parameters.

**Common Beginner Mistakes:** Assuming uniform processing eliminates all batch effects — biological and technical variation remains; Not checking sample metadata quality before including studies in meta-analysis.

**When to Use It:** Use recount3 for large-scale meta-analysis of RNA-seq data, for accessing uniformly processed public expression data, or for cross-study comparison.



**When NOT to Use It:** Do not use recount3 for studies requiring custom alignment parameters, for non-human/non-mouse species, or for single-cell RNA-seq data.

**Related Databases or Alternatives:** GEO (primary data source), SRA (raw data), GTEx (tissue expression), TCGA (cancer expression), ARCHS4 (alternative reprocessed resource).

**How It Connects to Other Resources:** recount3 integrates with SRA (data source), GTEx, TCGA, and Bioconductor packages.

**API / FTP / Bulk Download / Programmatic Access:** R/Bioconductor package recount3. Web interface at <https://rna.recount.bio>.

**Evidence or Curation Level:** Uniformly reprocessed from public data; data quality depends on original studies.

**Update Status:** Regularly updated; actively maintained by Johns Hopkins University.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Zhang Y et al. (2021). Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances*, 7(25):eabf9084. doi:10.1126/sciadv.abf9084 [Note: cite the recount3 paper: Wilks C et al. (2021). recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biology*, 22(1):323. doi:10.1186/s13059-021-02533-6]

**Beginner-Friendly Explanation:** recount3 is a resource that has taken hundreds of thousands of public RNA-seq datasets and processed them all in the same way, so you can compare gene expression across studies without worrying about differences in how the data was processed. It is like having a uniformly processed version of GEO, making it much easier to do large-scale analyses.

**Advanced Technical Explanation:** recount3 uses a uniform processing pipeline (STAR alignment to GRCh38/GRCm38, GENCODE annotation) to generate read counts at the gene, exon, and junction levels for all samples. The recount3 R/Bioconductor package provides programmatic access to the data and integrates with DESeq2, edgeR, and other differential expression tools. recount3 also provides coverage data (BigWig files) for visualization.

### **Practical Workflow Example:**

Step 1: Install the recount3 R package.

Step 2: Use `recount3::available_projects()` to find studies of interest.

Step 3: Download data using `recount3::create_rse_manual()`.

Step 4: Perform differential expression analysis using DESeq2 or edgeR.

## Short Index Entries — Category AE

### Tabula Sapiens

**Resource Type:** Dataset Collection (Single cell)

**Domain:** Single-cell transcriptomics / Human cell atlas

**Main Purpose:** A comprehensive single-cell RNA-seq atlas of human tissues, profiling over 480,000 cells from 24 organs from the same donors, enabling cross-tissue comparison of cell types.

**Best Used For:** Cross-tissue single-cell expression analysis; cell type identification; human cell atlas reference.

**Key Limitation:** Limited to a small number of donors; may not capture population diversity.

**Related Resources:** Human Cell Atlas (broader initiative), CellxGene (data portal), SCEA (single-cell expression atlas)

**Access / Licensing:** Open access; available through CellxGene and figshare.

### ARCHS4 (All RNA-seq and ChIP-seq Sample and Signature Search)

**Resource Type:** Dataset Collection / Tool

**Domain:** Transcriptomics / RNA-seq

**Main Purpose:** Large-scale uniformly processed human and mouse RNA-seq data from GEO, providing gene expression profiles for millions of samples. Enables gene expression queries and signature analysis.

**Best Used For:** Querying gene expression across thousands of public RNA-seq studies; gene signature analysis; co-expression analysis.

**Key Limitation:** Uniform processing may not be optimal for all studies; metadata quality varies.

**Related Resources:** recount3 (alternative reprocessed resource), GEO (primary data source), Expression Atlas (curated expression)

**Access / Licensing:** Open access; freely available at <https://maayanlab.cloud/archs4/>.

### BioGPS

**Resource Type:** Portal / Database (Gene Expression)

**Domain:** Transcriptomics / Gene expression

**Main Purpose:** Gene annotation portal providing gene expression profiles across tissues and cell types from multiple microarray and RNA-seq datasets. Enables comparison of gene expression across multiple datasets.

**Best Used For:** Quick exploration of gene expression across tissues; comparing expression across multiple datasets; gene annotation.

**Key Limitation:** Primarily microarray-based; some datasets may be outdated. Less comprehensive than GTEx or Human Protein Atlas for current data.

**Related Resources:** GTEx (human tissue expression), Human Protein Atlas (protein expression), Expression Atlas (curated expression)

**Access / Licensing:** Open access; freely available at <https://biogps.org>.

## Category AF: GWAS, Rare Disease, and Clinical Variant Interpretation Resources

### Category Overview

---

GWAS, rare disease, and clinical variant interpretation resources form the infrastructure for translating genomic variation into biological and clinical understanding. This category spans resources for genome-wide association studies, Mendelian disease genetics, clinical variant classification, and rare disease phenotyping.

### Critical distinctions:

---

- **GWAS associations (GWAS Catalog):** Statistical associations between common variants and complex traits. Do not imply causality or clinical actionability.
- **Mendelian disease causality (OMIM, ClinGen):** Curated evidence for genes causing Mendelian diseases. High clinical relevance.
- **Clinical variant classification (ClinVar):** Expert-reviewed pathogenicity classifications for specific variants. Directly relevant to clinical interpretation.
- **Rare disease phenotyping (DECIPHER, Orphanet):** Phenotypic characterization of rare disease patients and conditions.

**WARNING: Do not confuse GWAS associations with Mendelian disease causality. A GWAS hit indicates a statistical association between a variant and a trait in a population; it does not mean the variant causes the disease in individual patients. Do not use GWAS associations for clinical variant interpretation without additional evidence.**

## AF1 — NHGRI-EBI GWAS Catalog

**Official Website URL:** <https://www.ebi.ac.uk/gwas>

**Resource Type:** Database / Knowledgebase (GWAS)

**Main Biological Domain:** Human genetics / Complex disease / Genomics

**Short Definition:** The NHGRI-EBI GWAS Catalog is a comprehensive, manually curated catalog of published genome-wide association studies (GWAS), providing standardized data on genetic associations between variants and traits/diseases.

**What It Is Used For:** The GWAS Catalog is used to find published GWAS associations for specific traits or diseases, to identify genetic variants associated with complex traits, to access GWAS summary statistics, and to prioritize candidate genes for functional follow-up.

**What Data It Contains:** The GWAS Catalog contains manually curated data from over 6,000 publications (as of 2024), covering over 500,000 unique SNP-trait associations across thousands of traits and diseases. Data includes variant identifiers (rsIDs), p-values, effect sizes (odds ratios, beta coefficients), sample sizes, ancestry information, and links to full summary statistics where available.

**Main Scientific Question It Helps Answer:** What genetic variants are associated with this trait or disease, and what is the evidence from published GWAS?

**Typical Users:** Human geneticists; epidemiologists; bioinformaticians; clinical researchers; drug target researchers.

### Example Scientific Questions:

- What variants are associated with type 2 diabetes in published GWAS?
- What is the effect size and p-value for this variant-trait association?
- What traits are associated with this genomic region?
- What GWAS have been performed for this disease?

### Example Use Cases:

- Identifying candidate genes for a complex disease from GWAS hits.
- Accessing GWAS summary statistics for polygenic risk score development.
- Performing LD score regression using GWAS summary statistics.
- Prioritizing drug targets based on genetic evidence.

**Input Data Accepted:** Trait names, disease names, rsIDs, genomic coordinates, EFO terms.

**Output Data Provided:** SNP-trait associations, p-values, effect sizes, sample information, links to summary statistics.

### Strengths:

- Comprehensive, manually curated catalog of published GWAS.
- Standardized data format enabling cross-study comparison.
- Provides links to full summary statistics for many studies.
- Uses EFO (Experimental Factor Ontology) for standardized trait annotation.

- Integrates with Ensembl, Open Targets, and other resources.
- Freely accessible with bulk download available.

**Limitations:**

- Covers only published GWAS; unpublished or preprint studies are not included.
- GWAS associations are statistical, not causal — lead SNPs may be in LD with causal variants.
- Ancestry representation is historically biased toward European populations.
- Effect sizes are not directly comparable across studies with different designs.
- Summary statistics are not available for all studies.

**Common Beginner Mistakes:**

- Treating GWAS associations as causal variants — lead SNPs are often in LD with the causal variant.
- Assuming GWAS hits identify causal genes — the associated gene may not be the causal gene.
- Comparing effect sizes across studies without accounting for differences in study design.
- Not checking ancestry information — effect sizes may differ across populations.
- Confusing GWAS associations (common variants, complex traits) with Mendelian disease causality.

**When to Use It:** Use the GWAS Catalog when searching for published GWAS associations for a trait or disease, when accessing GWAS summary statistics, or when prioritizing candidate genes based on genetic evidence.

**When NOT to Use It:** Do not use the GWAS Catalog for clinical variant interpretation (use ClinVar), for Mendelian disease genetics (use OMIM), or for rare variant analysis (use gnomAD, ClinVar).

**Related Databases or Alternatives:** ClinVar (clinical variant interpretation), OMIM (Mendelian disease), gnomAD (population frequencies), Open Targets (drug target evidence), dbGaP (individual-level data), EGA (controlled-access data).

**How It Connects to Other Resources:** GWAS Catalog integrates with Ensembl (variant annotation), Open Targets (target prioritization), dbSNP (rsIDs), and EFO (trait ontology). Summary statistics link to EGA and dbGaP.

**API / FTP / Bulk Download / Programmatic Access:** GWAS Catalog REST API at <https://www.ebi.ac.uk/gwas/rest/api/>. Returns JSON. R package `gwasrapidd` available. Bulk download at <https://www.ebi.ac.uk/gwas/docs/file-downloads>.

**Evidence or Curation Level:** Manually curated from published GWAS; p-value threshold typically  $< 5 \times 10^{-8}$  for genome-wide significance.

**Update Status:** Weekly updates; actively maintained by NHGRI and EMBL-EBI.

**Licensing or Access Restrictions:** Open access; freely available under Creative Commons Attribution 4.0.

**Citation / Recommended Reference:** Sollis E et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985. doi:10.1093/nar/gkac1010

**Beginner-Friendly Explanation:** The GWAS Catalog is a database of published genome-wide association studies (GWAS). A GWAS looks for genetic variants (usually SNPs) that are more common in people with a particular disease or trait than in people without it. The GWAS Catalog collects the results of thousands of such studies, so

you can search for all the genetic variants associated with a disease of interest. It is important to understand that these are statistical associations, not proof that the variant causes the disease.

**Advanced Technical Explanation:** The GWAS Catalog curates GWAS results from published papers, extracting variant-trait associations with p-values below the genome-wide significance threshold (typically  $5 \times 10^{-8}$ ). Data is standardized using EFO (Experimental Factor Ontology) for trait annotation and mapped to GRCh38 coordinates. The catalog provides both top-line results (lead SNPs) and, increasingly, full summary statistics for complete GWAS datasets. The GWAS Catalog REST API supports programmatic access to associations, studies, and summary statistics.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://www.ebi.ac.uk/gwas> and search for your trait of interest.

Step 2: Review the association table for genome-wide significant hits.

Step 3: Download summary statistics if available.

Step 4: Use Ensembl VEP or Open Targets to annotate and prioritize candidate genes.

Step 5: Check gnomAD for population frequency of associated variants.

**Reproducibility Notes:** Record the GWAS Catalog version and access date. Note the p-value threshold used for filtering associations. Record the genome build (GRCh38). Cite the original GWAS publications, not just the GWAS Catalog.

**Quality-Control Notes:** Check ancestry information for each study. Verify that effect sizes are comparable across studies. Check for LD between lead SNPs and potential causal variants. Note that GWAS associations are statistical, not causal.

## AF2 — DECIPHER (Database of genomic variation and Phenotype in Humans using Ensembl Resources)

**Official Website URL:** <https://www.deciphergenomics.org>

**Resource Type:** Database / Knowledgebase (Rare Disease / Clinical Genomics)

**Main Biological Domain:** Clinical genomics / Rare disease / Developmental disorders

**Short Definition:** DECIPHER is a web-based database of genomic variants and associated phenotypes from patients with rare developmental disorders, enabling genotype-phenotype correlation and variant interpretation.

**What It Is Used For:** DECIPHER is used to interpret rare genomic variants (CNVs, SNVs) in patients with developmental disorders, to find patients with similar variants and phenotypes, and to access curated genotype-phenotype correlations for rare diseases.

**What Data It Contains:** DECIPHER contains genomic variants (copy number variants, SNVs, indels) and HPO-coded phenotypes from over 30,000 patients with rare developmental disorders, contributed by clinical genetics centers worldwide. Data includes variant coordinates, inheritance information, clinical features, and diagnostic outcomes.

**Main Scientific Question It Helps Answer:** Have other patients with similar genomic variants been reported, and what phenotypes are associated with this variant?

**Typical Users:** Clinical geneticists; genetic counselors; rare disease researchers; bioinformaticians.

### Example Scientific Questions:

- Have other patients with a deletion in this genomic region been reported?
- What phenotypes are associated with variants in this gene?
- Is this CNV likely pathogenic based on similar cases in DECIPHER?

### Example Use Cases:

- Interpreting a novel CNV in a patient with developmental delay.
- Finding similar cases to support variant pathogenicity.
- Exploring genotype-phenotype correlations for a rare disease.

**Input Data Accepted:** Genomic coordinates, gene names, HPO terms.

**Output Data Provided:** Patient variants, phenotypes, genotype-phenotype correlations, variant interpretation support.

### Strengths:

- Curated clinical data from real patients with rare developmental disorders.
- Uses HPO for standardized phenotype annotation.
- Enables genotype-phenotype correlation for rare variants.
- Integrates with Ensembl for variant annotation.
- Provides variant interpretation support for clinical genetics.

### Limitations:

- Access to individual patient data requires registration and data sharing agreement.



- Coverage is biased toward developmental disorders; not comprehensive for all rare diseases.
- Patient data is contributed voluntarily; coverage varies by institution.
- Not suitable for common variant analysis or GWAS.

#### Common Beginner Mistakes:

- Assuming DECIPHER contains all rare disease patients — coverage is voluntary and biased.
- Not registering for full access — some data requires registration.
- Confusing DECIPHER (rare disease patients) with ClinVar (variant classifications).

**When to Use It:** Use DECIPHER when interpreting rare genomic variants in patients with developmental disorders, when searching for similar cases, or when exploring genotype-phenotype correlations.

**When NOT to Use It:** Do not use DECIPHER for common variant analysis, for non-developmental disorders (use other resources), or as a substitute for ClinVar for variant pathogenicity classification.

**Related Databases or Alternatives:** ClinVar (variant classification), ClinGen (clinical validity), OMIM (Mendelian disease), Orphanet (rare diseases), gnomAD (population frequencies), HPO (phenotype ontology).

**How It Connects to Other Resources:** DECIPHER integrates with Ensembl (variant annotation), HPO (phenotype terms), and ClinVar (variant classifications).

**API / FTP / Bulk Download / Programmatic Access:** DECIPHER REST API available for registered users. Web interface for variant and patient queries.

**Evidence or Curation Level:** Clinical data from patients; curated by clinical genetics centers.

**Update Status:** Continuously updated with new patient data; actively maintained by Wellcome Sanger Institute.

**Licensing or Access Restrictions:** Open access for aggregate data; individual patient data requires registration and data sharing agreement.

**Citation / Recommended Reference:** Firth HV et al. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. American Journal of Human Genetics, 84(4):524–533. doi:10.1016/j.ajhg.2009.03.010

**Beginner-Friendly Explanation:** DECIPHER is a database of patients with rare genetic conditions, particularly developmental disorders. When a clinical genetics center finds an unusual genetic variant in a patient, they can share the variant and the patient's clinical features (anonymized) in DECIPHER. This allows other clinicians and researchers to find similar cases, which helps determine whether a variant is likely to be causing the patient's condition. DECIPHER is particularly useful for rare variants that have not been seen before.

**Advanced Technical Explanation:** DECIPHER implements a federated data model where clinical genetics centers contribute patient data (variants + HPO-coded phenotypes) while maintaining control over data sharing. Aggregate data is publicly accessible; individual patient data requires registration. DECIPHER uses Ensembl for variant annotation and HPO for phenotype standardization. The database supports CNV, SNV, and indel data, with inheritance information and diagnostic outcomes.

#### Practical Workflow Example:

Step 1: Navigate to <https://www.deciphergenomics.org>.

Step 2: Search for your variant by genomic coordinates or gene name.



Step 3: Review similar cases and associated phenotypes.

Step 4: Register for full access to individual patient data if needed.

Step 5: Use DECIPHER data as supporting evidence for variant interpretation.

**Reproducibility Notes:** Record the DECIPHER access date. Note the number of similar cases found and the phenotypes reported. Cite DECIPHER in your methods section.

**Quality-Control Notes:** Check the quality and completeness of phenotype data for similar cases. Note that DECIPHER coverage is voluntary and may not represent all rare disease patients.

## AF3 — ClinGen (Clinical Genome Resource)

---

**Official Website URL:** <https://clinicalgenome.org>

**Resource Type:** Knowledgebase (Clinical Genomics)

**Main Biological Domain:** Clinical genomics / Variant interpretation / Gene-disease validity

**Short Definition:** ClinGen is a NIH-funded resource that defines the clinical relevance of genes and variants for use in precision medicine and research, providing expert-curated gene-disease validity assessments and variant pathogenicity classifications.

**What It Is Used For:** ClinGen is used to assess the clinical validity of gene-disease relationships, to access expert-curated variant pathogenicity classifications, and to find ACMG/AMP variant interpretation guidelines for specific genes and diseases.

**What Data It Contains:** ClinGen provides gene-disease validity classifications (Definitive, Strong, Moderate, Limited, No Known Disease Relationship, Disputed, Refuted) for hundreds of gene-disease pairs, variant pathogenicity classifications from expert panels, and ACMG/AMP variant interpretation guidelines for specific genes.

**Main Scientific Question It Helps Answer:** Is there sufficient clinical evidence to classify this gene as causing this disease, and what is the pathogenicity of this variant?

**Typical Users:** Clinical geneticists; genetic counselors; clinical laboratory scientists; researchers.

**Example Scientific Questions:**

- Is there sufficient evidence that variants in this gene cause this disease?
- What is the ClinGen classification for this gene-disease relationship?
- What ACMG/AMP guidelines apply to variants in this gene?

**Example Use Cases:**

- Assessing gene-disease validity before reporting a variant in a clinical context.
- Finding expert-curated variant classifications for a specific gene.
- Accessing gene-specific ACMG/AMP variant interpretation guidelines.

**Input Data Accepted:** Gene names, disease names, variant identifiers.

**Output Data Provided:** Gene-disease validity classifications, variant pathogenicity classifications, ACMG/AMP guidelines.

**Strengths:** Expert-curated gene-disease validity assessments using standardized framework; Provides ACMG/AMP variant interpretation guidelines for specific genes; Integrates with ClinVar for variant classifications; Freely accessible.

**Limitations:** Coverage is not comprehensive for all genes and diseases; Gene-disease validity assessments are updated periodically; check for current status; Not a substitute for clinical laboratory variant interpretation.

**Common Beginner Mistakes:** Assuming all genes have ClinGen validity assessments — coverage is not comprehensive; Confusing ClinGen (gene-disease validity) with ClinVar (variant classifications).

**When to Use It:** Use ClinGen when assessing gene-disease validity, when accessing expert-curated variant classifications, or when finding gene-specific ACMG/AMP guidelines.

**When NOT to Use It:** Do not use ClinGen as a substitute for clinical laboratory variant interpretation or for genes without ClinGen assessments.

**Related Databases or Alternatives:** ClinVar (variant classifications), OMIM (Mendelian disease), DECIPHER (rare disease patients), gnomAD (population frequencies).

**How It Connects to Other Resources:** ClinGen integrates with ClinVar (variant classifications), OMIM (disease information), and HPO (phenotype terms).

**API / FTP / Bulk Download / Programmatic Access:** ClinGen REST API at <https://search.clinicalgenome.org/kb/api/>. Returns JSON.

**Evidence or Curation Level:** Expert-curated by ClinGen working groups using standardized frameworks.

**Update Status:** Regularly updated; actively maintained by NIH.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Rehm HL et al. (2015). ClinGen — The Clinical Genome Resource. New England Journal of Medicine, 372(23):2235–2242. doi:10.1056/NEJMsrl406261

**Beginner-Friendly Explanation:** ClinGen is a resource that evaluates the evidence for whether a gene causes a specific disease. Not every gene that has been associated with a disease in a publication is actually proven to cause it — ClinGen expert panels review all the evidence and assign a validity classification (Definitive, Strong, Moderate, Limited, etc.). This is important for clinical genetics: before reporting a variant in a gene as potentially disease-causing, you should check whether ClinGen has assessed the gene-disease relationship.

**Advanced Technical Explanation:** ClinGen uses a standardized framework (Strande et al., 2017) to assess gene-disease validity based on genetic evidence (case-level data, segregation, de novo variants), experimental evidence (functional studies, animal models), and replication. Classifications range from Definitive (overwhelming evidence) to Refuted (evidence against causality). ClinGen also coordinates expert variant curation panels (VCEPs) that develop gene-specific ACMG/AMP variant interpretation guidelines.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://clinicalgenome.org> and search for your gene-disease pair.

Step 2: Review the gene-disease validity classification.

Step 3: Check for gene-specific ACMG/AMP guidelines from ClinGen VCEPs.

Step 4: Use ClinGen classifications as evidence in variant interpretation.

## Short Index Entries — Category AF

### LOVD (Leiden Open Variation Database)

---

**Resource Type:** Database (Locus-Specific Variant)

**Domain:** Clinical genomics / Rare disease

**Main Purpose:** Open-source platform for locus-specific variant databases, hosting curated variant data for specific genes from clinical laboratories and research groups worldwide.

**Best Used For:** Finding curated variant data for specific genes; locus-specific variant databases for rare disease genes.

**Key Limitation:** Coverage varies by gene; not all genes have LOVD databases. Quality depends on contributing laboratories.

**Related Resources:** ClinVar (variant classifications), DECIPHER (rare disease patients), ClinGen (gene-disease validity)

**Access / Licensing:** Open access; freely available at <https://www.lovd.nl>.

**Citation / Documentation:** Fokkema IF et al. (2011). LOVD v.2.0: the next generation in gene variant databases. *Human Mutation*, 32(5):557–563. doi:10.1002/humu.21438

### HGMD (Human Gene Mutation Database)

---

**Resource Type:** Database (Disease Mutations) — RESTRICTED

**Domain:** Clinical genomics / Disease mutations

**Main Purpose:** Comprehensive database of published germline mutations in human genes causing or associated with human inherited disease. The most comprehensive catalog of disease-causing mutations.

**Best Used For:** Comprehensive variant literature review; clinical variant interpretation support.

**Key Limitation:** Full database requires institutional subscription; free tier is severely limited. Commercial resource. Not open access.

**Related Resources:** ClinVar (open-access variant classifications), LOVD (open-access locus-specific), ClinGen (gene-disease validity)

**Access / Licensing:** RESTRICTED: Full access requires institutional subscription (Cardiff University/QIAGEN). Free tier provides very limited access.

**Citation / Documentation:** Stenson PD et al. (2020). The Human Gene Mutation Database (HGMD): optimizing its use in a clinical diagnostic or research setting. *Human Genetics*, 139(10):1197–1207. doi:10.1007/s00439-020-02199-3

## CIViC (Clinical Interpretation of Variants in Cancer)

---

**Resource Type:** Knowledgebase (Cancer Variant Interpretation)

**Domain:** Cancer genomics / Clinical variant interpretation

**Main Purpose:** Open-access knowledgebase for clinical interpretation of variants in cancer, providing curated evidence for the clinical significance of somatic variants in cancer treatment and prognosis.

**Best Used For:** Clinical interpretation of somatic cancer variants; identifying actionable mutations; drug-variant associations in cancer.

**Key Limitation:** Coverage is not comprehensive for all cancer variants; curation is community-driven and may be incomplete for rare variants.

**Related Resources:** OncoKB (alternative cancer variant KB), COSMIC (somatic mutation catalog), ClinVar (germline variants), cBioPortal (cancer genomics)

**Access / Licensing:** Open access; freely available at <https://civicdb.org>.

**Citation / Documentation:** Griffith M et al. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174. doi:10.1038/ng.3774

## Orphanet

---

**Resource Type:** Database / Knowledgebase (Rare Disease)

**Domain:** Rare disease / Clinical genomics

**Main Purpose:** European rare disease database providing information on rare diseases including disease definitions, epidemiology, gene-disease relationships, and clinical resources. The primary European rare disease resource.

**Best Used For:** Finding information about rare diseases; gene-disease relationships for rare diseases; rare disease epidemiology; patient organization resources.

**Key Limitation:** Coverage is primarily European; some disease entries may be less comprehensive than OMIM for genetic details.

**Related Resources:** OMIM (Mendelian disease genetics), DECIPHER (rare disease patients), ClinGen (gene-disease validity), HPO (phenotype ontology)

**Access / Licensing:** Open access; freely available at <https://www.orpha.net>.

**Citation / Documentation:** Orphanet Report Series. (2024). Rare diseases collection. Available at <https://www.orpha.net>.

## Category AG: Ontology Ecosystem and Controlled Vocabularies

### Category Overview

Ontologies and controlled vocabularies are the semantic backbone of modern bioinformatics. They provide standardized, machine-readable definitions of biological concepts, enabling data integration, automated reasoning, and interoperability across databases and tools.

### Critical distinctions:

- **Ontology (GO, HPO, MONDO):** A formal representation of knowledge as a set of concepts and relationships. Enables automated reasoning and inference.
- **Controlled vocabulary (MeSH):** A standardized set of terms without formal logical relationships. Enables consistent indexing and retrieval.
- **Ontology repository (OLS, BioPortal):** A portal providing access to multiple ontologies. Not ontology itself.
- **Ontology framework (OBO Foundry):** A set of principles and standards for ontology development. Not ontology itself.

**WARNING: Ontology versions matter critically. The same term may have different definitions or relationships across versions. Always record the ontology version and release date used in any analysis. Use persistent identifiers (CURIEs) rather than term labels, which may change across versions.**



## AG1 — OBO Foundry (Open Biological and Biomedical Ontologies Foundry)

**Official Website URL:** <https://obofoundry.org>

**Resource Type:** Registry / Framework (Ontology)

**Main Biological Domain:** Bioinformatics / Ontology / Data standards

**Short Definition:** The OBO Foundry is a collaborative initiative to develop a family of interoperable ontologies for the biological and biomedical sciences, providing a registry of ontologies that adhere to shared principles of orthogonality, openness, and collaborative development.

**What It Is Used For:** The OBO Foundry is used to discover ontologies for specific biological domains, to find ontologies that adhere to interoperability standards, and to understand the relationships between different biological ontologies.

**What Data It Contains:** The OBO Foundry registry contains metadata for over 200 ontologies covering diverse biological and biomedical domains, including information on ontology scope, license, contact, and download links.

**Main Scientific Question It Helps Answer:** Which ontology should I use for annotating this type of biological data, and does it adhere to interoperability standards?

**Typical Users:** Bioinformaticians; database developers; data curators; ontology developers.

**Example Scientific Questions:**

- Which ontology should I use for annotating cell types?
- Is there an OBO Foundry ontology for chemical entities?
- Which ontologies are recommended for annotating phenotypes?

**Example Use Cases:**

- Selecting an appropriate ontology for data annotation.
- Finding ontologies that are interoperable with GO and HPO.
- Discovering ontologies for a specific biological domain.

**Input Data Accepted:** Domain keywords, ontology names.

**Output Data Provided:** Ontology metadata, download links, scope descriptions.

**Strengths:** Provides a curated registry of interoperable biological ontologies; Enforces shared principles (orthogonality, openness, collaborative development); Enables discovery of domain-specific ontologies; Freely accessible.

**Limitations:** Not all relevant ontologies are OBO Foundry members; Ontology quality varies; OBO Foundry membership does not guarantee completeness; The registry is a discovery tool, not an ontology itself.

**Common Beginner Mistakes:**

- Confusing OBO Foundry (registry/framework) with individual ontologies like GO or HPO.
- Assuming OBO Foundry membership guarantees ontology completeness or accuracy.

**When to Use It:** Use OBO Foundry when selecting an ontology for data annotation, when looking for interoperable ontologies, or when discovering ontologies for a specific domain.

**When NOT to Use It:** Do not use OBO Foundry as a substitute for individual ontologies — it is a registry, not an ontology itself.

**Related Databases or Alternatives:** OLS (ontology lookup service), BioPortal (ontology portal), GO (gene ontology), HPO (human phenotype ontology), MONDO (disease ontology).

**How It Connects to Other Resources:** OBO Foundry ontologies are accessible through OLS and BioPortal. Individual ontologies link to their respective databases.

**API / FTP / Bulk Download / Programmatic Access:** OBO Foundry registry available at <https://obofoundry.org/registry/ontologies.yml>. Individual ontologies accessible through OLS API.

**Evidence or Curation Level:** Registry of ontologies; quality depends on individual ontology development.

**Update Status:** Continuously updated; actively maintained by the OBO Foundry community.

**Licensing or Access Restrictions:** Registry is open access; individual ontologies have their own licenses (most are CC BY 4.0 or CC0).

**Citation / Recommended Reference:** Jackson R et al. (2021). The OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database, 2021:baab069. doi:10.1093/database/baab069

**Beginner-Friendly Explanation:** The OBO Foundry is like a quality-controlled directory of biological ontologies. An ontology is a formal vocabulary that defines biological concepts and their relationships in a machine-readable way. The OBO Foundry ensures that its member ontologies follow shared principles, making them interoperable — meaning they can be used together without conflicts. If you need to annotate your data with standardized terms, the OBO Foundry is a good starting point for finding the right ontology.

**Advanced Technical Explanation:** The OBO Foundry enforces principles including: unique identifiers (CURIEs), open licensing, orthogonality (each ontology covers a distinct domain), collaborative development, and use of OWL or OBO format. These principles enable ontology reuse and cross-ontology reasoning. The Foundry coordinates development of core ontologies (GO, HPO, MONDO, Uberon, ChEBI, etc.) and provides a registry of over 200 member and candidate ontologies.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://obofoundry.org>.

Step 2: Browse or search for ontologies relevant to your domain.

Step 3: Review the ontology scope and license.

Step 4: Download the ontology or access it through OLS.

Step 5: Use the ontology terms for data annotation.

## AG2 — MONDO Disease Ontology

**Official Website URL:** <https://mondo.monarchinitiative.org>

**Resource Type:** Ontology (Disease)

**Main Biological Domain:** Disease / Clinical genomics / Ontology

**Short Definition:** MONDO (Monarch Disease Ontology) is a semi-automatically constructed ontology that merges and harmonizes disease concepts from multiple disease resources (OMIM, Orphanet, DOID, MeSH, ICD, NCIT, and others) into a unified, logically consistent disease classification.

**What It Is Used For:** MONDO is used for standardized disease annotation, for cross-database disease concept mapping, for integrating disease information from multiple sources, and for disease classification in clinical and research contexts.

**What Data It Contains:** MONDO contains over 22,000 disease concepts with cross-references to OMIM, Orphanet, DOID, MeSH, ICD-10, ICD-11, NCIT, and other disease resources. Each concept includes synonyms, definitions, and logical axioms.

**Main Scientific Question It Helps Answer:** What is the standardized ontology term for this disease, and how does it map to other disease classification systems?

**Typical Users:** Bioinformaticians; database developers; clinical informaticists; rare disease researchers.

**Example Scientific Questions:**

- What is the MONDO term for this disease?
- How does this OMIM disease map to Orphanet and ICD-10?
- What are the subtypes of this disease in MONDO?

**Example Use Cases:**

- Standardizing disease annotations across multiple databases.
- Mapping disease concepts between OMIM, Orphanet, and ICD.
- Integrating disease data from multiple sources using MONDO as a common vocabulary.

**Input Data Accepted:** Disease names, OMIM IDs, Orphanet IDs, ICD codes, MONDO IDs.

**Output Data Provided:** MONDO disease terms, cross-references, synonyms, logical axioms.

**Strengths:** Harmonizes disease concepts from multiple major disease resources; Provides cross-references to OMIM, Orphanet, ICD, MeSH, and others; Logically consistent with OWL axioms enabling automated reasoning; Freely accessible and open license; Actively maintained by the Monarch Initiative.

**Limitations:** Semi-automatic construction means some mappings may be incorrect; Coverage of rare diseases may be incomplete; Disease classification is complex; some diseases may be classified differently in different systems.

**Common Beginner Mistakes:** Assuming MONDO mappings are always correct — semi-automatic construction introduces errors; Confusing MONDO (disease ontology) with OMIM (disease genetics database); Not checking for deprecated terms when using older MONDO versions.

**When to Use It:** Use MONDO for standardized disease annotation, for cross-database disease concept mapping, or for integrating disease data from multiple sources.

**When NOT to Use It:** Do not use MONDO as a substitute for OMIM (disease genetics) or Orphanet (rare disease information) — MONDO is an ontology, not a disease information database.

**Related Databases or Alternatives:** OMIM (disease genetics), Orphanet (rare diseases), DOID (Disease Ontology), HPO (phenotype ontology), ICD (disease classification), OLS (ontology lookup).

**How It Connects to Other Resources:** MONDO integrates with OMIM, Orphanet, DOID, MeSH, ICD-10, ICD-11, NCIT, and other disease resources through cross-references.

**API / FTP / Bulk Download / Programmatic Access:** MONDO available through OLS API at <https://www.ebi.ac.uk/ols4/>. Download at <https://mondo.monarchinitiative.org/pages/download/>.

**Evidence or Curation Level:** Semi-automatically constructed from multiple disease resources; manually curated for key concepts.

**Update Status:** Monthly releases; actively maintained by the Monarch Initiative.

**Licensing or Access Restrictions:** CC BY 4.0.

**Citation / Recommended Reference:** Vasilevsky NA et al. (2022). MONDO: Unifying diseases for the world, by the world. medRxiv. doi:10.1101/2022.04.13.22273750

**Beginner-Friendly Explanation:** MONDO is a disease ontology that tries to unify disease names and concepts from many different sources. The problem it solves is that the same disease may have different names and identifiers in OMIM, Orphanet, ICD, and other databases. MONDO maps all these together so that when you say 'type 2 diabetes', everyone knows you mean the same thing, regardless of which database they are using. It is particularly useful when integrating data from multiple sources.

**Advanced Technical Explanation:** MONDO uses a semi-automated pipeline to merge disease concepts from OMIM, Orphanet, DOID, MeSH, ICD-10, ICD-11, NCIT, and other sources, using lexical matching and manual curation to resolve conflicts. The ontology is represented in OWL, enabling automated reasoning. MONDO uses the Monarch Initiative's ontology development infrastructure and follows OBO Foundry principles. Cross-references are maintained as xrefs in the OWL file.

**Practical Workflow Example:** Step 1: Search for your disease in OLS (<https://www.ebi.ac.uk/ols4/>) using MONDO. Step 2: Find the MONDO term and its cross-references. Step 3: Use the MONDO ID (e.g., MONDO:0005148) for standardized annotation. Step 4: Download the MONDO OWL file for programmatic use.

## AG3 — Cell Ontology (CL)

---

**Official Website URL:** <https://cell-ontology.github.io>

**Resource Type:** Ontology (Cell Type)

**Main Biological Domain:** Cell biology / Single-cell genomics / Ontology

**Short Definition:** The Cell Ontology (CL) is an OBO Foundry ontology for the representation of cell types in animals, providing a controlled vocabulary for cell type annotation in single-cell genomics, immunology, and developmental biology.

**What It Is Used For:** The Cell Ontology is used for standardized cell type annotation in single-cell RNA-seq data, for integrating cell type information across datasets, and for automated cell type classification.

**What Data It Contains:** The Cell Ontology contains over 2,500 cell type terms with definitions, synonyms, and logical relationships (is\_a, part\_of, develops\_from). Terms are cross-referenced to other ontologies including Uberon (anatomy), GO (function), and PRO (protein).

**Main Scientific Question It Helps Answer:** What is the standardized ontology term for this cell type, and how does it relate to other cell types?

**Typical Users:** Single-cell genomicists; immunologists; developmental biologists; database developers.

**Example Scientific Questions:**

- What is the CL term for this cell type?
- What are the subtypes of T cells in the Cell Ontology?
- How does this cell type relate to other cell types in the ontology?

**Example Use Cases:**

- Annotating cell types in single-cell RNA-seq data using standardized terms.
- Integrating cell type annotations across multiple single-cell datasets.
- Automated cell type classification using CL terms.

**Input Data Accepted:** Cell type names, CL IDs.

**Output Data Provided:** CL terms, definitions, synonyms, logical relationships.

**Strengths:**

- Standardized cell type vocabulary enabling cross-dataset integration.
- Logically consistent with OWL axioms.
- Widely used in single-cell genomics (Human Cell Atlas, CellxGene).
- Freely accessible and open license.

**Limitations:**

- Coverage of rare or novel cell types may be incomplete.
- Cell type classification is complex; some cell types may be classified differently by different researchers.
- Requires familiarity with ontology concepts for effective use.

**Common Beginner Mistakes:**

- Using free-text cell type labels instead of CL terms — prevents cross-dataset integration.

- Not checking for the most specific applicable CL term.

**When to Use It:** Use the Cell Ontology for standardized cell type annotation in single-cell genomics, for cross-dataset integration, or for automated cell type classification.

**When NOT to Use It:** Do not use the Cell Ontology for non-animal cell types (use other ontologies) or as a substitute for experimental cell type characterization.

**Related Databases or Alternatives:** Uberon (anatomy), GO (gene ontology), HPO (phenotype), Human Cell Atlas (single-cell data), CellxGene (single-cell portal).

**How It Connects to Other Resources:** Cell Ontology integrates with Uberon (anatomical location), GO (cell function), and PRO (protein markers). Used by CellxGene and Human Cell Atlas for cell type annotation.

**API / FTP / Bulk Download / Programmatic Access:** Cell Ontology available through OLS API at <https://www.ebi.ac.uk/ols4/>. Download at <https://cell-ontology.github.io>.

**Evidence or Curation Level:** Expert-curated; follows OBO Foundry principles.

**Update Status:** Regularly updated; actively maintained.

**Licensing or Access Restrictions:** CC BY 4.0.

**Citation / Recommended Reference:** Diehl AD et al. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44. doi:10.1186/s13326-016-0088-7

**Beginner-Friendly Explanation:** The Cell Ontology is a standardized vocabulary for cell types. Instead of each researcher using their own names for cell types (which can vary widely), the Cell Ontology provides agreed-upon terms and identifiers. For example, instead of 'CD4+ T cell', 'helper T cell', or 'T helper cell', you would use the CL term CL:0000492. This standardization is essential for integrating single-cell RNA-seq data across different studies and databases.

**Advanced Technical Explanation:** The Cell Ontology uses OWL to represent cell types with logical axioms that enable automated reasoning. For example, a 'CD4-positive, alpha-beta T cell' is defined as a T cell that expresses CD4 and has an alpha-beta T cell receptor. These logical definitions enable automated cell type classification from marker gene expression data. The CL is a core ontology in the Human Cell Atlas and is used by CellxGene for standardized cell type annotation.

#### **Practical Workflow Example:**

Step 1: Search for your cell type in OLS (<https://www.ebi.ac.uk/ols4/>) using CL.

Step 2: Find the most specific applicable CL term.

Step 3: Use the CL ID (e.g., CL:0000492) for annotation.

Step 4: Use CL terms for cross-dataset integration in tools like Seurat or Scanpy.

## AG4 — ECO (Evidence and Conclusion Ontology)

---

**Official Website URL:** <https://evidenceontology.org>

**Resource Type:** Ontology (Evidence)

**Main Biological Domain:** Bioinformatics / Data annotation / Ontology

**Short Definition:** The Evidence and Conclusion Ontology (ECO) is an ontology for describing types of scientific evidence used to support biological assertions, enabling standardized annotation of evidence types in databases and publications.

**What It Is Used For:** ECO is used to annotate the type of evidence supporting a biological assertion (e.g., experimental evidence, computational prediction, manual curation), enabling filtering and evaluation of database annotations by evidence quality.

**What Data It Contains:** ECO contains over 1,500 evidence types organized in a hierarchy, from broad categories (experimental evidence, computational evidence) to specific types (RNA-seq evidence, protein interaction evidence from co-immunoprecipitation).

**Main Scientific Question It Helps Answer:** What type of evidence supports this biological assertion, and how reliable is it?

**Typical Users:** Database curators; bioinformaticians; researchers evaluating annotation quality.

### Example Scientific Questions:

- What evidence types are used to support GO annotations?
- How do I distinguish experimental from computational annotations?
- What ECO term should I use for this type of evidence?

### Example Use Cases:

- Filtering GO annotations by evidence type (experimental vs. computational).
- Annotating database entries with standardized evidence codes.
- Evaluating the reliability of biological assertions.

**Input Data Accepted:** Evidence type descriptions, ECO IDs.

**Output Data Provided:** ECO terms, definitions, evidence hierarchies.

### Strengths:

- Standardized evidence vocabulary enabling cross-database comparison.
- Widely used in GO, UniProt, and other major databases.
- Enables filtering annotations by evidence quality.
- Freely accessible and open license.

### Limitations:

- Requires familiarity with evidence types for effective use.
- Not all databases use ECO consistently.



**Common Beginner Mistakes:**

- Not filtering GO annotations by evidence type — IEA (inferred from electronic annotation) annotations are less reliable than experimental annotations.
- Confusing ECO (evidence types) with GO (gene function).

**When to Use It:** Use ECO when annotating biological assertions with evidence types, when filtering database annotations by evidence quality, or when evaluating the reliability of biological claims.

**When NOT to Use It:** Do not use ECO as a substitute for the actual evidence — it describes evidence types, not the evidence itself.

**Related Databases or Alternatives:** GO (uses ECO for annotation evidence), UniProt (uses ECO), OBO Foundry (ECO is an OBO ontology).

**How It Connects to Other Resources:** ECO is used by GO, UniProt, and other databases for evidence annotation.

**API / FTP / Bulk Download / Programmatic Access:** ECO available through OLS API. Download at <https://evidenceontology.org>.

**Evidence or Curation Level:** Expert-curated; follows OBO Foundry principles.

**Update Status:** Regularly updated; actively maintained.

**Licensing or Access Restrictions:** CC0 (public domain).

**Citation / Recommended Reference:** Giglio M et al. (2019). ECO, the Evidence and Conclusion Ontology: community standard for evidence information. *Nucleic Acids Research*, 47(D1):D1186–D1194. doi:10.1093/nar/gky1036

**Beginner-Friendly Explanation:** The Evidence and Conclusion Ontology (ECO) provides standardized terms for describing the type of evidence supporting a biological claim. For example, when a database says a gene has a particular function, ECO terms tell you whether that claim is based on direct experimental evidence, computational prediction, or manual curation. This is important because experimental evidence is generally more reliable than computational predictions. In GO annotations, you can filter by ECO terms to focus on experimentally supported annotations.

**Advanced Technical Explanation:** ECO provides a hierarchical classification of evidence types used in biological databases. Key distinctions include: experimental evidence (ECO:0000006) vs. computational evidence (ECO:0000501) vs. author statement (ECO:0000304). In GO annotations, the evidence code IEA (Inferred from Electronic Annotation, ECO:0000501) indicates computational annotation and is generally considered less reliable than experimental codes like IDA (Inferred from Direct Assay, ECO:0000314). ECO is used by GO, UniProt, IntAct, and other major databases.

**Practical Workflow Example:** Step 1: When retrieving GO annotations, check the evidence codes. Step 2: Filter for experimental evidence codes (IDA, IMP, IGI, IEP, IPI) to focus on high-confidence annotations. Step 3: Use ECO terms when annotating your own data.

## AG5 — EDAM (Bioinformatics Operations, Data, Topics, and Formats Ontology)

**Official Website URL:** <https://edamontology.org>

**Resource Type:** Ontology (Bioinformatics Operations/Data)

**Main Biological Domain:** Bioinformatics / Tool annotation / Data standards

**Short Definition:** EDAM is an ontology of bioinformatics operations, data types, topics, and formats, providing a controlled vocabulary for annotating bioinformatics tools, databases, and workflows.

**What It Is Used For:** EDAM is used to annotate bioinformatics tools and databases with standardized terms for their operations, input/output data types, topics, and formats, enabling tool discovery and workflow composition.

**What Data It Contains:** EDAM contains approximately 3,000 terms organized in four main branches: Topic (biological/bioinformatics domain), Operation (computational operation), Data (type of data), and Format (data format).

**Main Scientific Question It Helps Answer:** What standardized terms describe this bioinformatics tool's operations, data types, and formats?

**Typical Users:** Bioinformatics tool developers; workflow developers; database curators; bio.tools contributors.

**Example Scientific Questions:**

- What EDAM terms describe sequence alignment tools?
- What is the EDAM format term for FASTQ?
- What EDAM operation terms apply to variant calling?

**Example Use Cases:**

- Annotating bioinformatics tools in bio.tools with EDAM terms.
- Discovering tools for a specific bioinformatics operation.
- Composing workflows using EDAM-annotated tools.

**Input Data Accepted:** Tool names, operation descriptions, data type names, format names.

**Output Data Provided:** EDAM terms, definitions, hierarchies.

**Strengths:**

- Standardized vocabulary for bioinformatics tool annotation.
- Used by bio.tools and other tool registries.
- Enables tool discovery and workflow composition.
- Freely accessible and open license.

**Limitations:**

- Coverage of new tools and formats may lag behind developments.
- Requires familiarity with EDAM structure for effective use.

**Common Beginner Mistakes:** Confusing EDAM (tool/operation annotation) with GO (gene function) or ECO (evidence types).

**When to Use It:** Use EDAM when annotating bioinformatics tools, when discovering tools for a specific operation, or when composing bioinformatics workflows.



**When NOT to Use It:** Do not use EDAM for biological data annotation (use GO, HPO, or other domain ontologies).

**Related Databases or Alternatives:** bio.tools (uses EDAM for tool annotation), OBO Foundry (EDAM is an OBO ontology), OLS (ontology lookup).

**How It Connects to Other Resources:** EDAM is used by bio.tools, WorkflowHub, and other bioinformatics tool registries.

**API / FTP / Bulk Download / Programmatic Access:** EDAM available through OLS API. Download at <https://edamontology.org>.

**Evidence or Curation Level:** Expert-curated; follows OBO Foundry principles.

**Update Status:** Regularly updated; actively maintained.

**Licensing or Access Restrictions:** CC BY-SA 4.0.

**Citation / Recommended Reference:** Ison J et al. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10):1325–1332. doi:10.1093/bioinformatics/btt113

**Beginner-Friendly Explanation:** EDAM is a vocabulary for describing bioinformatics tools and their capabilities. It provides standardized terms for what a tool does (operation), what data it works with (data type), what domain it belongs to (topic), and what file formats it uses (format). This standardization makes it easier to find the right tool for a task and to build automated workflows. EDAM is used by bio.tools, the main registry of bioinformatics tools.

**Advanced Technical Explanation:** EDAM has four main branches: Topic (e.g., EDAM:topic\_0080 Sequence analysis), Operation (e.g., EDAM:operation\_0292 Sequence alignment), Data (e.g., EDAM:data\_2044 Sequence), and Format (e.g., EDAM:format\_1929 FASTA). These terms are used to annotate tools in bio.tools, enabling semantic search and workflow composition. EDAM is integrated with the Common Workflow Language (CWL) and Galaxy workflow systems.

**Practical Workflow Example:** Step 1: Navigate to <https://bio.tools> and search for tools using EDAM terms. Step 2: Filter by operation, data type, or format. Step 3: Use EDAM terms to annotate your own tools in bio.tools.

**Reproducibility Notes:** Record the EDAM version used for tool annotation.

**Quality-Control Notes:** Check that EDAM terms accurately describe the tool's operations and data types.

## AG6 — OLS (Ontology Lookup Service)

---

**Official Website URL:** <https://www.ebi.ac.uk/ols4>

**Resource Type:** Portal / Tool (Ontology Lookup)

**Main Biological Domain:** Bioinformatics / Ontology / Data standards

**Short Definition:** The Ontology Lookup Service (OLS) is a repository for biomedical ontologies hosted by EMBL-EBI, providing a unified interface for searching, browsing, and accessing over 300 ontologies.

**What It Is Used For:** OLS is used to search for ontology terms, to browse ontology hierarchies, to access ontology metadata, and to programmatically retrieve ontology terms via API.

**What Data It Contains:** OLS hosts over 300 biomedical ontologies including GO, HPO, MONDO, CL, Uberon, ChEBI, ECO, EDAM, and many others. Provides term definitions, synonyms, cross-references, and hierarchical relationships.

**Main Scientific Question It Helps Answer:** What is the ontology term for this concept, and what are its relationships to other terms?

**Typical Users:** Bioinformaticians; database curators; researchers annotating data; tool developers.

**Example Scientific Questions:**

- What is the GO term for this biological process?
- What are the child terms of this HPO term?
- What ontologies are available for annotating cell types?

**Example Use Cases:**

- Looking up ontology terms for data annotation.
- Browsing ontology hierarchies to find the most specific applicable term.
- Programmatically retrieving ontology terms via the OLS API.

**Input Data Accepted:** Term names, ontology IDs, CURIEs.

**Output Data Provided:** Term definitions, synonyms, cross-references, hierarchical relationships.

**Strengths:**

- Unified interface for over 300 biomedical ontologies.
- Powerful search and browse functionality.
- REST API for programmatic access.
- Hosted by EMBL-EBI with high availability.
- Freely accessible.

**Limitations:**

- Not all ontologies are hosted in OLS.
- OLS is a lookup service, not an ontology itself.
- API response times may vary for large queries.

**Common Beginner Mistakes:**

- Confusing OLS (lookup service) with individual ontologies like GO or HPO.

- Not using the OLS API for programmatic access — manual lookup is inefficient for large datasets.

**When to Use It:** Use OLS when searching for ontology terms, when browsing ontology hierarchies, or when programmatically accessing ontology data.

**When NOT to Use It:** Do not use OLS as a substitute for individual ontologies — it is a lookup service.

**Related Databases or Alternatives:** BioPortal (alternative ontology portal), OBO Foundry (ontology registry), individual ontologies (GO, HPO, MONDO, etc.).

**How It Connects to Other Resources:** OLS hosts ontologies from OBO Foundry and other sources. Integrates with Ensembl, UniProt, and other EMBL-EBI resources.

**API / FTP / Bulk Download / Programmatic Access:** OLS REST API at <https://www.ebi.ac.uk/ols4/api/>. Returns JSON. R package `rols` available.

**Evidence or Curation Level:** Lookup service; quality depends on individual ontologies.

**Update Status:** Regularly updated; OLS4 is the current version; actively maintained by EMBL-EBI.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Jupp S et al. (2015). A new Ontology Lookup Service at EMBL-EBI. SWAT4LS. Available at <https://www.ebi.ac.uk/ols4>.

**Beginner-Friendly Explanation:** The Ontology Lookup Service (OLS) is a website where you can search for terms from hundreds of biological ontologies in one place. Instead of going to the GO website for gene function terms, the HPO website for phenotype terms, and the MONDO website for disease terms separately, you can search all of them at once in OLS. It is the most convenient way to find the right ontology term for your data.

**Advanced Technical Explanation:** OLS4 (the current version) provides a unified REST API for accessing over 300 ontologies. The API supports term search, hierarchy traversal, cross-reference lookup, and ontology metadata retrieval. The `rols` R package provides programmatic access to OLS from R. OLS is integrated with EMBL-EBI resources including Ensembl, UniProt, and Expression Atlas.

**Practical Workflow Example:** Step 1: Navigate to <https://www.ebi.ac.uk/ols4>. Step 2: Search for your term across all ontologies or within a specific ontology. Step 3: Browse the hierarchy to find the most specific applicable term. Step 4: Use the OLS API for programmatic access in your analysis pipeline.

**Reproducibility Notes:** Record the OLS version and access date. Note the ontology version for each term used.

**Quality-Control Notes:** Check that the term definition matches your intended meaning. Verify the ontology version.

## AG7 — BioPortal

---

**Official Website URL:** <https://bioportal.bioontology.org>

**Resource Type:** Portal / Tool (Ontology Lookup)

**Main Biological Domain:** Bioinformatics / Ontology / Biomedical informatics

**Short Definition:** BioPortal is a comprehensive repository of biomedical ontologies hosted by the National Center for Biomedical Ontology (NCBO), providing access to over 1,000 ontologies with tools for search, visualization, and annotation.

**What It Is Used For:** BioPortal is used to search for ontology terms, to annotate biomedical text with ontology terms, to compare ontologies, and to access ontology metadata.

**What Data It Contains:** BioPortal hosts over 1,000 biomedical ontologies including clinical, genomic, and research ontologies. Provides term definitions, synonyms, mappings, and visualization tools.

**Main Scientific Question It Helps Answer:** What ontology terms are available for annotating this biomedical concept, and how do different ontologies represent it?

**Typical Users:** Biomedical informaticists; clinical researchers; bioinformaticians; database curators.

**Example Scientific Questions:**

- What ontologies cover this clinical concept?
- How do different ontologies represent this disease?
- What are the mappings between this term and terms in other ontologies?

**Example Use Cases:**

- Annotating clinical text with ontology terms using the NCBO Annotator.
- Comparing how different ontologies represent a concept.
- Finding clinical ontologies (SNOMED CT, LOINC, RxNorm) alongside research ontologies.

**Input Data Accepted:** Term names, ontology IDs, biomedical text.

**Output Data Provided:** Term definitions, synonyms, mappings, ontology comparisons.

**Strengths:**

- Largest collection of biomedical ontologies (1,000+).
- Includes clinical ontologies (SNOMED CT, LOINC, RxNorm) alongside research ontologies.
- NCBO Annotator for text annotation.
- REST API for programmatic access.

**Limitations:**

- Quality varies widely across ontologies.
- Some ontologies require registration or have restricted access.
- Interface can be overwhelming given the large number of ontologies.

**Common Beginner Mistakes:**

- Assuming all BioPortal ontologies are high quality — quality varies widely.
- Confusing BioPortal (US-based, clinical focus) with OLS (EMBL-EBI, research focus).



**When to Use It:** Use BioPortal when you need clinical ontologies (SNOMED CT, LOINC, RxNorm), when annotating clinical text, or when comparing ontologies.

**When NOT to Use It:** For research ontologies (GO, HPO, MONDO), OLS may be more convenient. Do not use BioPortal as a substitute for individual ontologies.

**Related Databases or Alternatives:** OLS (EMBL-EBI alternative), OBO Foundry (ontology registry), individual ontologies.

**How It Connects to Other Resources:** BioPortal integrates with NCBO Annotator for text annotation and provides mappings between ontologies.

**API / FTP / Bulk Download / Programmatic Access:** BioPortal REST API at <https://data.bioontology.org/>. Requires API key (free registration). R package `ontologyIndex` available.

**Evidence or Curation Level:** Lookup service; quality depends on individual ontologies.

**Update Status:** Regularly updated; actively maintained by NCBO.

**Licensing or Access Restrictions:** Open access for most ontologies; some require registration.

**Citation / Recommended Reference:** Whetzel PL et al. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl\_2):W541–W545. doi:10.1093/nar/gkr469

**Beginner-Friendly Explanation:** BioPortal is similar to OLS but hosted in the US and includes a larger number of ontologies, including clinical ontologies used in healthcare (like SNOMED CT and LOINC). If you are working with clinical data or need to annotate clinical text with ontology terms, BioPortal is a good resource. It also has a text annotation tool (NCBO Annotator) that can automatically identify ontology terms in biomedical text.

**Advanced Technical Explanation:** BioPortal hosts over 1,000 ontologies and provides a REST API for programmatic access. The NCBO Annotator uses ontology terms to annotate biomedical text, supporting applications like clinical NLP and literature mining. BioPortal provides ontology mappings (cross-references between terms in different ontologies) and visualization tools for ontology hierarchies.

**Practical Workflow Example:** Step 1: Navigate to <https://bioportal.bioontology.org>. Step 2: Search for your term or browse ontologies. Step 3: Use the NCBO Annotator for text annotation. Step 4: Access the BioPortal API for programmatic use (requires free API key).

**Reproducibility Notes:** Record the BioPortal access date and the ontology version used.

**Quality-Control Notes:** Check ontology quality and currency. Verify that the ontology is appropriate for your domain.



## Short Index Entries — Category AG

### Gene Ontology (GO)

---

**Resource Type:** Ontology (Gene Function)

**Domain:** Molecular biology / Cell biology / Ontology

**Main Purpose:** The Gene Ontology provides a controlled vocabulary for describing gene and gene product attributes across species, organized in three aspects: Molecular Function, Biological Process, and Cellular Component.

**Best Used For:** Functional annotation of genes; GO enrichment analysis; cross-species functional comparison.

**Key Limitation:** GO annotations vary in quality; IEA (electronic) annotations are less reliable than experimental annotations. GO terms may be too broad or too specific for some analyses.

**Related Resources:** ECO (evidence types for GO annotations), OLS (ontology lookup), UniProt (uses GO), Ensembl (uses GO)

**Access / Licensing:** Open access; freely available at <https://geneontology.org>. CC0 license.

**Citation / Documentation:** Gene Ontology Consortium. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031. doi:10.1093/genetics/iyad031

### HPO (Human Phenotype Ontology)

---

**Resource Type:** Ontology (Human Phenotype)

**Domain:** Clinical genetics / Rare disease / Ontology

**Main Purpose:** The Human Phenotype Ontology provides a standardized vocabulary for describing human phenotypic abnormalities, enabling phenotype-driven gene discovery and rare disease diagnosis.

**Best Used For:** Phenotype annotation in clinical genetics; phenotype-driven gene discovery; rare disease diagnosis support.

**Key Limitation:** Coverage of rare phenotypes may be incomplete. Phenotype annotation requires clinical expertise.

**Related Resources:** MONDO (disease ontology), ClinGen (gene-disease validity), DECIPHER (rare disease patients), Orphanet (rare diseases)

**Access / Licensing:** Open access; freely available at <https://hpo.jax.org>. CC BY 4.0 license.

**Citation / Documentation:** Köhler S et al. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217. doi:10.1093/nar/gkaa1043

## DOID (Disease Ontology)

---

**Resource Type:** Ontology (Disease)

**Domain:** Disease / Clinical genomics / Ontology

**Main Purpose:** The Human Disease Ontology provides a standardized ontology for human disease, integrating disease concepts from MeSH, ICD, NCI Thesaurus, SNOMED CT, and OMIM.

**Best Used For:** Standardized disease annotation; cross-database disease concept mapping; disease classification.

**Key Limitation:** Overlaps with MONDO; for new projects, MONDO is generally preferred as it is more actively maintained and has broader cross-references.

**Related Resources:** MONDO (more comprehensive disease ontology), OMIM (disease genetics), Orphanet (rare diseases)

**Access / Licensing:** Open access; freely available at <https://disease-ontology.org>. CC BY 4.0 license.

**Citation / Documentation:** Schriml LM et al. (2022). The Human Disease Ontology 2022 update. *Nucleic Acids Research*, 50(D1):D1255–D1261. doi:10.1093/nar/gkab1063

## Uberon (Uber-anatomy Ontology)

---

**Resource Type:** Ontology (Anatomy)

**Domain:** Anatomy / Developmental biology / Comparative genomics / Ontology

**Main Purpose:** Uberon is a cross-species anatomy ontology covering anatomical structures in animals, enabling cross-species anatomical comparison and integration of expression data across species.

**Best Used For:** Cross-species anatomical annotation; integrating expression data across species; developmental biology.

**Key Limitation:** Cross-species anatomical homology is complex; some mappings may be approximate.

**Related Resources:** Cell Ontology (cell types), GO (gene function), Bgee (uses Uberon for expression mapping)

**Access / Licensing:** Open access; freely available at <https://uberon.github.io>. CC BY 3.0 license.

**Citation / Documentation:** Mungall CJ et al. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1):R5. doi:10.1186/gb-2012-13-1-r5

## Sequence Ontology (SO)

---

**Resource Type:** Ontology (Sequence Features)

**Domain:** Genomics / Sequence annotation / Ontology

**Main Purpose:** The Sequence Ontology provides a controlled vocabulary for describing features and attributes of biological sequences, used for standardized annotation of genomic features in GFF3 and other formats.

**Best Used For:** Standardized annotation of genomic features; GFF3 file annotation; variant annotation.

**Key Limitation:** Primarily used for sequence feature annotation; not suitable for functional or phenotypic annotation.

**Related Resources:** GO (gene function), Ensembl (uses SO for feature annotation), GENCODE (uses SO)

**Access / Licensing:** Open access; freely available at <http://www.sequenceontology.org>. CC BY-SA 4.0 license.

**Citation / Documentation:** Eilbeck K et al. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44. doi:10.1186/gb-2005-6-5-r44

## MeSH (Medical Subject Headings)

---

**Resource Type:** Controlled Vocabulary (Biomedical)

**Domain:** Biomedical literature / Clinical / Ontology

**Main Purpose:** MeSH is the NLM controlled vocabulary for indexing biomedical literature in PubMed and other databases. Provides standardized terms for diseases, drugs, anatomy, and other biomedical concepts.

**Best Used For:** Literature search in PubMed; standardized biomedical term lookup; clinical concept annotation.

**Key Limitation:** MeSH is a controlled vocabulary, not a formal ontology — limited logical relationships. Less suitable for automated reasoning than OWL ontologies.

**Related Resources:** PubMed (uses MeSH for indexing), MONDO (disease ontology with MeSH cross-references), BioPortal (hosts MeSH)

**Access / Licensing:** Open access; freely available at <https://www.nlm.nih.gov/mesh/>. Public domain.

**Citation / Documentation:** National Library of Medicine. (2024). Medical Subject Headings (MeSH). Available at <https://www.nlm.nih.gov/mesh/>.

## Category AH: Structural Biology Data Resources

### Category Overview

Structural biology data resources provide access to experimentally determined and computationally predicted three-dimensional structures of biological macromolecules. This category expands on the existing Category K (Protein Structure) to cover electron microscopy, NMR, small-angle scattering, integrative modeling, and structural classification resources.

### Critical distinctions:

- Primary structure archives (PDB, EMDB, BMRB): Repositories of experimentally determined structures. The authoritative source for structural data.
- Integrative/hybrid modeling (PDB-Dev, ModelArchive): Repositories for structures determined by combining multiple experimental methods or computational modeling.
- Structural classification (CATH, SCOPe): Databases that classify protein structures into hierarchical categories based on structural similarity.
- Structure annotation (PDBsum): Databases that provide additional annotations and analysis for PDB structures.

**WARNING: AlphaFold and other AI-predicted structures are models, not experimental structures. They should not be cited as experimental evidence for protein structure. Always distinguish between experimentally determined structures (PDB, EMDB, BMRB) and computationally predicted models (AlphaFold DB, SWISS-MODEL, ModelArchive).**

## AH1 — EMDB (Electron Microscopy Data Bank)

---

**Official Website URL:** <https://www.ebi.ac.uk/emdb>

**Resource Type:** Primary Archive (Structural Biology — Electron Microscopy)

**Main Biological Domain:** Structural biology / Cryo-EM / Electron microscopy

**Short Definition:** The Electron Microscopy Data Bank (EMDB) is the global archive for three-dimensional electron microscopy (EM) density maps of biological macromolecules and subcellular structures, including cryo-EM, cryo-ET, and electron crystallography data.

**What It Is Used For:** EMDB is used to access and deposit cryo-EM density maps, to find EM structures of biological macromolecules, and to access raw EM data for reanalysis.

**What Data It Contains:** EMDB contains over 40,000 EM density maps (as of 2024) from cryo-EM single-particle analysis, cryo-electron tomography (cryo-ET), and electron crystallography. Each entry includes the density map, associated atomic model (if available, linked to PDB), metadata, and validation reports.

**Main Scientific Question It Helps Answer:** Is there a cryo-EM structure of this protein or complex, and what is the resolution?

**Typical Users:** Structural biologists; cryo-EM specialists; biochemists; drug discovery researchers.

**Example Scientific Questions:**

- Is there a cryo-EM structure of this protein complex?
- What is the resolution of the cryo-EM map for this entry?
- What is the associated atomic model for this EM map?

**Example Use Cases:**

- Finding cryo-EM structures of membrane proteins.
- Accessing EM density maps for model building.
- Comparing cryo-EM structures of the same protein in different states.

**Input Data Accepted:** Protein names, EMDB accession numbers, PDB IDs.

**Output Data Provided:** EM density maps (MRC/CCP4 format), metadata, validation reports, links to PDB atomic models.

**Strengths:** Global archive for cryo-EM data; comprehensive coverage; Linked to PDB for associated atomic models; Provides validation reports for map quality; Freely accessible with bulk download.

**Limitations:** EM maps require specialized software for visualization (UCSF ChimeraX, Coot); Resolution varies widely; not all maps are suitable for atomic model building; Raw data (micrographs) are not always deposited.

**Common Beginner Mistakes:** Confusing EMDB (EM density maps) with PDB (atomic coordinates); Not checking map resolution before using for structural analysis; Assuming all EMDB entries have associated atomic models.

**When to Use It:** Use EMDB when looking for cryo-EM structures, when accessing EM density maps for model building, or when depositing cryo-EM data.

**When NOT to Use It:** Do not use EMDB for X-ray crystallography or NMR structures (use PDB), or for predicted structures (use AlphaFold DB).

**Related Databases or Alternatives:** PDB (atomic coordinates, linked to EMDB), BMRB (NMR data), AlphaFold DB (predicted structures), EMPIAR (raw EM data).

**How It Connects to Other Resources:** EMDB is linked to PDB for associated atomic models. EMPIAR (Electron Microscopy Public Image Archive) hosts raw EM data.

**API / FTP / Bulk Download / Programmatic Access:** EMDB REST API at <https://www.ebi.ac.uk/emdb/api/>. Returns JSON. Bulk download available.

**Evidence or Curation Level:** Experimentally determined; peer-reviewed deposition.

**Update Status:** Continuously updated; actively maintained by EMBL-EBI, PDBe.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Lawson CL et al. (2021). EMDB: Empowering structural characterization of macromolecules and complexes. *Nucleic Acids Research*, 49(D1):D1019–D1026. doi:10.1093/nar/gkaa1062

**Beginner-Friendly Explanation:** The Electron Microscopy Data Bank (EMDB) is the archive for cryo-EM structures. Cryo-EM is a technique that uses electron microscopy to determine the three-dimensional structure of proteins and other biological molecules at near-atomic resolution. EMDB stores the density maps produced by cryo-EM experiments. These maps show the shape of the molecule but not the exact positions of individual atoms — for that, you need the associated atomic model, which is stored in the PDB.

**Advanced Technical Explanation:** EMDB stores EM density maps in MRC/CCP4 format, along with metadata including resolution, symmetry, and experimental conditions. Each entry is linked to associated PDB atomic models where available. EMDB provides validation reports assessing map quality (FSC curves, local resolution). The EMDB API supports programmatic access to map metadata and download links. EMPIAR (Electron Microscopy Public Image Archive) is a companion resource for raw EM data (micrographs, tilt series).

**Practical Workflow Example:** Step 1: Search EMDB at <https://www.ebi.ac.uk/emdb> for your protein. Step 2: Review the resolution and map quality. Step 3: Download the density map in MRC format. Step 4: Visualize in UCSF ChimeraX or Coot. Step 5: Access the associated PDB atomic model if available.

**Reproducibility Notes:** Record the EMDB accession number and the associated PDB ID. Note the map resolution and the software used for visualization.

**Quality-Control Notes:** Check the map resolution and validation report. Verify that the map quality is sufficient for your analysis. Note that resolution varies widely across EMDB entries.

## AH2 — BMRB (Biological Magnetic Resonance Data Bank)

---

**Official Website URL:** <https://bmr.io>

**Resource Type:** Primary Archive (Structural Biology — NMR)

**Main Biological Domain:** Structural biology / NMR spectroscopy

**Short Definition:** The Biological Magnetic Resonance Data Bank (BMRB) is the global archive for NMR spectroscopic data from biological macromolecules, including chemical shift assignments, relaxation data, and other NMR-derived parameters.

**What It Is Used For:** BMRB is used to access NMR spectroscopic data for biological macromolecules, to find chemical shift assignments for proteins and nucleic acids, and to deposit NMR data.

**What Data It Contains:** BMRB contains NMR data for over 12,000 entries (as of 2024), including chemical shift assignments, coupling constants, relaxation parameters, and other NMR-derived data for proteins, nucleic acids, and small molecules.

**Main Scientific Question It Helps Answer:** Is there NMR data available for this protein, and what are the chemical shift assignments?

**Typical Users:** NMR spectroscopists; structural biologists; biochemists.

**Example Scientific Questions:**

- Are there chemical shift assignments available for this protein?
- What NMR data is available for this protein?
- What is the BMRB accession for this NMR structure?

**Example Use Cases:**

- Accessing chemical shift assignments for NMR structure determination.
- Comparing NMR data across different conditions.
- Depositing NMR data for publication.

**Input Data Accepted:** Protein names, BMRB accession numbers, PDB IDs.

**Output Data Provided:** Chemical shift assignments, relaxation data, NMR spectra metadata.

**Strengths:**

- Global archive for NMR data; comprehensive coverage.
- Linked to PDB for associated atomic models.
- Freely accessible.

**Limitations:**

- NMR data requires specialized expertise to interpret.
- Coverage is limited to proteins and nucleic acids amenable to NMR.
- NMR structures are typically limited to smaller proteins (<50 kDa).

**Common Beginner Mistakes:**

- Confusing BMRB (NMR data) with PDB (atomic coordinates) or EMDB (EM maps).
- Not checking whether the NMR data is for the same construct as your protein of interest.



**When to Use It:** Use BMRB when looking for NMR data for a protein, when accessing chemical shift assignments, or when depositing NMR data.

**When NOT to Use It:** Do not use BMRB for X-ray crystallography or cryo-EM structures (use PDB or EMDB).

**Related Databases or Alternatives:** PDB (atomic coordinates, linked to BMRB), EMDB (EM maps), AlphaFold DB (predicted structures).

**How It Connects to Other Resources:** BMRB is linked to PDB for associated atomic models.

**API / FTP / Bulk Download / Programmatic Access:** BMRB REST API at <https://api.bmr.io/>. Returns JSON.

**Evidence or Curation Level:** Experimentally determined; peer-reviewed deposition.

**Update Status:** Continuously updated; actively maintained by University of Wisconsin-Madison.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Ulrich EL et al. (2008). BioMagResBank. Nucleic Acids Research, 36(suppl\_1):D402–D408. doi:10.1093/nar/gkm957

**Beginner-Friendly Explanation:** The BMRB is the archive for NMR (nuclear magnetic resonance) data from biological molecules. NMR is a technique that uses magnetic fields to determine the structure and dynamics of proteins and other molecules in solution. Unlike X-ray crystallography or cryo-EM, NMR can study proteins in their natural solution state. BMRB stores the raw NMR data (chemical shifts, relaxation times) that are used to determine protein structures.

**Advanced Technical Explanation:** BMRB stores NMR data in NMR-STAR format, including chemical shift assignments, coupling constants, relaxation parameters (T1, T2, NOE), and other NMR-derived data. Each entry is linked to associated PDB atomic models where available. BMRB also hosts data for small molecules and metabolites. The BMRB API supports programmatic access to NMR data.

**Practical Workflow Example:** Step 1: Search BMRB at <https://bmr.io> for your protein. Step 2: Review the available NMR data. Step 3: Download chemical shift assignments. Step 4: Access the associated PDB atomic model if available.

**Reproducibility Notes:** Record the BMRB accession number and the associated PDB ID.

**Quality-Control Notes:** Check that the NMR data is for the same protein construct as your protein of interest. Verify the experimental conditions.

## AH3 — CATH (Class, Architecture, Topology, Homologous Superfamily)

**Official Website URL:** <https://www.cathdb.info>

**Resource Type:** Database (Protein Structure Classification)

**Main Biological Domain:** Structural biology / Protein evolution / Structural classification

**Short Definition:** CATH is a hierarchical classification of protein domain structures from the PDB, organizing protein domains into four levels: Class (C), Architecture (A), Topology (T), and Homologous Superfamily (H).

**What It Is Used For:** CATH is used to classify protein domains by structural similarity, to find structurally related proteins, to study protein evolution, and to identify structural superfamilies.

**What Data It Contains:** CATH classifies over 500,000 protein domain structures from the PDB into approximately 5,000 superfamilies. Each domain is assigned to a CATH classification and linked to functional annotations.

**Main Scientific Question It Helps Answer:** What structural superfamily does this protein domain belong to, and what other proteins share this fold?

**Typical Users:** Structural biologists; evolutionary biologists; bioinformaticians.

**Example Scientific Questions:**

- What CATH superfamily does this protein domain belong to?
- What other proteins share this structural fold?
- What is the evolutionary relationship between these protein structures?

**Example Use Cases:**

- Identifying structural homologs for a protein of unknown function.
- Studying protein evolution through structural classification.
- Finding proteins with similar folds for drug design.

**Input Data Accepted:** PDB IDs, protein names, CATH IDs.

**Output Data Provided:** CATH classifications, structural superfamilies, domain boundaries.

**Strengths:** Comprehensive hierarchical classification of protein structures; Enables identification of structural homologs; Linked to functional annotations; Freely accessible.

**Limitations:** Classification is based on PDB structures; not all proteins have PDB structures; Structural classification may not always reflect evolutionary relationships; CATH and SCOPe may classify the same protein differently.

**Common Beginner Mistakes:** Confusing CATH (structural classification) with sequence-based classification (Pfam, InterPro); Assuming CATH and SCOPe classifications are equivalent — they use different criteria.

**When to Use It:** Use CATH when classifying protein structures, when finding structural homologs, or when studying protein evolution through structural similarity.

**When NOT to Use It:** Do not use CATH for sequence-based classification (use Pfam or InterPro) or for proteins without PDB structures.

**Related Databases or Alternatives:** SCOPe (alternative structural classification), PDB (source of structures), Pfam (sequence-based domain classification), InterPro (integrated domain classification).

**How It Connects to Other Resources:** CATH is linked to PDB (source structures), UniProt (protein sequences), and Gene Ontology (functional annotations).

**API / FTP / Bulk Download / Programmatic Access:** CATH REST API at <https://www.cathdb.info/api/>. Returns JSON.

**Evidence or Curation Level:** Computationally classified from PDB structures; manually curated for key superfamilies.

**Update Status:** Regularly updated; actively maintained by University College London.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Sillitoe I et al. (2021). CATH: increased structural coverage of functional space. *Nucleic Acids Research*, 49(D1):D266–D273. doi:10.1093/nar/gkaa1079

**Beginner-Friendly Explanation:** CATH is a database that classifies protein structures into a hierarchy based on their three-dimensional shape. The four levels are: Class (the overall shape — mainly alpha, mainly beta, or mixed), Architecture (the arrangement of secondary structures), Topology (the connectivity of secondary structures), and Homologous Superfamily (proteins that are evolutionarily related). CATH is useful for finding proteins that have similar structures to your protein of interest, even if they have very different sequences.

**Advanced Technical Explanation:** CATH uses a combination of automated algorithms and manual curation to classify protein domains from the PDB. The classification is hierarchical: Class (C) → Architecture (A) → Topology (T) → Homologous Superfamily (H). Each superfamily contains proteins that are evolutionarily related based on structural and functional similarity. CATH-Gene3D extends CATH classifications to unstructured proteins using sequence-based methods.

**Practical Workflow Example:** Step 1: Search CATH at <https://www.cathdb.info> for your protein or PDB ID. Step 2: Find the CATH classification for your domain. Step 3: Browse the superfamily to find structurally related proteins. Step 4: Use the CATH API for programmatic access.

**Reproducibility Notes:** Record the CATH version and the CATH ID for each domain classified.

**Quality-Control Notes:** Check that the domain boundaries are correct. Note that CATH and SCOPe may classify the same protein differently.

## AH4 — SCOPe (Structural Classification of Proteins — extended)

**Official Website URL:** <https://scop.berkeley.edu>

**Resource Type:** Database (Protein Structure Classification)

**Main Biological Domain:** Structural biology / Protein evolution / Structural classification

**Short Definition:** SCOPe (Structural Classification of Proteins — extended) is a hierarchical classification of protein structures from the PDB, organizing protein domains into Class, Fold, Superfamily, and Family levels, with an emphasis on evolutionary relationships.

**What It Is Used For:** SCOPe is used to classify protein domains by structural and evolutionary similarity, to find structurally related proteins, and to study protein evolution.

**What Data It Contains:** SCOPe classifies protein domain structures from the PDB into approximately 4,000 superfamilies. Each domain is assigned to a SCOPe classification with emphasis on evolutionary relationships.

**Main Scientific Question It Helps Answer:** What structural fold and evolutionary superfamily does this protein domain belong to?

**Typical Users:** Structural biologists; evolutionary biologists; bioinformaticians.

**Example Scientific Questions:**

- What SCOPe fold does this protein belong to?
- What proteins are in the same SCOPe superfamily?
- How does SCOPe classify this protein compared to CATH?

**Example Use Cases:**

- Identifying structural and evolutionary relationships between proteins.
- Studying protein fold evolution.
- Benchmarking structural comparison algorithms.

**Input Data Accepted:** PDB IDs, protein names, SCOPe IDs.

**Output Data Provided:** SCOPe classifications, fold assignments, superfamily memberships.

**Strengths:** Emphasis on evolutionary relationships in classification; Widely used benchmark for structural comparison algorithms; Freely accessible.

**Limitations:** Updates have been less frequent than CATH in recent years; Classification may differ from CATH for some proteins; Coverage limited to PDB structures.

**Common Beginner Mistakes:** Assuming SCOPe and CATH classifications are equivalent — they use different criteria and may disagree.

**When to Use It:** Use SCOPe when studying protein fold evolution, when benchmarking structural comparison algorithms, or when comparing with CATH classifications.

**When NOT to Use It:** For most current structural classification needs, CATH may be more actively maintained. Do not use SCOPe for sequence-based classification.

**Related Databases or Alternatives:** CATH (alternative structural classification), PDB (source of structures), Pfam (sequence-based domain classification).

**How It Connects to Other Resources:** SCOPe is linked to PDB (source structures) and UniProt (protein sequences).

**API / FTP / Bulk Download / Programmatic Access:** SCOPe data available for download at <https://scop.berkeley.edu/downloads/>.

**Evidence or Curation Level:** Computationally classified from PDB structures; manually curated.

**Update Status:** Periodically updated; maintained by University of California, Berkeley.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Fox NK et al. (2014). SCOPe: Structural Classification of Proteins — extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309. doi:10.1093/nar/gkt1240

**Beginner-Friendly Explanation:** SCOPe is similar to CATH — it classifies protein structures into a hierarchy based on their three-dimensional shape and evolutionary relationships. The four levels are: Class (overall shape), Fold (similar secondary structure arrangement), Superfamily (probable evolutionary relationship), and Family (clear evolutionary relationship). SCOPe and CATH are complementary resources; they sometimes classify the same protein differently because they use different criteria.

**Advanced Technical Explanation:** SCOPe extends the original SCOP (Structural Classification of Proteins) database by incorporating automated classification methods (ASTRAL) alongside manual curation. The classification emphasizes evolutionary relationships: proteins in the same superfamily are considered to have a common evolutionary origin based on structural and functional similarity. SCOPe is widely used as a benchmark for structural comparison and fold recognition algorithms.

**Practical Workflow Example:** Step 1: Search SCOPe at <https://scop.berkeley.edu> for your protein or PDB ID. Step 2: Find the SCOPe classification. Step 3: Browse the superfamily to find related proteins. Step 4: Compare with CATH classification if needed.

**Reproducibility Notes:** Record the SCOPe version and the SCOPe ID for each domain classified.

**Quality-Control Notes:** Compare with CATH classification for consistency. Note that SCOPe and CATH may disagree for some proteins.

## Short Index Entries — Category AH

### SASBDB (Small Angle Scattering Biological Data Bank)

**Resource Type:** Primary Archive (Structural Biology — SAXS/SANS)

**Domain:** Structural biology / Small-angle scattering

**Main Purpose:** Global archive for small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS) data from biological macromolecules, providing solution-state structural information.

**Best Used For:** Accessing SAXS/SANS data for proteins and complexes; solution-state structural characterization.

**Key Limitation:** SAXS/SANS provides low-resolution structural information; not suitable for atomic-level structural analysis.

**Related Resources:** PDB (atomic structures), EMDB (EM maps), BMRB (NMR data)

**Access / Licensing:** Open access; freely available at <https://www.sasbdb.org>.

**Citation / Documentation:** Valentini E et al. (2015). SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Research*, 43(D1):D357–D363. doi:10.1093/nar/gku1047

### PDB-Dev (Prototype System for Archiving Integrative/Hybrid Structural Models)

**Resource Type:** Primary Archive (Structural Biology — Integrative Modeling)

**Domain:** Structural biology / Integrative modeling

**Main Purpose:** Archive for structural models determined by integrative/hybrid methods that combine multiple experimental data types (cryo-EM, SAXS, crosslinking MS, etc.) with computational modeling.

**Best Used For:** Accessing integrative structural models of large complexes; depositing integrative models.

**Key Limitation:** Integrative models have varying levels of accuracy; always check the experimental data used and the model validation.

**Related Resources:** PDB (atomic structures), EMDB (EM maps), BMRB (NMR data), SASBDB (SAXS data)

**Access / Licensing:** Open access; freely available at <https://pdb-dev.wwpdb.org>.

**Citation / Documentation:** Burley SK et al. (2022). Protein Data Bank: a comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. *Biomolecules*, 12(10):1425. doi:10.3390/biom12101425

### ModelArchive

**Resource Type:** Archive (Computational Structural Models)

**Domain:** Structural biology / Computational modeling

**Main Purpose:** Archive for computationally predicted protein structure models that are not based on experimental data, including homology models and AI-predicted structures not deposited in AlphaFold DB.

**Best Used For:** Depositing and accessing computational protein structure models; models from methods not covered by AlphaFold DB.



**Key Limitation:** Models are computational predictions, not experimental structures. Accuracy varies widely. Always validate models before use.

**Related Resources:** AlphaFold DB (AI-predicted structures), PDB (experimental structures), SWISS-MODEL (homology modeling)

**Access / Licensing:** Open access; freely available at <https://modelarchive.org>.

**Citation / Documentation:** Haas J et al. (2023). ModelArchive: a resource for computationally predicted protein structures. Nucleic Acids Research, 51(D1):D368–D374. doi:10.1093/nar/gkac1062

## PDBsum

---

**Resource Type:** Database (Protein Structure Annotation)

**Domain:** Structural biology / Protein structure analysis

**Main Purpose:** Provides pictorial summaries and analyses of PDB structures, including secondary structure diagrams, ligand interactions, protein-protein interfaces, and structural quality metrics.

**Best Used For:** Quick visual summary of PDB structure features; ligand binding analysis; protein-protein interface analysis.

**Key Limitation:** Annotation is automated; may not capture all biologically relevant features. Not a substitute for manual structural analysis.

**Related Resources:** PDB (source structures), CATH (structural classification), SCOPe (structural classification)

**Access / Licensing:** Open access; freely available at <https://www.ebi.ac.uk/pdbsum>.

**Citation / Documentation:** Laskowski RA et al. (2018). PDBsum: Structural summaries of PDB entries. Protein Science, 27(1):129–134. doi:10.1002/pro.3289



## Category AI: Pharmacogenomics and Drug-Target Interaction Resources

### Category Overview

Pharmacogenomics and drug-target interaction resources provide the infrastructure for translating genomic and molecular data into drug discovery and precision medicine insights. This category covers resources for drug-target interactions, pharmacogenomics, drug mechanisms, and clinical drug-gene relationships.

### Critical distinctions:

- Drug-target interaction databases (DrugBank, ChEMBL, BindingDB): Curated or experimental data on drug-target binding. Vary in scope, curation level, and data type.
- Pharmacogenomics (PharmGKB): Curated evidence for how genetic variation affects drug response. Clinically relevant for precision medicine.
- Target prioritization (Open Targets): Integrates genetic, genomic, and clinical evidence to prioritize drug targets. Designed for drug discovery.
- Drug-gene interaction (DGIdb): Aggregates drug-gene interactions from multiple sources. Useful for identifying actionable targets.

**WARNING: Drug-target interaction data varies widely in quality and clinical relevance. Binding affinity data (IC<sub>50</sub>, K<sub>d</sub>) from biochemical assays does not necessarily predict clinical efficacy. Always distinguish between biochemical binding data, cellular activity data, and clinical evidence. For clinical drug-gene interactions, use PharmGKB or CPIC guidelines.**

## AI1 — Open Targets Platform

---

**Official Website URL:** <https://platform.opentargets.org>

**Resource Type:** Platform / Knowledgebase (Drug Target Prioritization)

**Main Biological Domain:** Drug discovery / Genomics / Translational medicine

**Short Definition:** The Open Targets Platform is a comprehensive, publicly available platform that integrates genetic, genomic, and clinical evidence to systematically identify and prioritize drug targets for human diseases.

**What It Is Used For:** Open Targets is used to identify and prioritize drug targets for a disease of interest, to assess the genetic evidence supporting a target-disease association, and to explore the drug pipeline for a target.

**What Data It Contains:** Open Targets integrates evidence from GWAS (GWAS Catalog), rare disease genetics (ClinVar, ClinGen), somatic mutations (COSMIC, TCGA), gene expression (GTEx, Expression Atlas), animal models, pathway analysis, and drug databases (ChEMBL, DrugBank) to score target-disease associations.

**Main Scientific Question It Helps Answer:** What are the best-supported drug targets for this disease, and what is the evidence?

**Typical Users:** Drug discovery researchers; translational scientists; bioinformaticians; pharmaceutical industry researchers.

**Example Scientific Questions:**

- What are the top drug targets for type 2 diabetes based on genetic evidence?
- What is the evidence supporting this gene as a target for this disease?
- What drugs are in development for this target?

**Example Use Cases:**

- Prioritizing drug targets for a disease based on genetic evidence.
- Assessing the druggability of a candidate target.
- Exploring the drug pipeline for a target of interest.

**Input Data Accepted:** Disease names, gene names, EFO terms, Ensembl IDs.

**Output Data Provided:** Target-disease association scores, evidence summaries, drug pipeline information.

**Strengths:** Integrates multiple evidence types for comprehensive target assessment; Genetic evidence from GWAS and rare disease genetics is particularly valuable; Provides drug pipeline information from ChEMBL; Freely accessible with bulk download; REST API for programmatic access.

**Limitations:** Association scores are composite metrics; individual evidence types should be reviewed; Coverage depends on available data; rare diseases may have limited evidence; Not a substitute for experimental validation of targets.

**Common Beginner Mistakes:** Treating Open Targets association scores as definitive evidence for target validity; Not reviewing the individual evidence types supporting a score; Confusing Open Targets Platform (target prioritization) with Open Targets Genetics (GWAS fine-mapping).

**When to Use It:** Use Open Targets when prioritizing drug targets for a disease, when assessing genetic evidence for a target-disease association, or when exploring the drug pipeline.

**When NOT to Use It:** Do not use Open Targets as a substitute for experimental validation. For GWAS fine-mapping, use Open Targets Genetics.

**Related Databases or Alternatives:** GWAS Catalog (GWAS evidence), ClinVar (clinical evidence), ChEMBL (drug data), PharmGKB (pharmacogenomics), DrugBank (drug information).

**How It Connects to Other Resources:** Open Targets integrates GWAS Catalog, ClinVar, COSMIC, GTEx, Expression Atlas, ChEMBL, and other resources.

**API / FTP / Bulk Download / Programmatic Access:** Open Targets GraphQL API at <https://api.platform.opentargets.org/api/v4/graphql>. Python and R clients available.

**Evidence or Curation Level:** Integrated from multiple sources; evidence quality varies by data type.

**Update Status:** Quarterly releases; actively maintained by Open Targets consortium (EMBL-EBI, Wellcome Sanger, GSK, Pfizer, Takeda, Bristol Myers Squibb).

**Licensing or Access Restrictions:** Open access; CC BY 4.0.

**Citation / Recommended Reference:** Ochoa D et al. (2023). The next-generation Open Targets Platform: reimaged, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1):D1353–D1359. doi:10.1093/nar/gkac1046

**Beginner-Friendly Explanation:** Open Targets is a platform that helps researchers find the best drug targets for a disease. It collects evidence from many different sources — genetic studies, clinical data, gene expression, animal models, and existing drugs — and combines them into a score for each gene-disease pair. A high score means there is strong evidence that a gene is involved in a disease and might be a good drug target. Open Targets is particularly useful because it emphasizes genetic evidence, which is considered the most reliable type of evidence for drug target validation.

**Advanced Technical Explanation:** Open Targets uses a weighted scoring system that combines evidence from multiple data types: genetic associations (GWAS, rare variants), somatic mutations, gene expression, animal models, pathway analysis, and clinical evidence. The platform uses Ensembl gene IDs and EFO disease terms for standardization. The GraphQL API provides programmatic access to all platform data. Open Targets Genetics is a companion platform focused on GWAS fine-mapping and variant-to-gene assignment.

**Practical Workflow Example:** Step 1: Navigate to <https://platform.opentargets.org>. Step 2: Search for your disease of interest. Step 3: Review the top-scoring targets and their evidence. Step 4: Click on a target to see detailed evidence. Step 5: Use the GraphQL API for programmatic access to association data.

## AI2 — PharmGKB (Pharmacogenomics Knowledgebase)

---

**Official Website URL:** <https://www.pharmgkb.org>

**Resource Type:** Knowledgebase (Pharmacogenomics)

**Main Biological Domain:** Pharmacogenomics / Precision medicine / Clinical genomics

**Short Definition:** PharmGKB is a pharmacogenomics knowledge resource that curates information about how genetic variation affects drug response, providing variant annotations, drug-gene relationships, and clinical pharmacogenomics guidelines.

**What It Is Used For:** PharmGKB is used to find curated evidence for how genetic variants affect drug response, to access clinical pharmacogenomics guidelines (CPIC), and to explore drug-gene relationships.

**What Data It Contains:** PharmGKB contains curated variant annotations (how specific variants affect drug response), drug-gene relationships, pathway diagrams, and links to CPIC (Clinical Pharmacogenomics Implementation Consortium) guidelines for clinical implementation.

**Main Scientific Question It Helps Answer:** How does this genetic variant affect the response to this drug, and are there clinical guidelines?

**Typical Users:** Clinical pharmacologists; clinical geneticists; precision medicine researchers; pharmacists.

**Example Scientific Questions:**

- How does the CYP2D6 genotype affect codeine metabolism?
- Are there CPIC guidelines for this drug-gene pair?
- What variants affect warfarin dosing?

**Example Use Cases:**

- Implementing pharmacogenomics testing in clinical practice.
- Identifying patients at risk for adverse drug reactions.
- Exploring drug-gene interactions for precision medicine.

**Input Data Accepted:** Drug names, gene names, variant identifiers (rsIDs).

**Output Data Provided:** Variant annotations, drug-gene relationships, CPIC guidelines, pathway diagrams.

**Strengths:**

- Curated pharmacogenomics evidence with clinical relevance.
- Links to CPIC guidelines for clinical implementation.
- Pathway diagrams for drug metabolism and response.
- Freely accessible.

**Limitations:**

- Coverage is not comprehensive for all drug-gene pairs.
- Evidence quality varies; not all annotations have clinical guidelines.
- Pharmacogenomics is complex; clinical implementation requires expert interpretation.

**Common Beginner Mistakes:**

- Assuming all PharmGKB annotations have clinical guidelines — many are research-level evidence only.



- Confusing PharmGKB (pharmacogenomics) with DrugBank (drug information) or Open Targets (target prioritization).

**When to Use It:** Use PharmGKB when exploring how genetic variants affect drug response, when looking for CPIC guidelines, or when implementing pharmacogenomics in clinical practice.

**When NOT to Use It:** Do not use PharmGKB for drug-target binding data (use ChEMBL or BindingDB) or for drug discovery target prioritization (use Open Targets).

**Related Databases or Alternatives:** CPIC (clinical pharmacogenomics guidelines), DrugBank (drug information), Open Targets (target prioritization), ClinVar (clinical variant interpretation).

**How It Connects to Other Resources:** PharmGKB integrates with CPIC, DrugBank, dbSNP, and ClinVar.

**API / FTP / Bulk Download / Programmatic Access:** PharmGKB REST API at <https://api.pharmgkb.org/>. Returns JSON. Bulk download available.

**Evidence or Curation Level:** Manually curated from published literature; evidence levels assigned (1A, 1B, 2A, 2B, 3, 4).

**Update Status:** Regularly updated; actively maintained by Stanford University.

**Licensing or Access Restrictions:** Open access; Creative Commons Attribution-ShareAlike 4.0.

**Citation / Recommended Reference:** Whirl-Carrillo M et al. (2021). An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 110(3):563–572. doi:10.1002/cpt.2350

**Beginner-Friendly Explanation:** PharmGKB is a database that collects information about how your genes affect how your body responds to drugs. For example, some people have a variant in the CYP2D6 gene that makes them metabolize certain drugs much faster or slower than normal, which can affect whether a drug works or causes side effects. PharmGKB curates this information from published research and links it to clinical guidelines (CPIC) that tell doctors how to adjust drug dosing based on a patient's genotype.

**Advanced Technical Explanation:** PharmGKB uses a tiered evidence system (1A–4) to classify the strength of pharmacogenomics evidence. Level 1A represents the highest evidence (CPIC guideline or FDA label with pharmacogenomics information). PharmGKB curates variant annotations, drug-gene relationships, and pathway diagrams. The CPIC (Clinical Pharmacogenomics Implementation Consortium) guidelines, developed in collaboration with PharmGKB, provide actionable clinical recommendations for specific drug-gene pairs.

**Practical Workflow Example:** Step 1: Navigate to <https://www.pharmgkb.org>. Step 2: Search for your drug or gene. Step 3: Review variant annotations and evidence levels. Step 4: Check for CPIC guidelines for clinical implementation. Step 5: Download data using the PharmGKB API.

## AI3 — DGIdb (Drug-Gene Interaction Database)

**Official Website URL:** <https://www.dgldb.org>

**Resource Type:** Database / Aggregator (Drug-Gene Interactions)



**Main Biological Domain:** Drug discovery / Pharmacogenomics / Cancer genomics

**Short Definition:** DGIdb is a database that aggregates drug-gene interaction data from multiple sources, providing a unified interface for querying drug-gene interactions and identifying druggable genes.

**What It Is Used For:** DGIdb is used to find drug-gene interactions for a gene of interest, to identify druggable genes in a gene list, and to explore the drug landscape for a target.

**What Data It Contains:** DGIdb aggregates drug-gene interaction data from over 40 sources including DrugBank, ChEMBL, PharmGKB, CIViC, and others. Provides interaction types, evidence levels, and links to source databases.

**Main Scientific Question It Helps Answer:** What drugs interact with this gene, and is this gene druggable?

**Typical Users:** Cancer genomicists; drug discovery researchers; bioinformaticians.

**Example Scientific Questions:**

- What drugs target this gene?
- Is this gene druggable?
- What drug-gene interactions are known for this list of genes?

**Example Use Cases:**

- Identifying actionable mutations in cancer genomics.
- Finding drugs for a list of differentially expressed genes.
- Assessing the druggability of candidate targets.

**Input Data Accepted:** Gene names, drug names, gene lists.

**Output Data Provided:** Drug-gene interactions, druggability categories, source database links.

**Strengths:**

- Aggregates data from 40+ sources for comprehensive coverage.
- Provides druggability categories for genes.
- Useful for identifying actionable mutations in cancer.
- Freely accessible with API.

**Limitations:**

- Data quality varies across source databases.
- Aggregation may introduce redundancy or inconsistencies.
- Not all interactions are clinically actionable.

**Common Beginner Mistakes:**

- Assuming all DGIdb interactions are clinically actionable — many are research-level.
- Not checking the source database for each interaction.

**When to Use It:** Use DGIdb when identifying drug-gene interactions for a gene or gene list, when assessing druggability, or when finding actionable mutations in cancer.

**When NOT to Use It:** For clinical pharmacogenomics, use PharmGKB. For detailed drug information, use DrugBank or ChEMBL.



**Related Databases or Alternatives:** DrugBank (drug information), ChEMBL (drug-target binding), PharmGKB (pharmacogenomics), Open Targets (target prioritization), CIViC (cancer variant interpretation).

**How It Connects to Other Resources:** DGIdb aggregates data from DrugBank, ChEMBL, PharmGKB, CIViC, and 40+ other sources.

**API / FTP / Bulk Download / Programmatic Access:** DGIdb GraphQL API at <https://www.dgldb.org/api>. Returns JSON.

**Evidence or Curation Level:** Aggregated from multiple sources; evidence quality varies.

**Update Status:** Regularly updated; actively maintained by Washington University in St. Louis.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Freshour SL et al. (2021). Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsourcing efforts. *Nucleic Acids Research*, 49(D1):D1144–D1151. doi:10.1093/nar/gkaa1084

**Beginner-Friendly Explanation:** DGIdb is a database that collects information about which drugs interact with which genes, pulling data from over 40 different sources. If you have a list of genes from a genomics study and want to know which ones have drugs that target them, DGIdb is a good starting point. It is particularly useful in cancer genomics for identifying 'actionable' mutations — mutations in genes that can be targeted with existing drugs.

**Advanced Technical Explanation:** DGIdb aggregates drug-gene interaction data from 40+ sources using a standardized data model. Interactions are categorized by type (inhibitor, activator, agonist, antagonist, etc.) and evidence level. The druggability categories (clinically actionable, drug resistance, druggable genome) help prioritize genes for drug development. The DGIdb GraphQL API supports batch queries for gene lists.

**Practical Workflow Example:** Step 1: Navigate to <https://www.dgldb.org>. Step 2: Enter your gene or gene list. Step 3: Review drug-gene interactions and druggability categories. Step 4: Use the GraphQL API for programmatic access.

**Reproducibility Notes:** Record the DGIdb version and access date. Note the source databases included in the query.

**Quality-Control Notes:** Check the source database for each interaction. Verify evidence levels. Note that not all interactions are clinically actionable.



## AI4 — DrugCentral

---

**Official Website URL:** <https://drugcentral.org>

**Resource Type:** Database / Knowledgebase (Drug Information)

**Main Biological Domain:** Drug discovery / Pharmacology / Clinical medicine

**Short Definition:** DrugCentral is an online drug information resource that integrates drug-target interactions, drug indications, mechanisms of action, and pharmacological data for FDA-approved and investigational drugs.

**What It Is Used For:** DrugCentral is used to find drug-target interactions, drug indications, mechanisms of action, and pharmacological data for approved drugs.

**What Data It Contains:** DrugCentral contains data for over 4,000 drugs, including drug-target interactions (with binding affinities), drug indications (from FDA labels), mechanisms of action, pharmacokinetics, and adverse effects.

**Main Scientific Question It Helps Answer:** What are the targets, indications, and mechanisms of action of this drug?

**Typical Users:** Drug discovery researchers; pharmacologists; clinical researchers; bioinformaticians.

**Example Scientific Questions:**

- What are the targets of this drug?
- What are the approved indications for this drug?
- What is the mechanism of action of this drug?

**Example Use Cases:**

- Drug repurposing — finding new indications for existing drugs.
- Identifying off-target effects of drugs.
- Exploring drug mechanisms for a disease of interest.

**Input Data Accepted:** Drug names, INN names, drug IDs.

**Output Data Provided:** Drug-target interactions, indications, mechanisms of action, pharmacokinetics.

**Strengths:**

- Integrates drug-target interactions with clinical indication data.
- Covers FDA-approved drugs with high-quality data.
- Useful for drug repurposing.
- Freely accessible.

**Limitations:**

- Coverage of investigational drugs is less comprehensive than approved drugs.
- Drug-target interaction data may not include all known targets.

**Common Beginner Mistakes:** Confusing DrugCentral (drug information) with DrugBank (more comprehensive drug database) or ChEMBL (drug-target binding data).

**When to Use It:** Use DrugCentral for drug repurposing, for finding drug targets and indications, or for exploring drug mechanisms.

**When NOT to Use It:** For comprehensive drug-target binding data, use ChEMBL or BindingDB. For pharmacogenomics, use PharmGKB.

**Related Databases or Alternatives:** DrugBank (comprehensive drug database), ChEMBL (drug-target binding), PharmGKB (pharmacogenomics), Open Targets (target prioritization).

**How It Connects to Other Resources:** DrugCentral integrates with FDA drug labels, UniProt (targets), and other drug databases.

**API / FTP / Bulk Download / Programmatic Access:** DrugCentral REST API at <https://drugcentral.org/api>. Returns JSON.

**Evidence or Curation Level:** Curated from FDA labels and published literature.

**Update Status:** Regularly updated; actively maintained by University of New Mexico.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Ursu O et al. (2019). DrugCentral 2018: an update. *Nucleic Acids Research*, 47(D1):D963–D970. doi:10.1093/nar/gky963

**Beginner-Friendly Explanation:** DrugCentral is a database of approved drugs that tells you what targets a drug binds to, what diseases it is approved to treat, and how it works. It is particularly useful for drug repurposing — finding new uses for existing drugs. For example, if you discover that a gene is important in a disease, you can search DrugCentral to find out if there are already approved drugs that target that gene.

**Advanced Technical Explanation:** DrugCentral integrates drug-target interaction data (with binding affinities where available), FDA-approved indications (from structured product labels), mechanisms of action, pharmacokinetics, and adverse effects. The database is particularly strong for FDA-approved drugs. DrugCentral is useful for drug repurposing analyses and for identifying off-target effects.

**Practical Workflow Example:** Step 1: Navigate to <https://drugcentral.org>. Step 2: Search for your drug or target. Step 3: Review drug-target interactions and indications. Step 4: Use the API for programmatic access.

**Reproducibility Notes:** Record the DrugCentral version and access date.

**Quality-Control Notes:** Check the evidence source for each drug-target interaction. Verify that the drug is approved for the indication of interest.

## Short Index Entries — Category AI

### IUPHAR/BPS Guide to Pharmacology

---

**Resource Type:** Knowledgebase (Pharmacology)

**Domain:** Pharmacology / Drug targets / Receptors

**Main Purpose:** Expert-curated database of pharmacological targets (receptors, ion channels, enzymes, transporters) and their ligands, providing quantitative pharmacological data and nomenclature.

**Best Used For:** Finding pharmacological data for receptors and drug targets; receptor nomenclature; quantitative pharmacology data.

**Key Limitation:** Coverage is primarily for well-characterized pharmacological targets; less comprehensive for novel targets.

**Related Resources:** DrugBank (drug information), ChEMBL (drug-target binding), Open Targets (target prioritization)

**Access / Licensing:** Open access; freely available at <https://www.guidetopharmacology.org>.

**Citation / Documentation:** Alexander SP et al. (2023). The Concise Guide to Pharmacology 2023/24. British Journal of Pharmacology, 180(S2):S1–S1440. doi:10.1111/bph.16176

### DrugBank

---

**Resource Type:** Database (Drug Information) — PARTIALLY RESTRICTED

**Domain:** Drug discovery / Pharmacology / Clinical medicine

**Main Purpose:** Comprehensive drug database combining detailed drug data with drug target information, covering approved, experimental, and illicit drugs with chemical, pharmacological, and clinical data.

**Best Used For:** Comprehensive drug information; drug-target interactions; drug metabolism; drug-drug interactions.

**Key Limitation:** Full database requires academic or commercial license. Free tier provides limited access. Some data may be outdated.

**Related Resources:** ChEMBL (open-access drug-target binding), DrugCentral (open-access drug information), PharmGKB (pharmacogenomics)

**Access / Licensing:** PARTIALLY RESTRICTED: Free tier available at <https://www.drugbank.com>; full access requires license.

**Citation / Documentation:** Wishart DS et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1):D1074–D1082. doi:10.1093/nar/gkx1037

## ChEMBL

---

**Resource Type:** Database (Drug-Target Binding)

**Domain:** Drug discovery / Medicinal chemistry / Pharmacology

**Main Purpose:** Large-scale bioactivity database containing drug-like molecules and their biological activities against protein targets, extracted from published medicinal chemistry literature.

**Best Used For:** Drug-target binding data; bioactivity data for drug discovery; QSAR modeling; target identification.

**Key Limitation:** Data is extracted from literature; quality depends on original publications. Not all bioactivity data is directly comparable across assays.

**Related Resources:** PubChem (chemical data), DrugBank (drug information), Open Targets (uses ChEMBL data), BindingDB (binding data)

**Access / Licensing:** Open access; freely available at <https://www.ebi.ac.uk/chembl>. CC BY-SA 3.0 license.

**Citation / Documentation:** Zdrazil B et al. (2024). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192. doi:10.1093/nar/gkad1004

## BindingDB

---

**Resource Type:** Database (Drug-Target Binding)

**Domain:** Drug discovery / Structural biology / Pharmacology

**Main Purpose:** Database of measured binding affinities (K<sub>d</sub>, K<sub>i</sub>, IC<sub>50</sub>, EC<sub>50</sub>) between proteins and drug-like molecules, extracted from published literature.

**Best Used For:** Quantitative binding affinity data; structure-activity relationship (SAR) analysis; target identification.

**Key Limitation:** Data is extracted from literature; assay conditions vary. Not all binding data is directly comparable.

**Related Resources:** ChEMBL (broader bioactivity data), PubChem (chemical data), DrugBank (drug information)

**Access / Licensing:** Open access; freely available at <https://www.bindingdb.org>.

**Citation / Documentation:** Gilson MK et al. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053. doi:10.1093/nar/gkv1072

## Category AJ: Cancer Model Systems and Cell Line Resources

### Category Overview

Cancer model system resources provide genomic, transcriptomic, and pharmacological data from cancer cell lines, patient-derived models, and clinical cohorts. These resources are essential for understanding cancer biology, identifying biomarkers, and developing cancer therapeutics.

### Critical distinctions:

- Cell line pharmacogenomics (DepMap, CCLE, GDSC): Genomic and drug sensitivity data from cancer cell lines. Useful for identifying biomarkers of drug response.
- Cell line registry (Cellosaurus): Authoritative registry of cell lines. Use for standardized cell line identification.
- Clinical cancer genomics (AACR GENIE, ICGC/ARGO): Genomic data from cancer patients. More clinically relevant than cell line data.

**WARNING: Cancer cell lines are imperfect models of human tumors. Drug sensitivity data from cell lines does not always translate to clinical efficacy. Always validate cell line findings in more clinically relevant models (patient-derived organoids, PDX models, clinical data) before drawing clinical conclusions.**

## AJ1 — DepMap (Cancer Dependency Map)

---

**Official Website URL:** <https://depmap.org>

**Resource Type:** Dataset Collection / Platform (Cancer Genomics)

**Main Biological Domain:** Cancer biology / Functional genomics / Drug discovery

**Short Definition:** The Cancer Dependency Map (DepMap) is a large-scale project that systematically identifies genetic dependencies and drug sensitivities in cancer cell lines using CRISPR screens, RNAi screens, and drug sensitivity assays, enabling identification of cancer vulnerabilities.

**What It Is Used For:** DepMap is used to identify genetic dependencies (essential genes) in cancer cell lines, to find biomarkers of drug sensitivity, and to prioritize cancer targets based on genetic dependency data.

**What Data It Contains:** DepMap contains CRISPR (Cas9) and RNAi gene dependency scores for over 1,000 cancer cell lines, drug sensitivity data (PRISM), genomic profiles (mutations, copy number, expression), and proteomics data for hundreds of cell lines.

**Main Scientific Question It Helps Answer:** Is this gene essential for cancer cell survival, and in which cancer types?

**Typical Users:** Cancer biologists; drug discovery researchers; bioinformaticians.

**Example Scientific Questions:**

- Is this gene a selective dependency in a specific cancer type?
- What are the biomarkers of sensitivity to this drug?
- What genes are co-dependencies with this gene?

**Example Use Cases:**

- Identifying cancer-selective dependencies for drug target prioritization.
- Finding biomarkers of drug sensitivity for patient stratification.
- Exploring synthetic lethal interactions in cancer.

**Input Data Accepted:** Gene names, drug names, cell line names, cancer type.

**Output Data Provided:** Gene dependency scores, drug sensitivity data, biomarker associations.

**Strengths:** Largest systematic cancer dependency dataset available; Integrates CRISPR, RNAi, drug sensitivity, and genomic data; Enables identification of cancer-selective dependencies; Freely accessible with bulk download.

**Limitations:** Cell lines are imperfect models of human tumors; CRISPR and RNAi screens have different biases and off-target effects; Drug sensitivity in cell lines does not always translate to clinical efficacy; Coverage of rare cancer types may be limited.

**Common Beginner Mistakes:** Treating cell line dependency data as direct evidence for clinical target validity; Not distinguishing between CRISPR (Chronos) and RNAi (DEMETER2) dependency scores; Confusing DepMap (dependency/drug sensitivity) with CCLE (genomic profiles only).

**When to Use It:** Use DepMap when identifying cancer dependencies, when finding biomarkers of drug sensitivity, or when prioritizing cancer targets based on functional genomics data.

**When NOT to Use It:** Do not use DepMap as a substitute for clinical evidence. For clinical cancer genomics, use TCGA, AACR GENIE, or ICGC/ARGO.

**Related Databases or Alternatives:** CCLE (genomic profiles, integrated with DepMap), GDSC (drug sensitivity, European), Cellosaurus (cell line registry), TCGA (clinical cancer genomics), Open Targets (target prioritization).

**How It Connects to Other Resources:** DepMap integrates CCLE genomic data, PRISM drug sensitivity data, and links to Cellosaurus for cell line identification.

**API / FTP / Bulk Download / Programmatic Access:** DepMap data available for download at <https://depmap.org/portal/download/>. R package depmap available.

**Evidence or Curation Level:** Experimental data from cancer cell lines; CRISPR and RNAi screens.

**Update Status:** Quarterly releases; actively maintained by Broad Institute.

**Licensing or Access Restrictions:** Open access; CC BY 4.0.

**Citation / Recommended Reference:** Tsherniak A et al. (2017). Defining a Cancer Dependency Map. Cell, 170(3):564–576.e16. doi:10.1016/j.cell.2017.06.010

**Beginner-Friendly Explanation:** DepMap (Cancer Dependency Map) is a project that systematically tests which genes are essential for the survival of different cancer cell lines. Using CRISPR technology, researchers knock out each gene one at a time in hundreds of cancer cell lines and measure whether the cells survive. Genes that are essential for cancer cell survival but not normal cells are called 'dependencies' and are potential drug targets. DepMap also tests hundreds of drugs against these cell lines to find which drugs kill which cancer types.

**Advanced Technical Explanation:** DepMap uses genome-scale CRISPR (Cas9) screens to generate gene dependency scores (Chronos algorithm) for over 1,000 cancer cell lines. RNAi data (DEMETER2 algorithm) is also available for comparison. Drug sensitivity data is generated using the PRISM multiplexed cell viability assay. Genomic profiles (mutations, copy number, expression, methylation) from CCLE are integrated with dependency data. The depmap R package provides programmatic access to all DepMap data.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://depmap.org/portal>.

Step 2: Search for your gene of interest.

Step 3: Review dependency scores across cancer types.

Step 4: Identify biomarkers of dependency using the correlation analysis. Step 5: Download data using the depmap R package.

**Reproducibility Notes:** Record the DepMap release version (e.g., 24Q2). Note the dependency score type (Chronos for CRISPR, DEMETER2 for RNAi). Record the cell lines and cancer types analyzed.

**Quality-Control Notes:** Check for cell line quality flags. Distinguish between pan-essential genes (essential in all cell lines) and selective dependencies. Validate key findings in independent datasets.



## AJ2 — GDSC (Genomics of Drug Sensitivity in Cancer)

**Official Website URL:** <https://www.cancerrxgene.org>

**Resource Type:** Dataset Collection (Cancer Pharmacogenomics)

**Main Biological Domain:** Cancer pharmacogenomics / Drug discovery

**Short Definition:** The Genomics of Drug Sensitivity in Cancer (GDSC) database provides drug sensitivity data for hundreds of anti-cancer compounds tested against a large panel of cancer cell lines, integrated with genomic profiles for biomarker discovery.

**What It Is Used For:** GDSC is used to find drug sensitivity data for cancer cell lines, to identify genomic biomarkers of drug response, and to compare drug sensitivity across cancer types.

**What Data It Contains:** GDSC contains drug sensitivity data (IC50, AUC) for over 500 compounds tested against over 1,000 cancer cell lines, integrated with genomic profiles (mutations, copy number, expression, methylation).

**Main Scientific Question It Helps Answer:** What is the sensitivity of this cancer cell line to this drug, and what genomic features predict sensitivity?

**Typical Users:** Cancer pharmacogenomicists; drug discovery researchers; bioinformaticians.

### Example Scientific Questions:

- What is the IC50 of this drug in this cancer cell line?
- What genomic features predict sensitivity to this drug?
- Which cancer types are most sensitive to this drug?

### Example Use Cases:

- Identifying biomarkers of drug sensitivity for patient stratification.
- Comparing drug sensitivity across cancer types.
- Drug repurposing based on genomic biomarkers.

**Input Data Accepted:** Drug names, cell line names, cancer types, gene names.

**Output Data Provided:** Drug sensitivity data (IC50, AUC), biomarker associations, genomic profiles.

**Strengths:** Large-scale drug sensitivity data with genomic integration; Enables biomarker discovery for drug response; Freely accessible with bulk download; Complementary to DepMap (US) with European perspective.

**Limitations:** Cell lines are imperfect models of human tumors; IC50 values are not directly comparable across different assay conditions; Drug sensitivity in cell lines does not always translate to clinical efficacy.

**Common Beginner Mistakes:** Comparing IC50 values across different assay conditions without normalization; Confusing GDSC (European, drug sensitivity) with DepMap (US, dependency + drug sensitivity).

**When to Use It:** Use GDSC for drug sensitivity data, for biomarker discovery, or for comparing drug sensitivity across cancer types.

**When NOT to Use It:** Do not use GDSC as a substitute for clinical evidence. For genetic dependencies, use DepMap.

**Related Databases or Alternatives:** DepMap (dependency + drug sensitivity, US), CCLE (genomic profiles), Cellosaurus (cell line registry), TCGA (clinical cancer genomics).

**How It Connects to Other Resources:** GDSC integrates with CCLE genomic data and Cellosaurus for cell line identification.

**API / FTP / Bulk Download / Programmatic Access:** GDSC data available for download at <https://www.cancerrxgene.org/downloads>. R package `gdscIC50` available.

**Evidence or Curation Level:** Experimental data from cancer cell lines.

**Update Status:** Regularly updated; actively maintained by Wellcome Sanger Institute.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Yang W et al. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961. doi:10.1093/nar/gks1111

**Beginner-Friendly Explanation:** GDSC (Genomics of Drug Sensitivity in Cancer) is a database that tests hundreds of anti-cancer drugs against hundreds of cancer cell lines and records how sensitive each cell line is to each drug. By combining this drug sensitivity data with genomic information about the cell lines (mutations, gene expression, etc.), researchers can identify which genomic features predict whether a cancer will respond to a particular drug. This is useful for developing precision medicine approaches to cancer treatment.

**Advanced Technical Explanation:** GDSC provides drug sensitivity data (IC50, AUC) for 500+ compounds across 1,000+ cancer cell lines, integrated with genomic profiles from CCLE. The database supports biomarker discovery through elastic net regression and other statistical methods. GDSC is complementary to DepMap (Broad Institute, US) and uses similar but not identical cell line panels and assay conditions.

**Practical Workflow Example:** Step 1: Navigate to <https://www.cancerrxgene.org>. Step 2: Search for your drug or cell line. Step 3: Review drug sensitivity data and biomarker associations. Step 4: Download data for further analysis.

**Reproducibility Notes:** Record the GDSC version (GDSC1 or GDSC2) and access date. Note the drug and cell line identifiers used.

**Quality-Control Notes:** Check for cell line quality flags. Note that GDSC1 and GDSC2 use different assay conditions. Validate key findings in independent datasets.

## AJ3 — Cellosaurus

---

**Official Website URL:** <https://www.cellosaurus.org>

**Resource Type:** Registry / Knowledgebase (Cell Lines)

**Main Biological Domain:** Cell biology / Cancer biology / Biomedical research

**Short Definition:** Cellosaurus is a comprehensive knowledge resource on cell lines, providing standardized identifiers, nomenclature, and metadata for over 150,000 cell lines from all organisms.

**What It Is Used For:** Cellosaurus is used to find standardized identifiers for cell lines, to check cell line authentication status, to find information about cell line origin and characteristics, and to resolve cell line name ambiguities.

**What Data It Contains:** Cellosaurus contains metadata for over 150,000 cell lines including standardized names, synonyms, accession numbers, species of origin, tissue of origin, disease, authentication status, cross-references to other databases, and references.

**Main Scientific Question It Helps Answer:** What is the standardized identifier for this cell line, and is it authenticated?

**Typical Users:** Cell biologists; cancer researchers; bioinformaticians; journal editors.

**Example Scientific Questions:**

- What is the Cellosaurus accession for HeLa cells?
- Is this cell line authenticated?
- What is the tissue of origin for this cell line?
- Are there known misidentified or contaminated cell lines in my study?

**Example Use Cases:** Standardizing cell line identifiers in publications; Checking cell line authentication status; Resolving cell line name ambiguities; Finding cross-references to DepMap, CCLE, and other databases.

**Input Data Accepted:** Cell line names, Cellosaurus accession numbers.

**Output Data Provided:** Standardized cell line metadata, authentication status, cross-references.

**Strengths:** Most comprehensive cell line registry available; Provides standardized identifiers for cell lines; Includes authentication status and contamination information; Cross-references to DepMap, CCLE, GDSC, and other databases; Freely accessible.

**Limitations:** Not all cell lines have authentication data; Coverage of very new or obscure cell lines may be incomplete.

**Common Beginner Mistakes:** Not checking Cellosaurus for cell line authentication status before using a cell line; Using non-standardized cell line names in publications — use Cellosaurus accessions.

**When to Use It:** Always use Cellosaurus to verify cell line identity and authentication status. Use Cellosaurus accessions for standardized cell line identification in publications.

**When NOT to Use It:** Cellosaurus is a registry, not a data resource — for genomic or drug sensitivity data, use DepMap, CCLE, or GDSC.

**Related Databases or Alternatives:** DepMap (cancer dependencies), CCLE (genomic profiles), GDSC (drug sensitivity), ATCC (cell line supplier).

**How It Connects to Other Resources:** Cellosaurus cross-references DepMap, CCLE, GDSC, ATCC, DSMZ, and many other cell line databases.

**API / FTP / Bulk Download / Programmatic Access:** Cellosaurus REST API at <https://api.cellosaurus.org/>. Returns JSON. Bulk download available.

**Evidence or Curation Level:** Manually curated from published literature and cell line suppliers.

**Update Status:** Regularly updated; actively maintained by SIB Swiss Institute of Bioinformatics.

**Licensing or Access Restrictions:** Open access; CC BY 4.0.

**Citation / Recommended Reference:** Bairoch A. (2018). The Cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques*, 29(2):25–38. doi:10.7171/jbt.18-2902-002

**Beginner-Friendly Explanation:** Cellosaurus is the authoritative registry for cell lines. It provides a standardized name and unique identifier (accession number) for over 150,000 cell lines. This is important because the same cell line may be known by many different names in different laboratories and publications. Cellosaurus also tracks which cell lines have been authenticated (confirmed to be what they claim to be) and which have been found to be misidentified or contaminated — a surprisingly common problem in cell biology research.

**Advanced Technical Explanation:** Cellosaurus uses a structured data model with standardized fields for cell line name, synonyms, accession number, species, tissue, disease, sex, age, authentication status, and cross-references. The database tracks cell line contamination and misidentification issues, which are a significant problem in biomedical research (estimated 15-20% of cell lines in use are misidentified). Cellosaurus accessions (e.g., CVCL\_0030 for HeLa) should be used in publications for unambiguous cell line identification.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://www.cellosaurus.org>.

Step 2: Search for your cell line by name.

Step 3: Verify the cell line identity and authentication status.

Step 4: Record the Cellosaurus accession for use in publications. Step 5: Check cross-references to DepMap, CCLE, and GDSC.

**Reproducibility Notes:** Always record the Cellosaurus accession number for each cell line used. Note the authentication status.

**Quality-Control Notes:** Check authentication status before using a cell line. Verify that the cell line is not listed as misidentified or contaminated.

## Short Index Entries — Category AJ

### CCLE (Cancer Cell Line Encyclopedia)

---

**Resource Type:** Dataset Collection (Cancer Genomics)

**Domain:** Cancer genomics / Pharmacogenomics

**Main Purpose:** Comprehensive genomic characterization of over 1,000 cancer cell lines, providing mutation, copy number, expression, methylation, and proteomics data. Now integrated with DepMap.

**Best Used For:** Genomic profiling of cancer cell lines; biomarker discovery; integration with DepMap dependency data.

**Key Limitation:** Cell lines are imperfect models of human tumors. CCLE is now integrated into DepMap; access through DepMap portal.

**Related Resources:** DepMap (dependency data, integrates CCLE), GDSC (drug sensitivity), Cellosaurus (cell line registry)

**Access / Licensing:** Open access; available through DepMap portal at <https://depmap.org/portal/download/>.

**Citation / Documentation:** Ghandi M et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, 569(7757):503–508. doi:10.1038/s41586-019-1186-3

### AACR Project GENIE

---

**Resource Type:** Dataset Collection (Clinical Cancer Genomics)

**Domain:** Clinical cancer genomics / Precision oncology

**Main Purpose:** Large-scale, real-world clinical cancer genomics dataset from multiple cancer centers, providing genomic data from cancer patients with clinical outcomes.

**Best Used For:** Real-world clinical cancer genomics; biomarker discovery with clinical outcomes; large-scale somatic mutation analysis.

**Key Limitation:** Access requires registration and data use agreement. Clinical data completeness varies by contributing center.

**Related Resources:** TCGA (research-grade cancer genomics), ICGC/ARGO (international cancer genomics), cBioPortal (data portal)

**Access / Licensing:** RESTRICTED: Registration required; data use agreement required. Available at <https://www.aacr.org/professionals/research/aacr-project-genie/>.

**Citation / Documentation:** AACR Project GENIE Consortium. (2017). AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discovery*, 7(8):818–831. doi:10.1158/2159-8290.CD-17-0151

## ICGC/ARGO (International Cancer Genome Consortium / Accelerating Research in Genomic Oncology)

---

**Resource Type:** Dataset Collection (Clinical Cancer Genomics) — CONTROLLED ACCESS

**Domain:** Clinical cancer genomics / Precision oncology

**Main Purpose:** International consortium providing high-quality genomic and clinical data from cancer patients across multiple cancer types and countries.

**Best Used For:** Large-scale international cancer genomics; cross-cancer type comparison; clinical outcome data.

**Key Limitation:** Controlled access; requires data access application. Data access may take weeks to months.

**Related Resources:** TCGA (US cancer genomics), AACR GENIE (real-world data), EGA (controlled-access archive), cBioPortal (data portal)

**Access / Licensing:** CONTROLLED ACCESS: Data access application required. Available at <https://daco.icgc-argo.org>.

**Citation / Documentation:** Zhang J et al. (2019). International Cancer Genome Consortium Data Portal — a one-stop shop for cancer genomics data. Database, 2019:baz026. doi:10.1093/database/baz026

## Category AK: Non-coding RNA Databases

### Category Overview

Non-coding RNA (ncRNA) databases provide sequence, annotation, and functional data for the diverse classes of RNA molecules that do not encode proteins. This category covers resources for microRNAs, long non-coding RNAs, ribosomal RNAs, transfer RNAs, and other ncRNA classes.

### Critical distinctions:

- Comprehensive ncRNA registry (RNAcentral): Unified access to sequences from all ncRNA databases. Use as the primary identifier resource.
- RNA family databases (Rfam): Sequence families and secondary structures for all RNA types. Use for RNA family annotation.
- miRNA-specific (miRBase, miRGeneDB): Curated miRNA annotations. miRGeneDB is more stringent in quality filtering.
- lncRNA-specific (LNCipedia, NONCODE): Long non-coding RNA annotations. Coverage and quality vary.

**WARNING: Non-coding RNA annotation is rapidly evolving. Many lncRNA annotations are based on computational prediction and lack experimental validation. miRNA annotations in miRBase include many low-confidence entries; miRGeneDB provides a more stringent, high-confidence set. Always check the evidence level for ncRNA annotations.**



## AK1 — RNAcentral

---

**Official Website URL:** <https://rnacentral.org>

**Resource Type:** Database / Registry (Non-coding RNA)

**Main Biological Domain:** Non-coding RNA / Transcriptomics / Bioinformatics

**Short Definition:** RNAcentral is a comprehensive database of non-coding RNA sequences that integrates data from over 40 expert databases, providing a unified identifier system and sequence repository for all types of non-coding RNA.

**What It Is Used For:** RNAcentral is used to find non-coding RNA sequences, to obtain unified identifiers for ncRNAs, to search for ncRNAs by sequence or annotation, and to access cross-database ncRNA information.

**What Data It Contains:** RNAcentral integrates ncRNA sequences from over 40 databases including miRBase, Rfam, Ensembl, GENCODE, tRNADB, snoDB, and others. Provides unified identifiers (URS IDs), sequences, secondary structures, and cross-references.

**Main Scientific Question It Helps Answer:** What is the sequence and annotation for this non-coding RNA, and what databases contain information about it?

**Typical Users:** RNA biologists; bioinformaticians; researchers working with ncRNA data.

**Example Scientific Questions:**

- What is the RNAcentral identifier for this miRNA?
- What databases contain information about this lncRNA?
- What is the sequence of this ncRNA?

**Example Use Cases:**

- Obtaining unified identifiers for ncRNAs across databases.
- Searching for ncRNAs by sequence similarity.
- Accessing cross-database ncRNA annotations.

**Input Data Accepted:** RNA sequences, ncRNA names, database accessions.

**Output Data Provided:** Unified ncRNA identifiers (URS IDs), sequences, cross-references, secondary structures.

**Strengths:** Unified identifier system for ncRNAs across 40+ databases; Comprehensive coverage of all ncRNA types; Sequence search and cross-database integration; Freely accessible with API.

**Limitations:** RNAcentral is an integrator, not a primary annotation resource; Quality of annotations depends on source databases; Not all ncRNAs have functional annotations.

**Common Beginner Mistakes:** Confusing RNAcentral (integrator) with primary ncRNA databases like miRBase or Rfam; Not using RNAcentral URS IDs for cross-database ncRNA identification.

**When to Use It:** Use RNAcentral as the primary resource for ncRNA sequence lookup and cross-database integration. Use RNAcentral URS IDs for standardized ncRNA identification.

**When NOT to Use It:** For detailed miRNA annotations, use miRBase or miRGeneDB. For RNA family annotations, use Rfam.

**Related Databases or Alternatives:** miRBase (miRNA), Rfam (RNA families), miRGeneDB (high-confidence miRNA), LNCipedia (lncRNA), Ensembl (ncRNA annotation).

**How It Connects to Other Resources:** RNAcentral integrates miRBase, Rfam, Ensembl, GENCODE, tRNAdb, snoDB, and 40+ other databases.

**API / FTP / Bulk Download / Programmatic Access:** RNAcentral REST API at <https://rnacentral.org/api/v1/>. Returns JSON. R package rnacentral available.

**Evidence or Curation Level:** Integrated from multiple expert databases; quality depends on source.

**Update Status:** Regularly updated; actively maintained by EMBL-EBI.

**Licensing or Access Restrictions:** Open access; CC0.

**Citation / Recommended Reference:** RNAcentral Consortium. (2021). RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220. doi:10.1093/nar/gkaa921

**Beginner-Friendly Explanation:** RNAcentral is a database that brings together information about non-coding RNAs from over 40 different databases into one place. Non-coding RNAs are RNA molecules that do not make proteins but have important functions in the cell, including microRNAs (which regulate gene expression), ribosomal RNAs (which make proteins), and long non-coding RNAs (which have diverse regulatory functions). RNAcentral gives each ncRNA a unique identifier so you can track it across different databases.

**Advanced Technical Explanation:** RNAcentral uses a Universal RNA Sequence (URS) identifier system that provides stable, unique identifiers for ncRNA sequences. The database integrates sequences from 40+ expert databases, providing cross-references, secondary structures (from Rfam), and functional annotations. The RNAcentral API supports sequence search (using nhmmer), identifier lookup, and cross-reference retrieval.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://rnacentral.org>.

Step 2: Search for your ncRNA by name or sequence.

Step 3: Find the URS identifier and cross-references.

Step 4: Use the API for programmatic access.

**Reproducibility Notes:** Record the RNAcentral version and the URS IDs used. Note the source databases for annotations.

**Quality-Control Notes:** Check the source database for each annotation. Verify that the ncRNA is the correct type and species.

## AK2 — Rfam

**Official Website URL:** <https://rfam.org>

**Resource Type:** Database (RNA Families)

**Main Biological Domain:** Non-coding RNA / RNA biology / Bioinformatics

**Short Definition:** Rfam is a database of RNA families, providing multiple sequence alignments, consensus secondary structures, and covariance models for all major classes of non-coding RNA and RNA structural elements.

**What It Is Used For:** Rfam is used to annotate RNA sequences with family classifications, to find RNA families in genomic sequences, and to access consensus secondary structures for RNA families.

**What Data It Contains:** Rfam contains over 4,000 RNA families with multiple sequence alignments, consensus secondary structures, and covariance models (CMs). Families cover all major ncRNA classes including rRNA, tRNA, miRNA, snRNA, snoRNA, lncRNA, riboswitches, and IRES elements.

**Main Scientific Question It Helps Answer:** What RNA family does this sequence belong to, and what is its consensus secondary structure?

**Typical Users:** RNA biologists; bioinformaticians; genome annotators.

**Example Scientific Questions:**

- What RNA family does this sequence belong to?
- What is the consensus secondary structure for this RNA family?
- What RNA families are present in this genomic region?

**Example Use Cases:**

- Annotating ncRNAs in a genome assembly.
- Finding RNA structural elements in sequences.
- Classifying novel RNA sequences into known families.

**Input Data Accepted:** RNA sequences, Rfam family IDs.

**Output Data Provided:** RNA family classifications, consensus secondary structures, covariance models.

**Strengths:** Comprehensive coverage of RNA families; Covariance models enable sensitive sequence-structure search; Used by Ensembl and GENCODE for ncRNA annotation; Freely accessible.

**Limitations:** Coverage of novel or poorly characterized RNA families may be incomplete; Covariance model searches require Infernal software.

**Common Beginner Mistakes:** Confusing Rfam (RNA families) with miRBase (miRNA-specific) or RNACentral (ncRNA integrator).

**When to Use It:** Use Rfam for RNA family annotation, for finding RNA structural elements, or for classifying novel RNA sequences.

**When NOT to Use It:** For miRNA-specific annotations, use miRBase or miRGeneDB. For comprehensive ncRNA lookup, use RNACentral.

**Related Databases or Alternatives:** RNACentral (ncRNA integrator), miRBase (miRNA), Ensembl (uses Rfam for annotation), Infernal (software for Rfam searches).

**How It Connects to Other Resources:** Rfam is integrated with RNACentral and used by Ensembl and GENCODE for ncRNA annotation.

**API / FTP / Bulk Download / Programmatic Access:** Rfam REST API at <https://rfam.org/api/>. Returns JSON. Bulk download available.

**Evidence or Curation Level:** Expert-curated RNA families with experimental and computational evidence.

**Update Status:** Regularly updated; actively maintained by EMBL-EBI.

**Licensing or Access Restrictions:** Open access; CC0.

**Citation / Recommended Reference:** Kalvari I et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1):D192–D200. doi:10.1093/nar/gkaa1047

**Beginner-Friendly Explanation:** Rfam is a database of RNA families — groups of RNA sequences that share a common evolutionary origin and secondary structure. Just as Pfam classifies protein domains into families, Rfam classifies RNA sequences into families. Each family has a consensus secondary structure (the typical folding pattern) and a covariance model that can be used to search for new members of the family in genomic sequences. Rfam is used by genome annotation pipelines to identify non-coding RNAs in newly sequenced genomes.

**Advanced Technical Explanation:** Rfam uses covariance models (CMs), implemented in the Infernal software package, to represent RNA families as probabilistic models of sequence and secondary structure. CMs are more sensitive than sequence-only models for finding divergent RNA family members. Rfam families cover rRNA, tRNA, miRNA, snRNA, snoRNA, lncRNA, riboswitches, IRES elements, and other RNA structural elements. Rfam is used by Ensembl, GENCODE, and other genome annotation pipelines.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://rfam.org>.

Step 2: Search for your RNA sequence or family.

Step 3: Download the covariance model for your family.

Step 4: Use Infernal (cmsearch) to search for family members in genomic sequences.

**Reproducibility Notes:** Record the Rfam version and the family IDs used. Note the Infernal version and search parameters.

**Quality-Control Notes:** Check the bit score threshold for CM searches. Verify that the family assignment is appropriate for your RNA type.

## AK3 — miRBase

**Official Website URL:** <https://www.mirbase.org>

**Resource Type:** Database (microRNA)

**Main Biological Domain:** Non-coding RNA / microRNA / Gene regulation

**Short Definition:** miRBase is the primary repository for microRNA (miRNA) sequence data and annotation, providing a standardized nomenclature, sequence data, and genomic coordinates for miRNAs across species.

**hat It Is Used For:** miRBase is used to find miRNA sequences and annotations, to obtain standardized miRNA names and identifiers, and to access miRNA genomic coordinates.

**What Data It Contains:** miRBase contains over 38,000 miRNA entries from over 270 species (as of release 22), including precursor and mature miRNA sequences, genomic coordinates, and cross-references.

**Main Scientific Question It Helps Answer:** What is the sequence and annotation for this microRNA?

**Typical Users:** RNA biologists; miRNA researchers; bioinformaticians.

**Example Scientific Questions:**

- What is the sequence of hsa-miR-21-5p?
- What is the genomic location of this miRNA?
- What miRNAs are encoded in this genomic region?

**Example Use Cases:**

- Obtaining miRNA sequences for experimental design.
- Annotating miRNA expression data.
- Finding miRNA genomic coordinates for ChIP-seq analysis.

**Input Data Accepted:** miRNA names, miRBase accessions, genomic coordinates.

**Output Data Provided:** miRNA sequences, genomic coordinates, cross-references.

**Strengths:**

- Primary repository for miRNA sequences; comprehensive coverage.
- Standardized miRNA nomenclature.
- Freely accessible.

**Limitations:**

- miRBase contains many low-confidence entries, particularly in later releases.
- Not all miRBase entries have experimental validation.
- For high-confidence miRNA annotations, use miRGeneDB instead.

**Common Beginner Mistakes:**

- Assuming all miRBase entries are experimentally validated — many are low-confidence.
- Not distinguishing between precursor (mir) and mature (miR) miRNA sequences.
- Using outdated miRBase releases — nomenclature and accessions may change between releases.

**When to Use It:** Use miRBase for miRNA sequence lookup and standardized nomenclature. For high-confidence miRNA annotations, use miRGeneDB.

**When NOT to Use It:** Do not use miRBase as the sole source for miRNA functional annotations — many entries lack experimental validation.

**Related Databases or Alternatives:** miRGeneDB (high-confidence miRNA), RNAcentral (ncRNA integrator), Rfam (RNA families), TargetScan (miRNA target prediction).

**How It Connects to Other Resources:** miRBase is integrated with RNAcentral and Rfam. Cross-references to Ensembl and NCBI.

**API / FTP / Bulk Download / Programmatic Access:** miRBase FTP download at <https://www.mirbase.org/ftp.shtml>. REST API limited.

**Evidence or Curation Level:** Varies; some entries are experimentally validated, others are computationally predicted.

**Update Status:** Last major update: Release 22 (2018). Updates have been infrequent since then.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Kozomara A et al. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162. doi:10.1093/nar/gky1141

**Beginner-Friendly Explanation:** miRBase is the main database for microRNAs (miRNAs). MicroRNAs are small RNA molecules (~22 nucleotides) that regulate gene expression by binding to messenger RNAs and preventing them from being translated into proteins. miRBase provides the sequences and names for all known miRNAs. The naming convention uses the species prefix (e.g., hsa for human), followed by miR and a number (e.g., hsa-miR-21-5p for human microRNA 21, mature form from the 5' arm).

**Advanced Technical Explanation:** miRBase uses a standardized nomenclature where precursor miRNAs are named with lowercase 'mir' (e.g., hsa-mir-21) and mature miRNAs with uppercase 'miR' (e.g., hsa-miR-21-5p, hsa-miR-21-3p). Each entry includes the precursor sequence, mature sequences, genomic coordinates, and cross-references. Note that miRBase release 22 (2018) was the last major update; for high-confidence annotations, miRGeneDB is preferred.

**Practical Workflow Example:** Step 1: Navigate to <https://www.mirbase.org>. Step 2: Search for your miRNA by name or accession. Step 3: Download sequences in FASTA format. Step 4: Use miRBase accessions for standardized annotation.

**Reproducibility Notes:** Record the miRBase release version. Note whether precursor or mature sequences are used.

**Quality-Control Notes:** Check the confidence level of miRNA annotations. For high-confidence annotations, cross-reference with miRGeneDB.

## Short Index Entries — Category AK

### miRGeneDB

**Resource Type:** Database (microRNA — High Confidence)

**Domain:** Non-coding RNA / microRNA

**Main Purpose:** High-confidence microRNA gene database providing a curated, filtered set of miRNA annotations based on stringent criteria for miRNA biogenesis and conservation.

**Best Used For:** High-confidence miRNA annotations; miRNA evolutionary analysis; filtering low-confidence miRBase entries.

**Key Limitation:** More restrictive than miRBase — some miRNAs in miRBase are not in miRGeneDB. Coverage is smaller but higher quality.

**Related Resources:** miRBase (comprehensive but lower confidence), RNAcentral (ncRNA integrator), Rfam (RNA families)

**Access / Licensing:** Open access; freely available at <https://mirgenedb.org>.

**Citation / Documentation:** Fromm B et al. (2020). MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Research*, 48(D1):D132–D141. doi:10.1093/nar/gkz885

### LNCipedia

**Resource Type:** Database (Long Non-coding RNA)

**Domain:** Non-coding RNA / lncRNA / Gene regulation

**Main Purpose:** Comprehensive database of human long non-coding RNA (lncRNA) transcripts, providing sequences, annotations, and secondary structure predictions.

**Best Used For:** Human lncRNA sequence lookup; lncRNA annotation; lncRNA secondary structure prediction.

**Key Limitation:** Many lncRNA annotations are computationally predicted; experimental validation is limited for most entries. lncRNA biology is rapidly evolving.

**Related Resources:** NONCODE (alternative lncRNA database), RNAcentral (ncRNA integrator), Ensembl (lncRNA annotation), GENCODE (lncRNA annotation)

**Access / Licensing:** Open access; freely available at <https://lncipedia.org>.

**Citation / Documentation:** Volders PJ et al. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Research*, 47(D1):D135–D139. doi:10.1093/nar/gky1031

### circBase

**Resource Type:** Database (Circular RNA)

**Domain:** Non-coding RNA / Circular RNA

**Main Purpose:** Database of circular RNAs (circRNAs) identified from RNA-seq data, providing sequences, genomic coordinates, and cross-species conservation information.





**Best Used For:** Circular RNA annotation; circRNA sequence lookup; cross-species circRNA comparison.

**Key Limitation:** circRNA annotation is challenging; many entries may be artifacts. Experimental validation is limited for most entries.

**Related Resources:** RNAcentral (ncRNA integrator), miRBase (miRNA), LNCipedia (lncRNA)

**Access / Licensing:** Open access; freely available at <http://www.circbase.org>.

**Citation / Documentation:** Glažar P et al. (2014). circBase: a database for circular RNAs. *RNA*, 20(11):1666–1670. doi:10.1261/rna.043687.113

## snoDB

---

**Resource Type:** Database (Small Nucleolar RNA)

**Domain:** Non-coding RNA / snoRNA

**Main Purpose:** Comprehensive database of human small nucleolar RNAs (snoRNAs), providing sequences, annotations, targets, and expression data.

**Best Used For:** Human snoRNA annotation; snoRNA target identification; snoRNA expression analysis.

**Key Limitation:** Coverage is primarily human; other species are less well covered.

**Related Resources:** RNAcentral (ncRNA integrator), Rfam (RNA families including snoRNA families)

**Access / Licensing:** Open access; freely available at <https://snodb.usherbrooke.ca>.

**Citation / Documentation:** Bouchard-Bourdon J et al. (2023). snoDB 2.0: an interactive database, specializing in human snoRNAs. *Nucleic Acids Research*, 51(D1):D291–D296. doi:10.1093/nar/gkac835

## Category AL: Comparative Genomics and Orthology Resources

### Category Overview

---

Comparative genomics and orthology resources provide the infrastructure for understanding gene evolution, identifying orthologous genes across species, and studying the evolutionary relationships between organisms. These resources are essential for functional annotation transfer, evolutionary analysis, and cross-species genomic comparison.

### Critical distinctions:

---

- Orthology databases (OrthoDB, OMA, eggNOG): Databases of orthologous gene groups. Differ in methodology and scope.
- Phylogenomics (PANTHER, TimeTree): Resources for phylogenetic analysis and evolutionary timing.
- Functional annotation transfer: Using orthology to transfer functional annotations from well-studied species to less-studied species. Requires careful consideration of orthology quality.

**WARNING: Orthology inference is not trivial. Different orthology databases may give different results for the same gene pair. One-to-one orthologs are more dependable for functional annotation transfer than one-to-many or many-to-many relationships. Always check the orthology type (1:1, 1:N, N:N) and the confidence score before transferring functional annotations.**

## AL1 — OrthoDB

---

**Official Website URL:** <https://www.orthodb.org>

**Resource Type:** Database (Orthology)

**Main Biological Domain:** Comparative genomics / Evolutionary biology / Functional genomics

**Short Definition:** OrthoDB is a hierarchical catalog of orthologs for all sequenced organisms, providing orthologous gene groups at multiple taxonomic levels from the universal root to species-specific groups.

**What It Is Used For:** OrthoDB is used to find orthologs of a gene across species, to assess gene conservation, to transfer functional annotations across species, and to study gene family evolution.

**What Data It Contains:** OrthoDB contains orthologous gene groups for over 5,000 species across all domains of life, organized hierarchically at multiple taxonomic levels. Each ortholog group includes gene members, functional annotations, and evolutionary statistics.

**Main Scientific Question It Helps Answer:** What are the orthologs of this gene across species, and how conserved is it?

**Typical Users:** Evolutionary biologists; comparative genomicists; bioinformaticians; genome annotators.

**Example Scientific Questions:**

- What are the orthologs of this human gene in mouse, zebrafish, and Drosophila?
- How conserved is this gene across vertebrates?
- What is the evolutionary history of this gene family?

**Example Use Cases:**

- Transferring functional annotations from human to model organisms.
- Assessing gene conservation for functional importance.
- Studying gene family evolution across the tree of life.

**Input Data Accepted:** Gene names, Ensembl IDs, UniProt IDs, species names.

**Output Data Provided:** Orthologous gene groups, gene members, functional annotations, evolutionary statistics.

**Strengths:** Hierarchical orthology at multiple taxonomic levels; Comprehensive coverage of all sequenced organisms; Used by BUSCO for genome assembly quality assessment; Freely accessible with API.

**Limitations:** Orthology inference is computationally intensive; some assignments may be incorrect; Coverage of poorly sequenced organisms may be incomplete; Different orthology databases may give different results.

**Common Beginner Mistakes:**

- Assuming all OrthoDB orthologs are one-to-one — many are one-to-many or many-to-many.
- Not checking the taxonomic level of the ortholog group.
- Confusing orthologs (same gene in different species) with paralogs (duplicated genes in the same species).

**When to Use It:** Use OrthoDB for ortholog identification, for functional annotation transfer, or for assessing gene conservation. OrthoDB is also used by BUSCO for genome quality assessment.

**When NOT to Use It:** For detailed phylogenetic analysis, use PANTHER or TimeTree. For protein family classification, use Pfam or InterPro.

**Related Databases or Alternatives:** OMA (alternative orthology), eggNOG (functional orthology), PANTHER (phylogenomics), Ensembl Compara (vertebrate orthology), BUSCO (uses OrthoDB).

**How It Connects to Other Resources:** OrthoDB integrates with Ensembl, UniProt, and NCBI Gene. Used by BUSCO for genome quality assessment.

**API / FTP / Bulk Download / Programmatic Access:** OrthoDB REST API at <https://www.orthodb.org/v10/>. Returns JSON.

**Evidence or Curation Level:** Computationally inferred using graph-based clustering of best reciprocal hits.

**Update Status:** Regularly updated; actively maintained by SIB Swiss Institute of Bioinformatics.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Kuznetsov D et al. (2023). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, 51(D1):D445–D451. doi:10.1093/nar/gkac998

**Beginner-Friendly Explanation:** OrthoDB is a database of orthologs — genes in different species that evolved from the same ancestral gene. For example, the human TP53 gene and the mouse Trp53 gene are orthologs because they both evolved from the same gene in the common ancestor of humans and mice. OrthoDB organizes orthologs hierarchically, so you can find orthologs at different levels of evolutionary distance (e.g., all vertebrates, all animals, all eukaryotes). This is useful for understanding how conserved a gene is and for transferring functional knowledge from well-studied species to less-studied ones.

**Advanced Technical Explanation:** OrthoDB uses a graph-based clustering algorithm to identify orthologous gene groups at multiple taxonomic levels. The hierarchical structure allows users to find orthologs at the appropriate evolutionary level for their analysis. OrthoDB is used by BUSCO (Benchmarking Universal Single-Copy Orthologs) for genome assembly quality assessment, where the presence of conserved single-copy orthologs is used as a proxy for genome completeness.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://www.orthodb.org>.

Step 2: Search for your gene by name or identifier.

Step 3: Select the taxonomic level of interest.

Step 4: Review the ortholog group members. Step 5: Download the ortholog group for further analysis.

**Reproducibility Notes:** Record the OrthoDB version and the taxonomic level used. Note the orthology type (1:1, 1:N, N:N) for each gene pair.

**Quality-Control Notes:** Check the orthology type — 1:1 orthologs are most reliable for functional annotation transfer. Verify the taxonomic level of the ortholog group.

## AL2 — eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups)

**Official Website URL:** <http://eggnog5.embl.de>

**Resource Type:** Database (Orthology / Functional Annotation)

**Main Biological Domain:** Comparative genomics / Functional genomics / Evolutionary biology

**Short Definition:** eggNOG is a database of orthologous groups and functional annotations for all sequenced organisms, providing hierarchical orthologous groups (HOGs) with functional annotations transferred from well-characterized species.

**What It Is Used For:** eggNOG is used for functional annotation of genes through orthology, for identifying orthologous groups, and for comparative genomics analysis. The eggNOG-mapper tool enables rapid functional annotation of new genomes.

**What Data It Contains:** eggNOG contains orthologous groups for over 5,000 organisms, with functional annotations (GO terms, KEGG pathways, COG categories) transferred from well-characterized species. eggNOG-mapper provides rapid annotation of query sequences.

**Main Scientific Question It Helps Answer:** What is the function of this gene based on its orthologous group, and what are its orthologs?

**Typical Users:** Genome annotators; comparative genomicists; bioinformaticians.

**Example Scientific Questions:**

- What is the functional annotation of this gene based on eggNOG?
- What COG category does this gene belong to?
- What are the orthologs of this gene in eggNOG?

**Example Use Cases:**

- Rapid functional annotation of a new genome using eggNOG-mapper.
- Assigning COG categories to genes for comparative genomics.
- Transferring GO annotations through orthology.

**Input Data Accepted:** Protein sequences, gene names, Ensembl IDs.

**Output Data Provided:** Orthologous groups, functional annotations (GO, KEGG, COG), eggNOG IDs.

**Strengths:** Functional annotations transferred through orthology; eggNOG-mapper enables rapid annotation of new genomes; COG categories for prokaryotic and eukaryotic genes; Freely accessible.

**Limitations:** Functional annotations are transferred through orthology; accuracy depends on orthology quality; Coverage of poorly characterized organisms may be limited.

**Common Beginner Mistakes:** Treating eggNOG functional annotations as experimentally validated — they are transferred through orthology; Confusing eggNOG (orthology + function) with OrthoDB (orthology only).

**When to Use It:** Use eggNOG for rapid functional annotation of new genomes, for COG category assignment, or for orthology-based functional annotation.

**When NOT to Use It:** For high-confidence functional annotations, use experimentally validated databases (UniProt/Swiss-Prot, GO with experimental evidence codes).

**Related Databases or Alternatives:** OrthoDB (orthology), PANTHER (phylogenomics), COG (clusters of orthologous groups), GO (gene ontology), KEGG (pathways).

**How It Connects to Other Resources:** eggNOG integrates with GO, KEGG, COG, and other functional annotation resources.

**API / FTP / Bulk Download / Programmatic Access:** eggNOG-mapper web server at <https://eggnog-mapper.embl.de/>. Command-line tool available for local installation.

**Evidence or Curation Level:** Computationally inferred through orthology; functional annotations transferred from well-characterized species.

**Update Status:** Regularly updated; actively maintained by EMBL.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Cantalapiedra CP et al. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, 38(12):5825–5829. doi:10.1093/molbev/msab293

**Beginner-Friendly Explanation:** eggNOG is a database that groups genes from different species into orthologous groups and assigns functional annotations to each group. The key feature of eggNOG is the eggNOG-mapper tool, which allows you to quickly annotate the genes in a new genome by finding their orthologous groups and transferring the functional annotations. This is particularly useful for newly sequenced organisms where experimental functional data is limited.

**Advanced Technical Explanation:** eggNOG uses a hierarchical orthologous group (HOG) structure similar to OrthoDB. The eggNOG-mapper tool uses DIAMOND for fast sequence search followed by HMM-based refinement to assign query sequences to eggNOG orthologous groups. Functional annotations (GO terms, KEGG pathways, COG categories, BRITE hierarchies) are transferred from the orthologous group. eggNOG-mapper v2 also supports metagenomic annotation.

#### **Practical Workflow Example:**

Step 1: Submit your protein sequences to eggNOG-mapper at <https://eggnog-mapper.embl.de/>.

Step 2: Review the orthologous group assignments and functional annotations.

Step 3: Download the annotation table.

Step 4: Use GO terms and KEGG pathways for downstream analysis.

## AL3 — PANTHER (Protein ANALYSIS THrough Evolutionary Relationships)

**Official Website URL:** <https://www.pantherdb.org>

**Resource Type:** Database / Tool (Phylogenomics / Functional Classification)

**Main Biological Domain:** Comparative genomics / Evolutionary biology / Functional genomics

**Short Definition:** PANTHER is a database and tool for classifying proteins and genes by their evolutionary relationships and biological functions, providing phylogenetic trees, ortholog identification, and statistical overrepresentation tests for GO enrichment analysis.

**What It Is Used For:** PANTHER is used for GO enrichment analysis, for protein family classification, for ortholog identification, and for evolutionary analysis of gene families.

**What Data It Contains:** PANTHER contains phylogenetic trees for over 15,000 protein families, with functional annotations (GO terms, PANTHER pathways) and ortholog assignments for genes from major model organisms.

**Main Scientific Question It Helps Answer:** What protein family does this gene belong to, and what are its evolutionary relationships?

**Typical Users:** Bioinformaticians; evolutionary biologists; researchers performing GO enrichment analysis.

**Example Scientific Questions:**

- What PANTHER family does this gene belong to?
- What GO terms are overrepresented in this gene list?
- What are the orthologs of this gene in PANTHER?

**Example Use Cases:**

- GO enrichment analysis using PANTHER's statistical overrepresentation test.
- Protein family classification.
- Evolutionary analysis of gene families.

**Input Data Accepted:** Gene lists, protein sequences, gene names.

**Output Data Provided:** GO enrichment results, protein family classifications, phylogenetic trees.

**Strengths:** Widely used for GO enrichment analysis; Phylogenetic trees for protein families; Integrates with GO for functional analysis; Freely accessible.

**Limitations:** GO enrichment results depend on the reference gene list used; Protein family classifications may differ from other databases.

**Common Beginner Mistakes:** Not specifying the correct reference gene list for GO enrichment analysis; Confusing PANTHER (phylogenomics + GO enrichment) with OrthoDB or eggNOG (orthology databases).

**When to Use It:** Use PANTHER for GO enrichment analysis, for protein family classification, or for evolutionary analysis of gene families.

**When NOT to Use It:** For comprehensive orthology databases, use OrthoDB or OMA. For detailed phylogenetic analysis, use dedicated phylogenetics tools.

**Related Databases or Alternatives:** GO (gene ontology), OrthoDB (orthology), eggNOG (functional orthology), TimeTree (species phylogeny).



**How It Connects to Other Resources:** PANTHER integrates with GO and is used by the GO Consortium for GO enrichment analysis.

**API / FTP / Bulk Download / Programmatic Access:** PANTHER REST API at <https://www.pantherdb.org/services/oai/pantherdb/>. Returns JSON.

**Evidence or Curation Level:** Computationally inferred phylogenetic trees; functional annotations from GO.

**Update Status:** Regularly updated; actively maintained by USC.

**Licensing or Access Restrictions:** Open access; freely available.

**Citation / Recommended Reference:** Mi H et al. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, 49(D1):D394–D403. doi:10.1093/nar/gkaa1106

**Beginner-Friendly Explanation:** PANTHER is a database and tool that classifies genes into protein families based on their evolutionary relationships. It is widely used for GO enrichment analysis — a statistical test that identifies which biological processes, molecular functions, or cellular components are overrepresented in a list of genes. For example, if you have a list of genes that are upregulated in cancer, PANTHER can tell you which biological processes are enriched in that list.

**Advanced Technical Explanation:** PANTHER uses maximum likelihood phylogenetic trees to classify proteins into families and subfamilies. The PANTHER overrepresentation test uses a binomial or Fisher's exact test to identify GO terms that are statistically overrepresented in a gene list compared to a reference list. PANTHER is the recommended tool for GO enrichment analysis by the GO Consortium. The PANTHER API supports programmatic access to enrichment analysis.

#### **Practical Workflow Example:**

Step 1: Navigate to <https://www.pantherdb.org>.

Step 2: Enter your gene list.

Step 3: Select the organism and reference list.

Step 4: Run the overrepresentation test.

Step 5: Review enriched GO terms and PANTHER pathways.

**Reproducibility Notes:** Record the PANTHER version and the reference gene list used. Note the statistical test and correction method.

**Quality-Control Notes:** Check that the reference gene list is appropriate for your analysis. Verify that the organism is correctly specified.

## Short Index Entries — Category AL

### OMA (Orthologous Matrix)

**Resource Type:** Database (Orthology)

**Domain:** Comparative genomics / Evolutionary biology

**Main Purpose:** Large-scale orthology database using a pairwise comparison approach to identify orthologs and paralogs across thousands of genomes.

**Best Used For:** Ortholog identification; comparative genomics; evolutionary analysis. OMA uses a different algorithm from OrthoDB and eggNOG, providing complementary results.

**Key Limitation:** Different orthology databases may give different results; OMA may be more conservative than OrthoDB for some gene families.

**Related Resources:** OrthoDB (hierarchical orthology), eggNOG (functional orthology), PANTHER (phylogenomics)

**Access / Licensing:** Open access; freely available at <https://omabrowser.org>.

**Citation / Documentation:** Altenhoff AM et al. (2021). OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Research*, 49(D1):D373–D379. doi:10.1093/nar/gkaa1007

### TimeTree

**Resource Type:** Database / Tool (Evolutionary Timescale)

**Domain:** Evolutionary biology / Phylogenomics

**Main Purpose:** Database and tool for estimating divergence times between species based on published molecular clock studies, providing a time-calibrated tree of life.

**Best Used For:** Estimating divergence times between species; time-calibrated phylogenetic analysis; evolutionary rate estimation.

**Key Limitation:** Divergence time estimates have uncertainty; different studies may give different estimates. Always report confidence intervals.

**Related Resources:** PANTHER (phylogenomics), OrthoDB (orthology), NCBI Taxonomy (species classification)

**Access / Licensing:** Open access; freely available at <http://www.timetree.org>.

**Citation / Documentation:** Kumar S et al. (2022). TimeTree 5: an expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8):msac174. doi:10.1093/molbev/msac174

### InParanoid

**Resource Type:** Database (Orthology)

**Domain:** Comparative genomics / Evolutionary biology

**Main Purpose:** Database of pairwise ortholog groups between species, using the InParanoid algorithm to identify orthologs and in-paralogs.

**Best Used For:** Pairwise ortholog identification; in-paralog analysis; comparative genomics.

**Key Limitation:** Pairwise approach; does not provide hierarchical ortholog groups. Less comprehensive than OrthoDB or OMA for multi-species analysis.

**Related Resources:** OrthoDB (hierarchical orthology), OMA (alternative orthology), eggNOG (functional orthology)

**Access / Licensing:** Open access; freely available at <https://inparanoid.sbc.su.se>.

**Citation / Documentation:** Sonnhammer EL & Östlund G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, 43(D1):D234–D239. doi:10.1093/nar/gku1203

## PhylomeDB

---

**Resource Type:** Database (Phylogenomics)

**Domain:** Comparative genomics / Evolutionary biology

**Main Purpose:** Database of complete gene phylogenies (phylomes) for multiple organisms, providing evolutionary analysis of gene families including duplication and speciation events.

**Best Used For:** Gene family evolution; duplication/speciation event analysis; phylogenetic tree visualization.

**Key Limitation:** Coverage is limited to organisms with complete phylomes; not all organisms are represented.

**Related Resources:** PANTHER (phylogenomics), OrthoDB (orthology), TimeTree (divergence times)

**Access / Licensing:** Open access; freely available at <https://phylomedb.org>.

**Citation / Documentation:** Fuentes D et al. (2022). PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Research*, 50(D1):D1062–D1068. doi:10.1093/nar/gkab966

## Part V: Practical Reproducibility and Quality-Control Guidelines for Database-Based Research

### Introductory Note

This part complements the earlier Database Release and Verification Policy. The earlier policy defines how database entries in this atlas are documented, verified, and cited. This part provides practical guidance for researchers on how to report, reproduce, audit, and quality-control database-based analyses in real research workflows. It is intended for use when writing methods sections, designing reproducible pipelines, reusing public datasets, checking database-derived results, or preparing publication-quality bioinformatics reports.

### Section 1: The Reproducibility Crisis in Bioinformatics Database Usage

A significant proportion of bioinformatics analyses are not reproducible due to inadequate documentation of database versions, access dates, and query parameters. This section provides a framework for reproducible database usage.

#### Key reproducibility failures in database-dependent analyses:

- Database version not recorded: Databases are updated continuously; results may differ between versions.
- Access date not recorded: For databases without versioning, the access date is the only way to identify the data used.
- Query parameters not documented: Different query parameters can produce substantially different results.
- Deprecated identifiers: Gene, protein, and variant identifiers may be deprecated or reassigned between database versions.
- Annotation version mismatch: Using different genome or annotation versions across tools in the same pipeline produces inconsistent results.

## Section 2: Minimum Reproducibility Requirements

---

The following minimum requirements should be met for any analysis that uses bioinformatics databases:

### 2.1 For All Database Queries

1. Record the database name, version (or release date), and URL.
2. Record the access date for databases without explicit versioning.
3. Document all query parameters, filters, and thresholds used.
4. Save a copy of the query results (or a representative sample) for verification.
5. Use stable identifiers (accession numbers, CURIEs) rather than names, which may change.

### 2.2 For Genome-Based Analyses

1. Specify the genome assembly version (e.g., GRCh38.p14, GRCm39).
2. Specify the gene annotation version (e.g., GENCODE v44, Ensembl 110, RefSeq GCF\_000001405.40).
3. Use the same genome assembly and annotation version throughout the analysis pipeline.
4. Document any coordinate conversions (e.g., liftOver from GRCh37 to GRCh38).

### 2.3 For Variant Analysis

1. Specify the genome build for all variant coordinates (GRCh37/hg19 or GRCh38/hg38).
2. Record the ClinVar version and submission date for clinical variant interpretations.
3. Record the gnomAD version for population frequency data.
4. Use rsIDs for common variants; use HGVS notation for clinical variants.

### 2.4 For Functional Annotation

1. Record the GO version and release date.
2. Specify the evidence codes included (experimental only, or including IEA).
3. Record the annotation source (UniProt, Ensembl, NCBI Gene).
4. For pathway analysis, record the pathway database version (KEGG, Reactome, WikiPathways).

## Section 3: Database-Specific QC Warnings

The following QC warnings apply to specific databases and should be checked before using data in analyses.

### 3.1 Sequence Databases

GenBank/ENA/DDBJ: Third-party sequence submissions are not peer-reviewed. Always verify sequences from primary literature before use in critical analyses. Check for retracted or superseded entries.

RefSeq: RefSeq sequences are curated but may include predicted sequences (XM\_, XP\_ prefixes) that are computationally generated. Prefer experimentally validated sequences (NM\_, NP\_ prefixes) for critical analyses.

UniProt/TrEMBL: TrEMBL entries are computationally annotated and not manually reviewed. For high-confidence functional annotations, use Swiss-Prot (reviewed) entries only.

### 3.2 Variant Databases

ClinVar contains submissions from multiple laboratories that may conflict. Always check for conflicting interpretations (Conflicting interpretations of pathogenicity). Do not use ClinVar classifications without reviewing the supporting evidence.

gnomAD population frequencies are from a specific set of cohorts and may not represent all populations equally. Allele frequencies may differ substantially between populations. Always check population-specific frequencies.

COSMIC contains both somatic mutations from cancer studies and germline variants. Ensure you are using the correct data type for your analysis. COSMIC Tier 1 genes have the strongest evidence for cancer causality.

### 3.3 Functional Annotation Databases

Gene Ontology (GO): IEA (Inferred from Electronic Annotation) annotations are computationally generated and less reliable than experimental annotations. For high-confidence analyses, filter to experimental evidence codes (IDA, IMP, IGI, IEP, IPI, IBA, IBD, IKR, IRD, ISS, ISO, ISA, ISM, IGC, RCA, TAS, NAS, IC, ND).

KEGG pathway annotations are manually curated but may not be comprehensive for all organisms. KEGG requires a subscription for bulk download; use Reactome or WikiPathways as open alternatives.

STRING protein interaction scores are composite scores that include both experimental and predicted interactions. Always check the evidence types supporting each interaction. High-confidence interactions (score > 0.7) are more reliable.

### 3.4 Clinical Databases

OMIM gene-disease relationships vary in evidence strength. Check the OMIM entry type (Phenotype, Gene, Phenotype series) and the evidence level. Not all OMIM entries represent confirmed Mendelian disease genes.

ClinGen gene-disease validity classifications are periodically updated. Always check the current classification and date. A 'Limited' classification means there is insufficient evidence, not that the gene does not cause the disease.

PharmGKB evidence levels range from 1A (highest, CPIC guideline) to 4 (lowest, case report). Only Level 1A and 1B annotations have clinical guidelines. Do not use lower-level annotations for clinical decision-making without expert review.

### 3.5 Structural Biology Databases

PDB structures vary widely in quality. Always check the resolution, R-factor, and validation report before using a structure for analysis. Structures with resolution > 3.5 Å may not be suitable for detailed structural analysis.

AlphaFold structures are computational predictions, not experimental structures. The pLDDT score indicates per-residue confidence; regions with pLDDT < 70 are low-confidence and should be interpreted cautiously. Do not cite AlphaFold structures as experimental evidence.

EMDB: Cryo-EM map resolution varies widely. Always check the FSC (Fourier Shell Correlation) curve and local resolution map. Maps with resolution > 4 Å may not support atomic model building.



## Section 4: Applying FAIR Principles in Database-Based Research

---

The FAIR principles (Findable, Accessible, Interoperable, Reusable) provide a framework for data management that supports reproducibility and reuse.

### 4.1 Findable

To make database-derived results findable, report stable identifiers wherever possible. These include database accession numbers, study IDs, gene IDs, transcript IDs, protein IDs, variant IDs, pathway IDs, ontology term IDs, publication DOIs, and dataset accession numbers. Avoid reporting only free-text names because names can change, merge, split, or become ambiguous across databases.

### 4.2 Accessible

To make database-derived results accessible, report the database URL, download page, API endpoint, access date, and access conditions. For controlled-access data, report the data access approval number, data access committee, and relevant usage restrictions. For bulk downloads, record the exact file name, download date, and archive or release location.

### 4.3 Interoperable

To make database-derived results interoperable, report the identifier system, genome build, annotation release, ontology version, file format, and mapping method used. Do not mix Ensembl, RefSeq, UniProt, HGNC, dbSNP, ClinVar, or ontology identifiers without recording how they were mapped. For cross-database integration, document the mapping table, tool, and date used.

### 4.4 Reusable

To make database-derived results reusable, report all filters, thresholds, preprocessing steps, excluded records, quality-control decisions, software versions, database releases, and licensing restrictions. A result is not reusable if another researcher cannot reconstruct the same query, retrieve the same data version, and understand why records were included or excluded.

## Common Mistakes and Misconceptions

### COMMON MISTAKES AND MISCONCEPTIONS IN BIOINFORMATICS DATABASE USE

The following is a list of 20 common mistakes that beginners make when using bioinformatics databases. For each mistake, the error is described, the reason it is problematic is explained, and the correct approach is provided.

#### **MISTAKE: Using outdated database versions or deprecated databases**

**Why is it wrong:** Bioinformatics databases are continuously updated with new data, corrected annotations, and improved algorithms. Using an outdated version can lead to incorrect or incomplete results. Some databases (e.g., DIP, MINT, RDP) have not been updated in years and may contain outdated information.

**Correct approach:** Always use the most current version of a database. Check the database website for the latest release date and version number. For deprecated databases, use the recommended replacement (e.g., use SILVA instead of RDP for ribosomal RNA sequences). When reporting results, always specify the database version used.

#### **MISTAKE: Confusing gene names across species**

**Why is it wrong:** The same gene name can refer to different genes in different organisms. For example, "p53" in humans (TP53) is different from "p53" in *Drosophila* (p53). Gene names are not standardized across species, and the same name can have different functions in different organisms.

**Correct approach:** Always specify the organism when referring to a gene. Use official gene symbols from the appropriate model organism database (FlyBase for *Drosophila*, WormBase for *C. elegans*, MGI for mouse, etc.). When translating findings across species, use ortholog databases (DIOPT, OrthoFinder, InParanoid) to identify the correct ortholog.

#### **MISTAKE: Not distinguishing between somatic mutations and germline variants**

**Why is it wrong:** Somatic mutations are acquired during a person's lifetime and are found only in specific cells (e.g., cancer cells). Germline variants are inherited and present in all cells. These two types of variation are stored in different databases (COSMIC for somatic mutations; ClinVar, gnomAD for germline variants) and have different clinical implications.

**Correct approach:** Use COSMIC for somatic mutation data in cancer. Use ClinVar for clinically significant germline variants. Use gnomAD for population-level germline variant frequencies. Never use gnomAD frequencies to assess the pathogenicity of somatic mutations.

#### **MISTAKE: Not accounting for database access restrictions**

**Why is it wrong?** Many databases have tiered access models, with some data freely available and other data requiring registration, institutional access, or commercial licensing. Attempting to access restricted data without authorization is not only ineffective but may violate terms of service.

**Correct approach:** Before using a database, check its access policy. For databases with tiered access (DrugBank, OncoKB, COSMIC), register for a free academic account if available. For controlled-access genomic data (TCGA, ICGC), apply for access through the appropriate data access committee (dbGaP, DACO).

### **MISTAKE: Not citing databases correctly**

**Why is it wrong:** Databases require citation just like any other scientific resource. Not citing databases makes it impossible for readers to verify your data sources, and it deprives database developers of the recognition they need to secure funding. Many journals require specific database citations.

**Correct approach:** Always cite the primary publication for each database you use. For most databases, the correct citation is the most recent NAR database issue paper. Use the NAR Database Collection to find the correct citation. Include the database version and access date in your methods section.

### **MISTAKE: Treating computationally predicted annotations as experimentally verified**

**Why is it wrong:** Many database annotations are computationally predicted rather than experimentally verified. For example, most UniProt/TrEMBL entries are computationally annotated, while UniProt/Swiss-Prot entries are manually reviewed. Using computationally predicted annotations as if they were experimentally verified can lead to incorrect conclusions.

**Correct approach:** Always check the evidence level for database annotations. In UniProt, use Swiss-Prot (manually reviewed) for high-confidence annotations. In GO, check the evidence codes (EXP, IDA, IPI for experimental; IEA for inferred from electronic annotation). In STRING, check the interaction score and evidence type.

### **MISTAKE: Not understanding the difference between sequence similarity and functional similarity**

**Why is it wrong?** Sequence similarity does not always imply functional similarity. Two proteins can have similar sequences but different functions (paralogs), or similar functions but different sequences (convergent evolution). Using sequence similarity alone to infer function can lead to incorrect annotations.

**Correct approach:** Use multiple lines of evidence to infer function, including sequence similarity, structural similarity, phylogenetic analysis, and experimental data. When using BLAST to find homologs, check the alignment quality (E-value, percent identity, query coverage) and verify the function of the top hits experimentally.

### **MISTAKE: Using the wrong BLAST program for the query type**

**Why it is wrong:** NCBI BLAST has multiple programs for different query and database types: BLASTn (nucleotide vs. nucleotide), BLASTp (protein vs. protein), BLASTx (translated nucleotide vs. protein), tBLASTn (protein vs. translated nucleotide), and tBLASTx (translated nucleotide vs. translated nucleotide). Using the wrong program can give incorrect or no results.

**Correct approach:** Use BLASTp for protein queries against protein databases. Use BLASTn for nucleotide queries against nucleotide databases. Use BLASTx to find protein homologs of a nucleotide sequence. Use tBLASTn to find nucleotide sequences encoding a protein of interest.

### **MISTAKE: Not filtering BLAST results by E-value and percent identity**

**Why is it wrong:** BLAST returns all hits above a threshold, including many that are not biologically meaningful. Without filtering by E-value and percent identity, researchers may include spurious hits in their analyses.

**Correct approach:** Use an E-value threshold of  $1e-5$  or lower for most analyses. For functional annotation, use a percent identity threshold of at least 30% (and ideally 50% or higher) and a query coverage of at least 70%. For phylogenetic analysis, use more stringent thresholds.

### **MISTAKE: Confusing gene expression databases with protein databases**

**Why is it wrong:** Gene expression databases (GEO, ArrayExpress) contain mRNA expression data, while protein databases (UniProt, PDB) contain protein sequence and structure data. mRNA expression does not always correlate with protein abundance due to post-transcriptional regulation.

**Correct approach:** Use GEO or ArrayExpress for gene expression data. Use UniProt for protein sequence and function data. Use PDB for protein structure data. For proteomics data, use PRIDE or ProteomicsDB. Remember that mRNA and protein levels may not correlate.

### **MISTAKE: Not using controlled vocabularies and ontologies for data annotation**

**Why it is wrong:** Using free-text descriptions for biological concepts makes data difficult to search, integrate, and compare. Without controlled vocabularies, the same concept may be described in many ways, making it impossible to find all relevant data.

**Correct approach:** Use controlled vocabularies and ontologies for data annotation. Use GO for gene function, HPO for human phenotypes, DO for diseases, UBERON for anatomy, and MeSH for medical concepts. When submitting data to databases, use the recommended ontology terms.

**MISTAKE: Not checking for database redundancy**

**Why is it wrong:** Many databases contain overlapping data, and the same information may be available in multiple databases. Using multiple databases without checking for redundancy can lead to double-counting and inflated results.

**Correct approach:** Understand the relationships between databases before using them. For example, UniProt integrates data from Swiss-Prot and TrEMBL; Ensembl and NCBI Gene both provide genome annotations; STRING integrates data from multiple interaction databases. Choose the most appropriate database for your analysis and be aware of data overlapping.

**MISTAKE: Not validating database downloads**

**Why is it wrong:** Database downloads can be corrupted during transfer, and some databases provide checksums (MD5, SHA256) to verify file integrity. Using corrupted data can lead to incorrect results.

**Correct approach:** Always verify the integrity of downloaded files using the checksums provided by the database. Use md5sum or sha256sum to verify file integrity. If a checksum is not provided, compare the file size with the expected size.

**MISTAKE: Not considering database coverage biases**

**Why is it wrong:** Databases are not uniformly comprehensive across all organisms, tissues, or conditions. Well-studied organisms (human, mouse, yeast) and tissues (blood, brain) are better represented than less-studied ones. This coverage bias can affect the interpretation of results.

**Correct approach:** Be aware of coverage biases in the databases you use. When analyzing data from less-studied organisms or tissues, consider that the absence of data may reflect a lack of study rather than a true biological absence. Use model organism databases for well-studied organisms and general databases for less-studied ones.

**MISTAKE: Not using the appropriate identifier system**

**Why is it wrong:** Different databases use different identifier systems for genes, proteins, and other biological entities. Using the wrong identifier can lead to incorrect data retrieval or failed database queries.

**Correct approach:** Use the identifier system appropriate for each database. For NCBI databases, use Entrez Gene IDs or RefSeq accessions. For Ensembl, use Ensembl gene IDs (ENSG). For UniProt, use UniProt accessions (P12345). Use identifier mapping tools (BioMart, UniProt ID mapping, DAVID) to convert between identifier systems.

### **MISTAKE: Not understanding the difference between primary and secondary databases**

**Why is it wrong:** Primary databases contain original experimental data (GenBank, PDB, GEO), while secondary databases contain derived or curated data (UniProt, Pfam, KEGG). Secondary databases may lag primary databases in incorporating new data.

**Correct approach:** Use primary databases for the most current data. Use secondary databases for curated, integrated information. When using secondary databases, check the data source and update frequency to ensure the data is current.

### **MISTAKE: Ignoring data provenance and evidence codes**

**Why is it wrong:** Database annotations vary in quality and reliability. Annotations based on experimental evidence are more reliable than those based on computational prediction or sequence similarity. Ignoring evidence codes can lead to incorrect conclusions.

**Correct approach:** Always check the evidence codes for database annotations. In GO, use experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP) for high-confidence annotations. In UniProt, use Swiss-Prot (manually reviewed) for high-confidence annotations. In STRING, check the interaction score and evidence type.

### **MISTAKE: Not using programmatic access for large-scale analyses**

**Why is it wrong:** Manual database queries are not scalable for large-scale analyses. Attempting to manually retrieve data for thousands of genes or proteins is time-consuming and error prone.

**Correct approach:** Use programmatic access (APIs, FTP downloads, R/Python packages) for large-scale analyses. Most major databases provide REST APIs, FTP downloads, and language-specific packages. Use Bioconductor packages (biomaRt, rentrez, UniProt.ws) for R-based analyses and Biopython for Python-based analyses.

### **MISTAKE: Not keeping up with database updates**

**Why it is wrong:** Databases are continuously updated with new data, corrected annotations, and improved algorithms. Analyses performed with an older version of a database may give different results than analyses performed with the current version.

**Correct approach:** Subscribe to database newsletters or RSS feeds to stay informed about updates. When repeating analyses, use the same database version for consistency. When reporting results, always specify the database version and access date.

**MISTAKE: Not understanding the limitations of pathway databases**

**Why it is wrong:** Pathway databases (KEGG, Reactome, WikiPathways) represent biological pathways as static diagrams, but biological pathways are dynamic and context-dependent. The same pathway may behave differently in different cell types, tissues, or conditions.

**Correct approach:** Use pathway databases as a starting point for analysis, not as a definitive representation of biology. Validate pathway analysis results with experimental data. Consider using multiple pathway databases, as they may have different coverage and curation approaches. Be aware that pathway enrichment analysis results depend on the gene set used and the statistical method applied.



## Recommended Learning Path

### RECOMMENDED LEARNING PATH FOR BIOINFORMATICS DATABASES

This section provides structured learning paths for researchers at three different levels of experience. Each path is designed to build competency progressively, starting with foundational concepts and moving toward advanced applications.

**LEVEL A: COMPLETE BEGINNERS (No bioinformatics background)** Target audience: Undergraduate students, researchers from non-computational fields, clinicians new to genomics.

**Goal:** Develop a basic understanding of bioinformatics databases and the ability to perform simple database queries.

- **Step 1: Understand the biological context**

Before using bioinformatics databases, develop a basic understanding of molecular biology: DNA, RNA, proteins, genes, and genomes. Resources: Khan Academy Biology, NCBI's "A Science Primer" (<https://www.ncbi.nlm.nih.gov/books/NBK20363/>). Start with NCBI Gene (<https://www.ncbi.nlm.nih.gov/gene>) to look up genes you are familiar with.

- **Step 2: Learn to use NCBI databases**

NCBI provides a suite of interconnected databases that are ideal for beginners. Start with PubMed (<https://pubmed.ncbi.nlm.nih.gov>) for literature searching, then explore NCBI Gene for gene information, and NCBI Nucleotide for sequence data. Practice searching for genes you know (e.g., BRCA1, TP53) and reading the database records.

- **Step 3: Explore UniProt for protein information**

Navigate to UniProt (<https://www.uniprot.org>) and search for a protein of interest. Read the Swiss-Prot entry for a well-characterized protein (e.g., human p53, P04637) and understand the different sections: function, subcellular location, disease associations, and sequence. Learn to distinguish between Swiss-Prot (manually reviewed) and TrEMBL (computationally annotated) entries.

- **Step 4: Use BLAST for sequence similarity searching**

Navigate to NCBI BLAST (<https://blast.ncbi.nlm.nih.gov>) and perform a simple BLASTp search with a protein sequence. Learn to interpret the results: E-value, percent identity, query coverage, and alignment. Practice finding homologs of a protein in different organisms.

- **Step 5: Explore a disease database**



Navigate to OMIM (<https://www.omim.org>) and search for a disease you are familiar with. Read the disease entry and understand the gene-disease associations. Then check ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar>) for variants associated with the disease.

- **Step 6: Learn about gene ontology**

Navigate to the Gene Ontology website (<https://geneontology.org>) and search for a gene of interest. Understand the three GO aspects: Molecular Function, Biological Process, and Cellular Component. Learn to interpret GO annotations and evidence codes.

- **Step 7: Practice with a complete workflow**

Choose a gene of interest and perform a complete workflow: find the gene in NCBI Gene, look up the protein in UniProt, find homologs using BLAST, check disease associations in OMIM and ClinVar, and annotate the gene with GO terms. Document your findings and the databases you used.

## LEVEL B: BIOLOGY STUDENTS WITH SOME COMPUTATIONAL EXPOSURE

**Target audience:** Graduate students in biology, researchers with basic programming skills, wet-lab scientists learning bioinformatics.

**Goal:** Develop proficiency in using major bioinformatics databases for research, including programmatic access.

**Step 1: Master the major sequence databases:** Develop proficiency in using NCBI databases (Gene, Nucleotide, Protein, SRA), Ensembl (<https://www.ensembl.org>), and UniProt. Learn to use BioMart (<https://www.ensembl.org/biomart>) for bulk data retrieval. Practice downloading gene lists, sequences, and annotations for downstream analysis.

**Step 2: Learn programmatic database access:** Learn to use the NCBI E-utilities API (<https://www.ncbi.nlm.nih.gov/books/NBK25497/>) for programmatic data retrieval. Install and use Biopython (<https://biopython.org>) for Python-based database access. Learn to use the biomaRt R package for Ensembl data retrieval. Practice retrieving data for a list of genes programmatically.

**Step 3: Explore protein structure databases:** Navigate to the PDB (<https://www.rcsb.org>) and explore protein structures. Learn to use PyMOL or UCSF ChimeraX for structure visualization. Explore AlphaFold DB (<https://alphafold.ebi.ac.uk>) for predicted structures. Understand the relationship between sequence, structure, and function.

**Step 4: Use pathway and interaction databases:** Explore KEGG (<https://www.kegg.jp>) for pathway analysis. Use STRING (<https://string-db.org>) for protein interaction networks. Learn to perform pathway enrichment analysis using DAVID (<https://david.ncifcrf.gov>) or g:Profiler (<https://biit.cs.ut.ee/gprofiler>). Practice interpreting pathway analysis results.

**Step 5: Analyze gene expression data:** Navigate to GEO (<https://www.ncbi.nlm.nih.gov/geo>) and find a dataset relevant to your research. Download the data and perform a basic differential expression analysis using DESeq2 or edgeR in R. Interpret the results using GO and pathway databases.

**Step 6: Explore variant databases:** Learn to use gnomAD (<https://gnomad.broadinstitute.org>) for population variant frequencies. Use ClinVar for clinical variant interpretation. Practice interpreting variant data using ACMG/AMP guidelines. Understand the difference between pathogenic, likely pathogenic, and variants of uncertain significance.

**Step 7: Develop a data management strategy:** Learn about FAIR principles and data management. Use FAIRsharing (<https://fairsharing.org>) to find appropriate repositories for your data. Register samples in BioSamples and projects in BioProject. Practice depositing data in a public repository (GEO, ArrayExpress, SRA).

## LEVEL C: EXPERIENCED RESEARCHERS NEW TO A SPECIFIC DOMAIN

**Target audience:** Established researchers moving into a new area.

**Goal:** Rapidly develop domain-specific database expertise & integrate databases into existing workflows.

**Step 1: Identify the key databases for the new domain:** Use NAR Database Collection (<https://www.nucleicacidsresearch.com/database-issue>) & bio.tools (<https://bio.tools>) to identify primary databases for the new domain. Read the NAR papers for the top 5-10 databases to understand their content, methods & limitations. Consult FAIRsharing for data standards and deposition requirements.

**Step 2: Understand the data types and formats:** Each domain has specific data types and formats. For single-cell genomics: AnnData (h5ad), Seurat objects (rds). For microbiome: FASTQ, BIOM, OTU tables. For structural biology: PDB format, mmCIF. Learn the standard formats for your new domain and the tools used to process them.

**Step 3: Access domain-specific databases:** For single-cell genomics: Human Cell Atlas (<https://www.humancellatlas.org>), CellxGene (<https://cellxgene.cziscience.com>), Single Cell Expression Atlas (<https://www.ebi.ac.uk/gxa/sc>). For microbiome: MGnify (<https://www.ebi.ac.uk/metagenomics>), SILVA (<https://www.arb-silva.de>), GTDB (<https://gtdb.ecogenomic.org>). For cancer genomics: TCGA (via GDC), cBioPortal (<https://www.cbioportal.org>), COSMIC (<https://cancer.sanger.ac.uk/cosmic>). For epigenomics: ENCODE (<https://www.encodeproject.org>), JASPAR (<https://jaspar.elixir.no>).

**Step 4: Learn domain-specific analysis tools:** Each domain has standard analysis tools. For single-cell genomics: Scanpy (Python), Seurat (R). For microbiome: QIIME2, mothur. For cancer genomics: maftools (R), GATK. For epigenomics: deepTools, MACS2. Use bio.tools to find the standard tools for your domain.

**Step 5: Integrate new databases with existing workflows:** Identify how the new domain's databases connect to databases you already use. For example, cancer genomics databases connect to variant databases (ClinVar, gnomAD), pathway databases (KEGG, Reactome), and drug databases (DrugBank, ChEMBL). Use identifier mapping tools (BioMart, UniProt ID mapping) to integrate data across databases.

**Step 6: Engage with the community:** Join domain-specific communities and mailing lists. Attend workshops and training courses (EMBL-EBI training, Galaxy training). Read recent review articles and methods papers to understand current best practices. Contribute to database curation if possible.

**Step 7: Develop domain-specific expertise:** After gaining familiarity with the databases & tools, develop expertise in domain-specific analysis methods. For single-cell genomics: trajectory analysis, cell type annotation, spatial transcriptomics. For microbiome: diversity analysis, taxonomic classification, functional annotation. For cancer genomics: mutational signature analysis, copy number analysis, tumor evolution. Use the databases you have learned to perform original analyses.

## References and Further Reading

The following references include key papers for major databases covered in this atlas, methodology papers, and review articles on bioinformatics resources. All references are real and verifiable.

### PRIMARY DATABASE PAPERS

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any of 14,000+ organisms. *Nucleic Acids Research*, 51(D1):D638–D646. <https://doi.org/10.1093/nar/gkac1000>

Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200. <https://doi.org/10.1002/pro.3978>

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363. <https://doi.org/10.1093/nar/gkt1115>

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>

Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E, Magarinos MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Maranon M, Hunter F, Junco H, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940. <https://doi.org/10.1093/nar/gky1075>

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. (2023). PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380. <https://doi.org/10.1093/nar/gkac956>

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29. <https://doi.org/10.1038/75556>

Kohler S, Gargano M, Matentzoglou N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ, Chute CG, Est JL, Galer PD, Ganesan S, Griesse M, Haimel M, Pazmandi J, Hanauer M, Harris NL, Hartnell MJ, Hastreiter M, Hauck F, He Y, Jeske T, Kearney H, Kindle G, Klein C, Knoflach K, Krause R, Lagorce D, McMurry JA, Miller JA, Munoz-Torres MC, Peters RL, Rapp CK, Rath AM, Rind SA, Rosenberg AZ, Segal MM, Seidel MG, Smedley D, Talmy T, Thomas Y, Wiafe SA, Xian J, Yunsel O, Zeltner V, Schriml LM, Haendel MA, Valentini G, Robinson PN. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217. <https://doi.org/10.1093/nar/gkaa1043>

- ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74. <https://doi.org/10.1038/nature11247>
- Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Gerstein G, Precup D, Kim A, Malin JA, Bhatt DL, Bhatt DL, Bhatt DL. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330. <https://doi.org/10.1038/nature14248>
- Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Perez N, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, Mathelier A. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1):D165–D173. <https://doi.org/10.1093/nar/gkab1113>
- Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Gottgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klennerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N; Human Cell Atlas Meeting Participants. (2017). The Human Cell Atlas. *eLife*, 6:e27041. <https://doi.org/10.7554/eLife.27041>
- Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A, Shlemov A, Kunyavskaya O, Lapidus A, Finn RD. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, 48(D1):D570–D578. <https://doi.org/10.1093/nar/gkz1035>
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596. <https://doi.org/10.1093/nar/gks1219>
- Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794. <https://doi.org/10.1093/nar/gkab776>
- Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947. <https://doi.org/10.1093/nar/gky1015>
- Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, 2(5):401–404. <https://doi.org/10.1158/2159-8290.CD-12-0095>
- Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, Hantash FM, Grupe A, Baxi SS, Callahan MK, Snyder A, Chi P, Danila D, Gounder M, Harding JJ, Hellmann MD, Iyer G, Janjigian Y, Kaley T, Levine DA, Lowery M, Omuro A, Postow MA, Rathkopf D, Shoushtari AN, Shukla N, Voss M, Paraiso E, Zehir A, Berger MF, Taylor BS, Saltz LB, Riely GJ, Ladanyi M, Hyman DM, Baselga J, Sabbatini P, Solit DB, Schultz N. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, 1:1–16. <https://doi.org/10.1200/PO.17.00011>



- Larkin A, Marygold SJ, Antonazzo G, Attrill H, Dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, Tabone CJ, Thurmond J; FlyBase Consortium. (2021). FlyBase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Research*, 49(D1):D899–D907. <https://doi.org/10.1093/nar/gkaa1026>
- Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, Davis P, Gao S, Grove CA, Kishore R, Lee RYN, Muller HM, Nakamura C, Nuin P, Paulini M, Raciti D, Rodgers FH, Russell M, Schindelman G, Auken KV, Wang Q, Williams G, Wright AJ, Yook K, Howe KL, Schedl T, Stein L, Sternberg PW. (2020). WormBase: a modern Model Organism Information Resource. *Nucleic Acids Research*, 48(D1):D762–D767. <https://doi.org/10.1093/nar/gkz920>
- Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE; Mouse Genome Database Group. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, 47(D1):D801–D806. <https://doi.org/10.1093/nar/gky1056>
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. (2012). *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705. <https://doi.org/10.1093/nar/gkr1029>
- Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M; FAIRsharing Community. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4):358–367. <https://doi.org/10.1038/s41587-019-0080-8>
- Ison J, Kalaš M, Jonassen I, Bolser D, Uludag M, McWilliam H, Malone J, Lopez R, Pettifer S, Rice P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10):1325–1332. <https://doi.org/10.1093/bioinformatics/btt113>
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018. <https://doi.org/10.1038/sdata.2016.18>

## REVIEW ARTICLES AND METHODOLOGY PAPERS

- Rigden DJ and Fernandez XM. (2024). The 2024 *Nucleic Acids Research* database issue and the online molecular biology database collection. *Nucleic Acids Research*, 52(D1):D1–D9. <https://doi.org/10.1093/nar/gkad1173>
- Perez-Riverol Y, Bai J, Bandla C, Garcia-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A, Vizcaino JA. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50(D1):D543–D552. <https://doi.org/10.1093/nar/gkab1038>



Appendix: Master Comparison Table of All Databases

The following master table consolidates all databases covered in this atlas. Each row represents one database or resource. Columns provide a quick-reference summary.

Category A

Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
NCBI	<a href="https://www.ncbi.nlm.nih.gov">https://www.ncbi.nlm.nih.gov</a>	Portal	Omics (broad)	Omics (broad)	Yes	E-utilities REST API	Mixed (community to manually reviewed)	Interface complexity; variable data quality across databases	Researcher / Student
EMBL-EBI	<a href="https://www.ebi.ac.uk">https://www.ebi.ac.uk</a>	Portal	Omics (broad)	Omics (broad)	Moderate	Multiple REST APIs per database	Mixed (community to manually reviewed)	Interface complexity; some databases restructured/moved	Researcher / Student
DDBJ	<a href="https://www.ddbj.nig.ac.jp">https://www.ddbj.nig.ac.jp</a>	Portal/Database	DNA sequences / Omics	DNA sequences / Omics	Moderate (better for Japanese users)	Web API, FTP	Community-submitted (INSDC standard)	Less comprehensive English documentation; primarily Asia-Pacific focus	Researcher / Student
ExPASy	<a href="https://www.expasy.org">https://www.expasy.org</a>	Portal	Proteins / Systems biology	Proteins / Systems biology	Moderate	UniProt REST API, Swiss-Model API	Mixed (Swiss-Prot manually reviewed; TrEMBL computationally predicted)	Primarily protein-focused; less useful for nucleotide or literature queries	Researcher / Student
Ensembl	<a href="https://www.ensembl.org">https://www.ensembl.org</a>	Genome Browser/Database/Portal	DNA sequences / Variants / Transcriptomics	DNA sequences / Variants / Transcriptomics	Moderate	Ensembl REST API, BioMart, FTP	Computationally predicted (with manual curation for human/mouse via GENCODE)	Gene IDs change between releases; annotation quality varies by species	Researcher / Student

Category B

Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
PubMed	<a href="https://pubmed.ncbi.nlm.nih.gov">https://pubmed.ncbi.nlm.nih.gov</a>	Literature Search Engine	Free	Free	40M+ biomedical citations	Links to PMC/publishers	E-utilities (free)	Biomedical literature search	Researcher / Student
PubMed Central (PMC)	<a href="https://www.ncbi.nlm.nih.gov/pmc">https://www.ncbi.nlm.nih.gov/pmc</a>	Literature Repository	Free	Free	9M+ full-text articles (open access)	Yes (full text)	E-utilities (free)	Full-text access and text mining	Researcher / Student
Europe PMC	<a href="https://europepmc.org">https://europepmc.org</a>	Literature Search Engine/Repository	Free	Free	42M+ abstracts, 9M+ full text	Yes (open access)	REST API (free)	Preprints + published literature; European funding	Researcher / Student
Google Scholar	<a href="https://scholar.google.com">https://scholar.google.com</a>	Literature Search Engine	Free	Free	Broadest (all disciplines, preprints, theses)	Links to sources	None (no official API)	Broad discovery; citation tracking	Researcher / Student
Semantic Scholar	<a href="https://www.semanticscholar.org">https://www.semanticscholar.org</a>	Literature Search Engine	Free	Free	200M+ papers (all disciplines)	Links to sources	Academic Graph API (free)	AI-assisted discovery; programmatic access	Researcher / Student
Scopus	<a href="https://www.scopus.com">https://www.scopus.com</a>	Citation Database	COMMERCIAL (institutional)	COMMERCIAL (institutional)	90M+ records (all disciplines)	Links to publishers	Scopus API (institutional)	Systematic reviews; citation analysis	Researcher / Student
Web of Science	<a href="https://www.webofscience.com">https://www.webofscience.com</a>	Citation Database	COMMERCIAL (institutional)	COMMERCIAL (institutional)	21,000+ journals (all disciplines)	Links to publishers	WoS API (institutional)	Citation analysis; Journal Impact Factor	Researcher / Student



Category C

Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
GenBank	<a href="https://www.ncbi.nlm.nih.gov/genbank">https://www.ncbi.nlm.nih.gov/genbank</a>	Primary Database	Community-submitted	Community-submitted	All organisms, assembled sequences	No (use SRA)	E-utilities	Comprehensive sequence archive; submission	Researcher / Student
ENA	<a href="https://www.ebi.ac.uk/ena">https://www.ebi.ac.uk/ena</a>	Primary Database	Community-submitted	Community-submitted	All organisms, assembled + raw reads	Yes	ENA Portal API	European submission; raw read download	Researcher / Student
DDBJ	<a href="https://www.ddbj.nig.ac.jp">https://www.ddbj.nig.ac.jp</a>	Primary Database	Community-submitted	Community-submitted	All organisms (INSDC)	Yes (DRA)	Web API, FTP	Japanese/Asia-Pacific submission and access	Researcher / Student
RefSeq	<a href="https://www.ncbi.nlm.nih.gov/refseq">https://www.ncbi.nlm.nih.gov/refseq</a>	Curated Database	Mixed (manual + computational)	Mixed (manual + computational)	All organisms, reference sequences only	No	E-utilities, FTP	Reference sequences for analysis; clinical reporting	Researcher / Student
NCBI Nucleotide	<a href="https://www.ncbi.nlm.nih.gov/nucleotide">https://www.ncbi.nlm.nih.gov/nucleotide</a>	Search Interface	Mixed (GenBank + RefSeq)	Mixed (GenBank + RefSeq)	All organisms (GenBank + RefSeq)	No	E-utilities	Unified search across NCBI nucleotide databases	Researcher / Student

Category D

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
D	NCBI BLAST	<a href="https://blast.ncbi.nlm.nih.gov">https://blast.ncbi.nlm.nih.gov</a>	Heuristic local alignment	Fast	Fast	Moderate	Yes	General-purpose sequence identification	Single to moderate queries	Researcher / Student
D	EBI BLAST	<a href="https://www.ebi.ac.uk/Tools/sss/ncbiblast">https://www.ebi.ac.uk/Tools/sss/ncbiblast</a>	Heuristic local alignment (NCBI BLAST)	Fast	Fast	Moderate	Yes	Searching EBI databases (UniProt, ENA)	Single to moderate queries	Researcher / Student
D	HMMER	<a href="https://hmmer.org">https://hmmer.org</a> <a href="https://www.ebi.ac.uk/Tools/hmmer">https://www.ebi.ac.uk/Tools/hmmer</a>	Profile HMM	Moderate	Moderate	High (remote homologs)	Yes (EBI)	Protein family classification; remote homologs	Single to moderate queries	Researcher / Student
D	PSI-BLAST	<a href="https://blast.ncbi.nlm.nih.gov">https://blast.ncbi.nlm.nih.gov</a>	Iterative PSSM-based BLAST	Moderate	Moderate	High (iterative)	Yes (NCBI)	Detecting distant protein homologs	Single queries	Researcher / Student
D	DIAMOND	<a href="https://github.com/bbuchfink/diamond">https://github.com/bbuchfink/diamond</a>	Double-index heuristic	Very fast (100x BLAST)	Very fast (100x BLAST)	Moderate-High	No (command-line only)	Large-scale metagenomics; genome annotation	Millions of queries	Researcher / Student
D	FASTA (EBI)	<a href="https://www.ebi.ac.uk/Tools/sss/fasta">https://www.ebi.ac.uk/Tools/sss/fasta</a>	Smith-Waterman + ktup	Moderate	Moderate	Moderate-High	Yes (EBI)	Cross-validation of BLAST; accurate alignments	Single to moderate queries	Researcher / Student

Category E

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
E	Ensembl	<a href="https://www.ensembl.org">https://www.ensembl.org</a>	Hosted Genome Browser/DB	Web (hosted)	Web (hosted)	Yes (300+ species)	Limited	Vertebrate genome annotation; VEP; BioMart	Researcher/Bioinformatician	Researcher / Student
E	UCSC Genome Browser	<a href="https://genome.ucsc.edu">https://genome.ucsc.edu</a>	Hosted Genome Browser	Web (hosted)	Web (hosted)	Yes (100+ species, 100s of tracks)	Yes (custom tracks)	Regulatory annotation; ENCODE data; Table Browser	Researcher/Bioinformatician	Researcher / Student
E	NCBI Genome Data Viewer	<a href="https://www.ncbi.nlm.nih.gov/genome/gdv">https://www.ncbi.nlm.nih.gov/genome/gdv</a>	Hosted Genome Browser	Web (hosted)	Web (hosted)	Yes (NCBI assemblies)	Limited	RefSeq annotations; NCBI variant data	Researcher/Clinician	Researcher / Student



Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
E	JBrowse	<a href="https://jbrowse.org">https://jbrowse.org</a>	Deployable Genome Browser	Self-hosted	Self-hosted	No (user provides data)	Yes (primary purpose)	Deploying custom genome browsers; sharing data	Bioinformatician/DB developer	Researcher / Student
E	IGV	<a href="https://igv.org">https://igv.org</a>	Desktop Genome Browser	Desktop/Web app	Desktop/Web app	Reference genomes only	Yes (primary purpose)	Visualizing personal sequencing data (BAM, VCF)	Bioinformatician/Researcher	Researcher / Student

Category F

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
F	NCBI Gene	<a href="https://www.ncbi.nlm.nih.gov/gene">https://www.ncbi.nlm.nih.gov/gene</a>	Database/Knowledgebase	All organisms	All organisms	Mixed (automated + manual)	Yes (OMIM links)	E-utilities	Gene IDs; integrated gene information; all organisms	Researcher / Student
F	GeneCards	<a href="https://www.genecards.org">https://www.genecards.org</a>	Knowledgebase/Portal	Human only	Human only	Mixed (automated from 150+ sources)	Yes (comprehensive)	Suite API (subscription)	Quick human gene overview; disease associations	Researcher / Student
F	HGNC	<a href="https://www.genenames.org">https://www.genenames.org</a>	Database	Human only	Human only	Manually curated	No (nomenclature only)	REST API	Official human gene symbols; ID conversion	Researcher / Student
F	OMIM	<a href="https://www.omim.org">https://www.omim.org</a>	Knowledgebase	Human (Mendelian diseases)	Human (Mendelian diseases)	Manually curated	Yes (Mendelian diseases)	OMIM API (registration)	Mendelian disease genetics; clinical genetics	Researcher / Student
F	Ensembl Genes	<a href="https://www.ensembl.org">https://www.ensembl.org</a>	Database/Genome Browser	300+ species (vertebrates)	300+ species (vertebrates)	Mixed (computational + GENCODE manual)	Limited (OMIM links)	REST API, BioMart	Genome-wide annotations; RNA-seq; comparative genomics	Researcher / Student

Category G

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
G	GEO	<a href="https://www.ncbi.nlm.nih.gov/geo">https://www.ncbi.nlm.nih.gov/geo</a>	Repository	Microarray, RNA-seq, ChIP-seq, etc.	Microarray, RNA-seq, ChIP-seq, etc.	As submitted (variable)	Open	E-utilities, GEOquery R	Finding expression datasets; data deposition	Researcher / Student
G	ArrayExpress/BioStudies	<a href="https://www.ebi.ac.uk/biostudies/arrayexpress">https://www.ebi.ac.uk/biostudies/arrayexpress</a>	Repository	Microarray, RNA-seq, etc.	Microarray, RNA-seq, etc.	As submitted (MIAME/MINSEQE)	Open	BioStudies API	European data deposition; European datasets	Researcher / Student
G	Expression Atlas	<a href="https://www.ebi.ac.uk/gxa">https://www.ebi.ac.uk/gxa</a>	Curated Database	RNA-seq, microarray (curated subset)	RNA-seq, microarray (curated subset)	Uniformly reprocessed	Open	Expression Atlas REST API	Cross-experiment comparisons; tissue expression	Researcher / Student
G	GTEx	<a href="https://gtexportal.org">https://gtexportal.org</a>	Database/Dataset Collection	RNA-seq (human tissues)	RNA-seq (human tissues)	Uniformly processed	Open (processed); dbGaP (raw)	GTEx Portal API	Human tissue-specific expression; eQTL data	Researcher / Student
G	SRA	<a href="https://www.ncbi.nlm.nih.gov/sra">https://www.ncbi.nlm.nih.gov/sra</a>	Repository	Raw sequencing reads (all types)	Raw sequencing reads (all types)	Raw (FASTQ/BAM)	Open (most); dbGaP (controlled)	SRA Toolkit, E-utilities	Raw sequencing data download and deposition	Researcher / Student



Category H

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
H	SRA (Sequence Read Archive)	<a href="https://www.ncbi.nlm.nih.gov/sra">https://www.ncbi.nlm.nih.gov/sra</a>	Repository	All organisms; >50 petabases; largest collection	All organisms; >50 petabases; largest collection	Community-submitted; automated QC	SRA Toolkit (fasterq-dump, prefetch); NCBI E-utilities; AWS/GCP cloud	Open; free	Largest raw sequencing archive; NCBI ecosystem integration	Researcher / Student
H	ENA (European Nucleotide Archive)	<a href="https://www.ebi.ac.uk/ena">https://www.ebi.ac.uk/ena</a>	Repository	All organisms; mirrors SRA; often faster European access	All organisms; mirrors SRA; often faster European access	Community-submitted; automated QC	ENA REST API; FTP; Aspera; ENA Browser Tools	Open; free	European access; direct FASTQ download without toolkit; EBI integration	Researcher / Student
H	DRA (DDBJ Sequence Read Archive)	<a href="https://www.ddbj.nig.ac.jp/dra">https://www.ddbj.nig.ac.jp/dra</a>	Repository	All organisms; mirrors SRA; strong Asian dataset coverage	All organisms; mirrors SRA; strong Asian dataset coverage	Community-submitted; automated QC	DRA REST API; FTP; Aspera	Open; free	Asian datasets; DDBJ/INSDC ecosystem; Japanese research data	Researcher / Student

Category I

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
I	UniProt (Universal Protein Resource)	<a href="https://www.uniprot.org">https://www.uniprot.org</a>	Database/Knowledgebase	>250M sequences (TrEMBL + Swiss-Prot); all organisms	>250M sequences (TrEMBL + Swiss-Prot); all organisms	Mixed: Swiss-Prot manually reviewed; TrEMBL computationally annotated	UniProt REST API; FTP; Python requests; Biopython	Open; CC BY 4.0	Primary protein sequence and function resource; all organisms	Researcher / Student
I	Swiss-Prot	<a href="https://www.uniprot.org/uniprotkb?query=reviewed:true">https://www.uniprot.org/uniprotkb?query=reviewed:true</a>	Database/Knowledgebase	~570K manually reviewed entries; all organisms	~570K manually reviewed entries; all organisms	Manually reviewed by expert curators	Via UniProt API (filter reviewed:true)	Open; CC BY 4.0	High-quality manually reviewed protein annotations	Researcher / Student
I	TrEMBL	<a href="https://www.uniprot.org/uniprotkb?query=reviewed:false">https://www.uniprot.org/uniprotkb?query=reviewed:false</a>	Database	>250M computationally annotated entries	>250M computationally annotated entries	Computationally predicted; not manually reviewed	Via UniProt API (filter reviewed:false)	Open; CC BY 4.0	Comprehensive coverage including newly sequenced proteins	Researcher / Student
I	InterPro	<a href="https://www.ebi.ac.uk/interpro">https://www.ebi.ac.uk/interpro</a>	Database/Knowledgebase	>40K entries integrating 13 member databases	>40K entries integrating 13 member databases	Mixed: member databases vary; InterPro integration is curated	InterPro REST API; InterProScan command-line	Open; free	Integrated protein family/domain/motif classification	Researcher / Student
I	PROSITE	<a href="https://prosite.expasy.org">https://prosite.expasy.org</a>	Database	>1,800 patterns and profiles; manually curated	>1,800 patterns and profiles; manually curated	Manually curated by ExPASy team	ExPASy API; ScanProsite web tool; FASTA input	Open; free	Protein patterns and profiles; functional site detection	Researcher / Student
I	SMART (Simple Modular Architecture Research Tool)	<a href="https://smart.embl.de">https://smart.embl.de</a>	Database/Tool	>1,300 domain families; focus on signaling domains	>1,300 domain families; focus on signaling domains	Manually curated; HMM-based	SMART web interface; batch submission	Open; free	Signaling domain analysis; domain architecture visualization	Researcher / Student



Category J

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
J	Pfam (now in InterPro)	<a href="https://www.ebi.ac.uk/interpro">https://www.ebi.ac.uk/interpro</a>	Database	>20K protein families; HMM-based; integrated into InterPro	>20K protein families; HMM-based; integrated into InterPro	Manually curated; HMM-based	Via InterPro API; InterProScan; HMMER	Open; free	Protein family classification; HMM-based domain detection	Researcher / Student
J	InterPro	<a href="https://www.ebi.ac.uk/interpro">https://www.ebi.ac.uk/interpro</a>	Database/Knowledgebase	>40K entries integrating 13 member databases	>40K entries integrating 13 member databases	Mixed; integration is curated	InterPro REST API; InterProScan	Open; free	Comprehensive integrated domain/family/motif annotation	Researcher / Student
J	CDD (Conserved Domain Database)	<a href="https://www.ncbi.nlm.nih.gov/cdd">https://www.ncbi.nlm.nih.gov/cdd</a>	Database	>60K domain models; includes Pfam, SMART, COG, etc.	>60K domain models; includes Pfam, SMART, COG, etc.	Mixed: curated NCBI domains + imported from other databases	NCBI E-utilities; RPS-BLAST; CD-Search web tool	Open; free	NCBI ecosystem integration; RPS-BLAST domain search	Researcher / Student
J	PROSITE	<a href="https://prosite.expasy.org">https://prosite.expasy.org</a>	Database	>1,800 patterns and profiles	>1,800 patterns and profiles	Manually curated	ExPASy API; ScanProsite	Open; free	Short functional motifs and patterns; active site detection	Researcher / Student
J	PRINTS	<a href="https://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS">https://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS</a>	Database	~2,000 protein fingerprints; LIMITED UPDATES since ~2012	~2,000 protein fingerprints; LIMITED UPDATES since ~2012	Manually curated (legacy)	Limited; web interface only	Open; free	LEGACY: protein fingerprints; largely superseded by InterPro	Researcher / Student
J	SUPERFAMILY	<a href="https://supfam.org">https://supfam.org</a>	Database	SCOP superfamily HMMs; structural domain classification	SCOP superfamily HMMs; structural domain classification	Computationally derived from SCOP structural classification	SUPERFAMILY API; batch submission	Open; free	Structural domain classification; SCOP superfamily assignment	Researcher / Student

Category K

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
K	RCSB PDB (Protein Data Bank)	<a href="https://www.rcsb.org">https://www.rcsb.org</a>	Repository/Database	>220K experimentally determined structures; all organisms	>220K experimentally determined structures; all organisms	Community-submitted; automated validation; some manual curation	RCSB PDB REST API; FTP; GraphQL API; mmCIF/PDB format	Open; free; CC0	Primary structure repository; experimental structures; US portal	Researcher / Student
K	PDBe (Protein Data Bank in Europe)	<a href="https://www.ebi.ac.uk/pdbe">https://www.ebi.ac.uk/pdbe</a>	Repository/Database	Same as RCSB PDB (wwPDB mirror); enhanced annotations	Same as RCSB PDB (wwPDB mirror); enhanced annotations	Community-submitted; enhanced EBI annotations	PDBe REST API; FTP; SIFTS cross-references	Open; free	European access; SIFTS sequence-structure mapping; EBI integration	Researcher / Student
K	PDBj (Protein Data Bank Japan)	<a href="https://pdbj.org">https://pdbj.org</a>	Repository/Database	Same as RCSB PDB (wwPDB mirror); enhanced annotations	Same as RCSB PDB (wwPDB mirror); enhanced annotations	Community-submitted; enhanced PDBj annotations	PDBj REST API; FTP; Mine2 SQL interface	Open; free	Asian access; Mine2 SQL queries; Japanese research data	Researcher / Student
K	AlphaFold Protein Structure Database	<a href="https://alphafold.ebi.ac.uk">https://alphafold.ebi.ac.uk</a>	Database	>200M predicted structures; nearly all known proteins	>200M predicted structures; nearly all known proteins	Computationally predicted by AlphaFold2; not experimentally verified	AlphaFold REST API; FTP bulk download; EBI API	Open; CC BY 4.0	Predicted structures for proteins without experimental data	Researcher / Student





Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
K	SWISS-MODEL Repository	<a href="https://swissmodel.expasy.org/repository">https://swissmodel.expasy.org/repository</a>	Database	Millions of homology models; UniProt-based coverage	Millions of homology models; UniProt-based coverage	Computationally generated homology models; automated pipeline	SWISS-MODEL API; FTP; web interface	Open; free for academic use	Homology models; template-based structure prediction	Researcher / Student

Category L

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
L	dbSNP	<a href="https://www.ncbi.nlm.nih.gov/snp">https://www.ncbi.nlm.nih.gov/snp</a>	Database	>1 billion variants; all organisms; germline	>1 billion variants; all organisms; germline	Community-submitted; automated rsID assignment	NCBI E-utilities; dbSNP REST API; FTP VCF downloads	Open; free	rsID assignment; variant catalog; population frequency overview	Researcher / Student
L	ClinVar	<a href="https://www.ncbi.nlm.nih.gov/clinvar">https://www.ncbi.nlm.nih.gov/clinvar</a>	Database/Knowledgebase	>2M variant-condition pairs; human germline	>2M variant-condition pairs; human germline	Community-submitted; 0-4 star review status system	NCBI E-utilities; ClinVar FTP; VCF downloads	Open; free	Clinical significance of germline variants; standard clinical reference	Researcher / Student
L	gnomAD	<a href="https://gnomad.broadinstitute.org">https://gnomad.broadinstitute.org</a>	Database	>730K exomes + 76K genomes; human; germline	>730K exomes + 76K genomes; human; germline	Computationally processed with rigorous QC	gnomAD GraphQL API; Google Cloud FTP; VCF downloads	Open; CC BY 4.0	High-quality population frequencies; gene constraint metrics (pLI/LOEUF)	Researcher / Student
L	COSMIC	<a href="https://cancer.sanger.ac.uk/cosmic">https://cancer.sanger.ac.uk/cosmic</a>	Database/Knowledgebase	>17M somatic mutations; >1.5M tumor samples; human cancer	>17M somatic mutations; >1.5M tumor samples; human cancer	Mixed: Cancer Gene Census manually curated; mutations from literature	COSMIC REST API (registration required); FTP (registration required)	Free registration required; commercial license for commercial use	Somatic cancer mutations; Cancer Gene Census; mutational signatures	Researcher / Student
L	Ensembl VEP	<a href="https://www.ensembl.org/vep">https://www.ensembl.org/vep</a>	Tool	Integrates multiple databases; human and other organisms	Integrates multiple databases; human and other organisms	Tool integrating data from multiple sources	VEP REST API; command-line tool; Docker; Conda	Open; Apache 2.0	Variant functional consequence annotation; multi-database integration	Researcher / Student
L	LOVD	<a href="https://www.lovd.nl">https://www.lovd.nl</a>	Database	Gene-specific databases; thousands of genes; human	Gene-specific databases; thousands of genes; human	Community-submitted; quality varies by database	LOVD API (per installation); data export	Open; free	Gene-specific variant databases; detailed phenotype information	Researcher / Student
L	ClinGen	<a href="https://clinicalgenome.org">https://clinicalgenome.org</a>	Knowledgebase/Portal	Expert panel-reviewed genes and variants; human	Expert panel-reviewed genes and variants; human	Manually curated by expert panels; highest quality available	ClinGen API; data downloads	Open; free	Highest-quality gene-disease validity and variant classifications	Researcher / Student

Category M

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
M	OMIM	<a href="https://www.omim.org">https://www.omim.org</a>	Knowledgebase	>26K entries; Mendelian diseases; human	>26K entries; Mendelian diseases; human	Manually curated by expert editors; literature-based	OMIM API (free key required); genemap2.txt download	Free registration; commercial license required for commercial use	Mendelian disease genetics; molecular mechanisms; MIM numbers	Researcher / Student
M	Orphanet	<a href="https://www.orpha.net">https://www.orpha.net</a>	Knowledgebase/Portal	>10K rare diseases; ORPHA codes; European focus	>10K rare diseases; ORPHA codes; European focus	Manually curated by rare disease experts	Orphanet API; Orphadata XML downloads	Open; free for academic use	Rare disease clinical information; ORPHA codes; expert centers	Researcher / Student



Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
M	ClinGen	<a href="https://clinicalgenome.org">https://clinicalgenome.org</a>	Knowledgebase/Portal	Expert panel-reviewed genes; human	Expert panel-reviewed genes; human	Manually curated by expert panels; highest quality	ClinGen API; data downloads	Open; free	Evidence-graded gene-disease validity; actionability assessments	Researcher / Student
M	DisGeNET	<a href="https://www.disgenet.org">https://www.disgenet.org</a>	Database/Knowledgebase	>1.5M gene-disease associations; >30K diseases; human	>1.5M gene-disease associations; >30K diseases; human	Mixed: curated + GWAS + animal models + text mining	DisGeNET REST API; disgenet2r R package; data downloads	Free registration; commercial license required	Broad gene-disease associations; complex diseases; network analysis	Researcher / Student
M	MalaCards	<a href="https://www.malacards.org">https://www.malacards.org</a>	Database/Portal	>20K diseases; integrated from dozens of sources	>20K diseases; integrated from dozens of sources	Aggregated; not independently curated	MalaCards API (subscription); GeneCards Suite API	Free basic access; subscription for full API	Rapid disease overview; disease aliases; multi-source integration	Researcher / Student

Category N

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
N	KEGG	<a href="https://www.kegg.jp">https://www.kegg.jp</a>	Knowledgebase/Database	>500 organisms; metabolic + signaling + disease pathways	>500 organisms; metabolic + signaling + disease pathways	Manually curated pathway maps	KEGG REST API (limited academic); KEGGREST R package; pathview	Academic web use free; API/FTP requires license	Metabolic pathways; cross-species analysis; KEGG ORTHOLOGY	Researcher / Student
N	Reactome	<a href="https://reactome.org">https://reactome.org</a>	Knowledgebase/Database	>15K human reactions; >2,500 pathways; 20+ species	>15K human reactions; >2,500 pathways; 20+ species	Manually curated; peer-reviewed; literature evidence codes	Reactome REST API; FTP; ReactomePA R package	Open; CC BY 4.0	Detailed human signaling pathways; reaction-level information	Researcher / Student
N	BioCyc	<a href="https://biocyc.org">https://biocyc.org</a>	Knowledgebase/Database	>20K organism-specific databases; metabolic focus	>20K organism-specific databases; metabolic focus	Mixed: EcoCyc/MetaCyc highly curated; others computationally predicted	BioCyc API (subscription); Pathway Tools software	Tiered: basic free; full access subscription; academic license	Organism-specific metabolic pathways; microbial metabolism; EcoCyc	Researcher / Student
N	WikiPathways	<a href="https://www.wikipathways.org">https://www.wikipathways.org</a>	Database/Knowledgebase	>3,000 pathways; >30 species; community-curated	>3,000 pathways; >30 species; community-curated	Community-curated; quality varies	WikiPathways REST API; GMT downloads; rWikiPathways R package	Open; CC BY 4.0	Community pathways; disease-specific pathways; freely reusable	Researcher / Student
N	STRING	<a href="https://string-db.org">https://string-db.org</a>	Database/Knowledgebase	>3B interactions; >12K organisms; protein interactions	>3B interactions; >12K organisms; protein interactions	Mixed: experimental curated + computational predictions + text mining	STRING REST API; FTP; STRINGdb R package; Cytoscape stringApp	Academic free; commercial license required	Protein interaction networks; network enrichment analysis	Researcher / Student

Category O

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
O	Primary focus	Functional associations	Genetic & physical interactions	Molecular interactions	Molecular interactions	Experimental interactions	Molecular interactions	—	—	Researcher / Student
O	Interaction types	Physical + functional	Physical + genetic	Physical only	Physical only	Physical only	Physical only	—	—	Researcher / Student
O	Evidence types	Experimental + predicted	Experimental only	Experimental only	Experimental only	Experimental only	Experimental only	—	—	Researcher / Student
O	Organisms covered	14,000+	~70	~700	~700	~200	~100	—	—	Researcher / Student
O	Confidence scores	Yes (0-1000)	No	MI-score	MI-score	No	No	—	—	Researcher / Student
O	Genetic interactions	No	Yes	No	No	No	No	—	—	Researcher / Student





Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
O	Update status	Active	Active	Active	Active	Limited (~2017)	Limited	—	—	Researcher / Student
O	API quality	Excellent	Good	Good	Good	Limited	Limited	—	—	Researcher / Student
O	Access	Free	Free	Free	Free	Free	Free	—	—	Researcher / Student
O	Best for	Network analysis	Genetic interactions	Curated physical	Curated physical	Legacy data	Legacy data	—	—	Researcher / Student

Category P

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
P	Primary focus	Drug-target-disease	Bioactive compounds	Chemical information	Chemical information	Binding affinities	Therapeutic targets	—	—	Researcher / Student
P	Compound count	~15,000 drugs	~2.4M compounds	~115M compounds	~115M compounds	~2.8M compounds	~3,600 targets	—	—	Researcher / Student
P	Binding data	Yes	Yes	Yes	Yes	Yes (primary)	Yes	—	—	Researcher / Student
P	Drug approval status	Yes	Yes	Limited	Limited	No	Yes	—	—	Researcher / Student
P	ADMET data	Yes	Limited	Limited	Limited	No	Limited	—	—	Researcher / Student
P	Clinical data	Yes	Yes	Limited	Limited	No	Yes	—	—	Researcher / Student
P	Access	Tiered (free academic)	Free	Free	Free	Free	Free	—	—	Researcher / Student
P	API quality	Good	Excellent	Excellent	Excellent	Good	Limited	—	—	Researcher / Student
P	Best for	Drug information	Bioactivity data	Chemical data	Chemical data	Binding constants	Target-disease links	—	—	Researcher / Student

Category Q

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
Q	Domain	Gene function	Sequence features	Human phenotypes	Human phenotypes	Diseases	Anatomy	Medical concepts	—	Researcher / Student
Q	Scope	All organisms	All organisms	Human	Human	Human diseases	Multi-species	Biomedical	—	Researcher / Student
Q	Term count	~44,000	~2,400	~17,000	~17,000	~11,000	~25,000	~30,000	—	Researcher / Student
Q	Format	OBO/OWL	OBO/OWL	OBO/OWL	OBO/OWL	OBO/OWL	OBO/OWL	MeSH XML	—	Researcher / Student
Q	Used for annotation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Researcher / Student
Q	Clinical use	Limited	No	Yes	Yes	Yes	Limited	Yes	—	Researcher / Student
Q	Update frequency	Regular	Regular	Regular	Regular	Regular	Regular	Regular	—	Researcher / Student
Q	Access	Free	Free	Free	Free	Free	Free	Free	—	Researcher / Student
Q	Best for	Gene function	Sequence annotation	Clinical phenotypes	Clinical phenotypes	Disease annotation	Anatomy	Literature indexing	—	Researcher / Student

Category R

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
R	Primary focus	Functional elements	Reference epigenomes	ChIP-seq analysis	ChIP-seq analysis	TF binding profiles	ChIP-seq integration	—	—	Researcher / Student
R	Data types	ChIP, ATAC, RNA, etc.	ChIP, WGBS, RNA	ChIP-seq	ChIP-seq	PWMs/PFM	ChIP-seq	—	—	Researcher / Student
R	Organisms	Human, mouse, others	Human	Human, mouse	Human, mouse	Multiple	Human, mouse, others	—	—	Researcher / Student
R	Update status	Active	Complete (no new data)	Active	Active	Active	Active	—	—	Researcher / Student
R	TF motifs	Limited	No	No	No	Yes (primary)	No	—	—	Researcher / Student



Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
R	Peak data	Yes	Yes	Yes	Yes	No	Yes	—	—	Researcher / Student
R	API quality	Excellent	Limited	Good	Good	Good	Good	—	—	Researcher / Student
R	Access	Free	Free	Free	Free	Free	Free	—	—	Researcher / Student
R	Best for	Comprehensive epigenomics	Reference epigenomes	ChIP-seq analysis	ChIP-seq analysis	TF motifs	ChIP-seq integration	—	—	Researcher / Student

Category S

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
S	Primary focus	Human cell atlas	Single-cell data portal	Cell type markers	Cell type markers	Single-cell expression	—	—	—	Researcher / Student
S	Organisms	Human (primary)	Human, mouse, others	Human, mouse	Human, mouse	Multiple	—	—	—	Researcher / Student
S	Data types	scRNA-seq, spatial, etc.	scRNA-seq, spatial	scRNA-seq	scRNA-seq	scRNA-seq	—	—	—	Researcher / Student
S	Cell type annotations	Yes	Yes	Yes (markers)	Yes (markers)	Yes	—	—	—	Researcher / Student
S	Spatial data	Yes	Yes	No	No	Limited	—	—	—	Researcher / Student
S	Interactive viewer	Yes	Yes (primary)	Limited	Limited	Yes	—	—	—	Researcher / Student
S	API quality	Good	Excellent	Limited	Limited	Good	—	—	—	Researcher / Student
S	Access	Free	Free	Free	Free	Free	—	—	—	Researcher / Student
S	Best for	Human cell reference	Data exploration	Cell type markers	Cell type markers	EBI single-cell data	—	—	—	Researcher / Student

Category T

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
T	Primary focus	Metagenomics analysis	rRNA sequences	Genome taxonomy	Genome taxonomy	rRNA sequences	Metagenomics analysis	—	—	Researcher / Student
T	Data types	Metagenomes, amplicons	16S/18S/23S/28S rRNA	Bacterial/archaeal genomes	Bacterial/archaeal genomes	16S rRNA	Metagenomes	—	—	Researcher / Student
T	Taxonomy system	SILVA-based	Own (SILVA)	GTDB (own)	GTDB (own)	Own (RDP)	Multiple	—	—	Researcher / Student
T	Update status	Active	Active	Active	Active	Limited	Active	—	—	Researcher / Student
T	Analysis tools	Yes	Yes (ARB)	Yes	Yes	Yes (Classifier)	Yes	—	—	Researcher / Student
T	API quality	Good	Limited	Good	Good	Limited	Good	—	—	Researcher / Student
T	Access	Free	Free	Free	Free	Free	Free	—	—	Researcher / Student
T	Best for	Metagenomics	rRNA taxonomy	Genome taxonomy	Genome taxonomy	16S classification (legacy)	Metagenomics analysis	—	—	Researcher / Student

Category U

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
U	Primary focus	Sequence taxonomy	Protein taxonomy	Species catalog	Species catalog	—	—	—	—	Researcher / Student
U	Organisms covered	~500,000	~1,000,000	~2,000,000	~2,000,000	—	—	—	—	Researcher / Student
U	Taxonomic ranks	All ranks	All ranks	All ranks	All ranks	—	—	—	—	Researcher / Student
U	Integration	NCBI databases	UniProt databases	Multiple sources	Multiple sources	—	—	—	—	Researcher / Student
U	Nomenclature	ICZN/ICBN	ICZN/ICBN	ICZN/ICBN	ICZN/ICBN	—	—	—	—	Researcher / Student
U	API quality	Good (E-utilities)	Good	Good	Good	—	—	—	—	Researcher / Student



Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
U	Access	Free	Free	Free	Free	—	—	—	—	Researcher / Student
U	Best for	Sequence taxonomy	Protein taxonomy	Species catalog	Species catalog	—	—	—	—	Researcher / Student

Category V

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
V	Peptide count	~22,000	~3,500	~12,000	~12,000	~10,000	~15,000	—	—	Researcher / Student
V	Experimental data	Yes	Yes	Yes	Yes	Yes	Yes	—	—	Researcher / Student
V	Structural data	Limited	Limited	Limited	Limited	Limited	Yes	—	—	Researcher / Student
V	Prediction tools	Yes	Yes	Yes	Yes	Yes	Yes	—	—	Researcher / Student
V	Update status	Active	Active	Verify status	Verify status	Active	Active	—	—	Researcher / Student
V	API quality	Limited	Limited	Limited	Limited	Limited	Limited	—	—	Researcher / Student
V	Access	Free	Free	Free	Free	Free	Free	—	—	Researcher / Student
V	Best for	Comprehensive AMP data	AMP properties	AMP prediction	AMP prediction	AMP collection	Structure-activity	—	—	Researcher / Student

Category W

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
W	Primary focus	Multi-omics cancer data	International cancer genomes	Interactive analysis portal	Interactive analysis portal	Somatic mutation catalog	Clinical mutation annotation	—	—	Researcher / Student
W	Sample count	~11,000	~25,000	~100,000+	~100,000+	~1.5M tumor samples	~6,000 annotated alterations	—	—	Researcher / Student
W	Cancer types	33	50+	300+ studies	300+ studies	All cancers	Clinically relevant	—	—	Researcher / Student
W	Data types	Multi-omics	Genomics + expression	Multi-omics	Multi-omics	Mutations + signatures	Clinical annotations	—	—	Researcher / Student
W	Clinical data	Yes	Yes	Yes	Yes	Limited	Yes (therapeutic)	—	—	Researcher / Student
W	Interactive interface	Limited (GDC)	Limited	Excellent	Excellent	Moderate	Good	—	—	Researcher / Student
W	API quality	Good (GDC)	Good	Excellent	Excellent	Good	Good	—	—	Researcher / Student
W	Access	Open + controlled	Open + controlled	Free	Free	Free (registration)	Free academic	—	—	Researcher / Student
W	Best for	Raw data analysis	International cohorts	Interactive exploration	Interactive exploration	Mutation annotation	Clinical interpretation	—	—	Researcher / Student

Category X

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
X	Organism	D. melanogaster	C. elegans	M. musculus	M. musculus	D. rerio	S. cerevisiae	A. thaliana	S. pombe	Researcher / Student
X	Kingdom	Animal	Animal	Animal	Animal	Animal	Fungi	Plant	Fungi	Researcher / Student
X	Genome size	~180 Mb	~100 Mb	~2.7 Gb	~2.7 Gb	~1.4 Gb	~12 Mb	~135 Mb	~14 Mb	Researcher / Student
X	Genetic interactions	Moderate	Moderate	Limited	Limited	Limited	Comprehensive	Limited	Moderate	Researcher / Student
X	Phenotype data	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Extensive	Researcher / Student
X	Disease models	Yes	Yes	Yes (primary)	Yes (primary)	Yes	Limited	Limited	Limited	Researcher / Student
X	Expression data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Researcher / Student
X	API quality	Good	Good	Good	Good	Good	Good	Limited	Good	Researcher / Student



Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
X	Alliance member	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Researcher / Student
X	Access	Free	Free	Free	Free	Free	Free	Free	Free	Researcher / Student

Category Y

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
Y	Primary function	Registry of standards/databases	Sample metadata repository	Project metadata repository	Project metadata repository	Study data repository	—	—	—	Researcher / Student
Y	Data stored	Metadata about resources	Sample metadata	Project metadata	Project metadata	Complete study datasets	—	—	—	Researcher / Student
Y	Persistent IDs	Yes (DOIs)	Yes (SAMEA/SAMN)	Yes (PRJNA/PRJEB)	Yes (PRJNA/PRJEB)	Yes (S-BSST)	—	—	—	Researcher / Student
Y	Data types	All (registry)	Biological samples	Research projects	Research projects	Any biological data	—	—	—	Researcher / Student
Y	Required by journals	Yes (for standards)	Yes (for data submission)	Yes (for NCBI submission)	Yes (for NCBI submission)	Increasingly	—	—	—	Researcher / Student
Y	API quality	Good	Good	Good	Good	Good	—	—	—	Researcher / Student
Y	Access	Free	Free	Free	Free	Free	—	—	—	Researcher / Student
Y	Best for	Finding standards/databases	Sample metadata	NCBI data submission	NCBI data submission	Complete study datasets	—	—	—	Researcher / Student

Category Z

Category	Database/Resource	Official URL	Main Purpose	Data Type	Best Used For	Beginner Friendly?	Programmatic Access?	Curation Level	Major Limitation	Recommended For
Z	Primary focus	Peer-reviewed databases	Bioinformatics tools	Global database catalog	Global database catalog	FAIR standards/databases	—	—	—	Researcher / Student
Z	Coverage	~1,800 databases	~20,000 tools/databases	~5,000 databases	~5,000 databases	~4,000 databases/standards	—	—	—	Researcher / Student
Z	Peer-reviewed	Yes	No	No	No	No	—	—	—	Researcher / Student
Z	Tools included	No	Yes	Yes	Yes	Yes	—	—	—	Researcher / Student
Z	Standards included	No	No	No	No	Yes	—	—	—	Researcher / Student
Z	EDAM ontology	No	Yes	No	No	No	—	—	—	Researcher / Student
Z	API quality	Limited	Good	Limited	Limited	Good	—	—	—	Researcher / Student
Z	Update frequency	Annual	Continuous	Regular	Regular	Continuous	—	—	—	Researcher / Student
Z	Access	Free	Free	Free	Free	Free	—	—	—	Researcher / Student
Z	Best for	Finding databases with citations	Finding tools	International databases	International databases	FAIR compliance	—	—	—	Researcher / Student

Appendix A: Master Database Table

This master table is an index for the atlas. It does not replace the detailed database cards; it helps readers locate the relevant resource, identify its resource type, and see whether special caution is needed.

A. General portals

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
NCBI	Integrated portal	Multi-domain	Best Used For: Core cross-database entry point	Free	Full-depth retained	<a href="https://www.ncbi.nlm.nih.gov">https://www.ncbi.nlm.nih.gov</a>
EMBL-EBI	Integrated portal	Multi-domain	Best Used For: European biological data hub	Free	Full-depth retained	<a href="https://www.ebi.ac.uk">https://www.ebi.ac.uk</a>
DDBJ	Primary archive / portal	Nucleotide sequences	Best Used For: INSDC Japanese node	Free	Full-depth retained	<a href="https://www.ddbj.nig.ac.jp">https://www.ddbj.nig.ac.jp</a>
ExPASy	Portal	Proteomics	Best Used For: SIB bioinformatics/proteomics portal	Free	Full-depth retained	<a href="https://www.expasy.org">https://www.expasy.org</a>

B. Literature



Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
PubMed	Literature database	Biomedical literature	<a href="#">Best Used For</a> : Citation/abstract search	Free	Corrected	URL: <a href="https://pubmed.ncbi.nlm.nih.gov">https://pubmed.ncbi.nlm.nih.gov</a>
PubMed Central	Full-text archive	Biomedical literature	<a href="#">Best Used For</a> : Free full-text articles	Free	Corrected	URL: <a href="https://pmc.ncbi.nlm.nih.gov">https://pmc.ncbi.nlm.nih.gov</a>
Europe PMC	Literature database/archive	Biomedical literature	<a href="#">Best Used For</a> : Literature + full text + preprints	Free	Full-depth retained	URL: <a href="https://europepmc.org">https://europepmc.org</a>
Google Scholar	Academic search engine	All disciplines	<a href="#">Best Used For</a> : Discovery/citation chasing, not reproducible systematic search	Free basic	QC: Flagged limitations	URL: <a href="https://scholar.google.com">https://scholar.google.com</a>
Semantic Scholar	Academic search engine	All disciplines	<a href="#">Best Used For</a> : AI-assisted literature discovery	Free	Index entry	URL: <a href="https://www.semanticscholar.org">https://www.semanticscholar.org</a>
Scopus	Commercial citation database	All disciplines	<a href="#">Best Used For</a> : Citation analysis and systematic searches	Access: Subscription	QC: Restricted	URL: <a href="https://www.scopus.com">https://www.scopus.com</a>
Web of Science	Commercial citation database	All disciplines	<a href="#">Best Used For</a> : Citation analysis and indexed literature search	Access: Subscription	QC: Restricted	URL: <a href="https://www.webofscience.com">https://www.webofscience.com</a>
Cochrane Library	Evidence database	Clinical evidence	<a href="#">Best Used For</a> : Systematic reviews and trials	Access: Mixed/subscription	QC: Restricted	URL: <a href="https://www.cochranelibrary.com">https://www.cochranelibrary.com</a>

C. Nucleotide sequences

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
GenBank	Primary nucleotide archive	DNA/RNA sequences	<a href="#">Best Used For</a> : Submitted nucleotide sequences	Free	Full-depth retained	URL: <a href="https://www.ncbi.nlm.nih.gov/genbank">https://www.ncbi.nlm.nih.gov/genbank</a>
ENA	Primary nucleotide/read archive	DNA/RNA/raw reads	<a href="#">Best Used For</a> : European INSDC access and raw reads	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/ena">https://www.ebi.ac.uk/ena</a>
DDBJ	Primary nucleotide archive	DNA/RNA sequences	<a href="#">Best Used For</a> : Japanese INSDC node	Free	Full-depth retained	URL: <a href="https://www.ddbj.nig.ac.jp">https://www.ddbj.nig.ac.jp</a>
RefSeq	Curated reference database	Genome/transcript/protein references	<a href="#">Best Used For</a> : Non-redundant reference sequences	Free	Full-depth retained	URL: <a href="https://www.ncbi.nlm.nih.gov/refseq">https://www.ncbi.nlm.nih.gov/refseq</a>

D. Similarity tools

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
NCBI BLAST	Sequence similarity tool	DNA/protein sequences	<a href="#">Best Used For</a> : Similarity searching	Free	Full-depth retained	URL: <a href="https://blast.ncbi.nlm.nih.gov/Blast.cgi">https://blast.ncbi.nlm.nih.gov/Blast.cgi</a>
HMMER	Profile-HMM search tool	Protein families	<a href="#">Best Used For</a> : Sensitive homology detection	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/Tools/hmmer/">https://www.ebi.ac.uk/Tools/hmmer/</a>
PSI-BLAST	Iterative similarity tool	Protein sequences	<a href="#">Best Used For</a> : Remote homology detection	Free	Index entry	URL: <a href="https://blast.ncbi.nlm.nih.gov/Blast.cgi">https://blast.ncbi.nlm.nih.gov/Blast.cgi</a>

E. Genome browsers

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
Ensembl	Genome browser/annotation DB	Genomes/transcripts/variants	<a href="#">Best Used For</a> : Gene models, coordinates, VEP	Free	Corrected/retained	URL: <a href="https://www.ensembl.org">https://www.ensembl.org</a>
UCSC Genome Browser	Genome browser	Genome coordinates/tracks	<a href="#">Best Used For</a> : Genome visualization and tracks	Free academic	Full-depth retained	URL: <a href="https://genome.ucsc.edu">https://genome.ucsc.edu</a>
NCBI Genome Data Viewer	Genome browser	Genome annotation	<a href="#">Best Used For</a> : NCBI-centric genome visualization	Free	Index entry	URL: <a href="https://www.ncbi.nlm.nih.gov/genome/gdv">https://www.ncbi.nlm.nih.gov/genome/gdv</a>
IGV	Local genome browser/tool	BAM/VCF/genomics	<a href="#">Best Used For</a> : Local alignment and variant visualization	Free	Index entry	URL: <a href="https://igv.org">https://igv.org</a>

F. Gene databases

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
NCBI Gene	Gene database	Genes/genomes	<a href="#">Best Used For</a> : Gene-centered NCBI records	Free	Full-depth retained	URL: <a href="https://www.ncbi.nlm.nih.gov/gene">https://www.ncbi.nlm.nih.gov/gene</a>
GeneCards	Integrated gene portal	Human genes	<a href="#">Best Used For</a> : Exploration/cross-references, not primary source	Free/restricted	Flagged as portal	URL: <a href="https://www.genecards.org">https://www.genecards.org</a>
HGNC	Nomenclature authority	Human gene names	<a href="#">Best Used For</a> : Approved human gene symbols	Free	Full-depth retained	URL: <a href="https://www.genenames.org">https://www.genenames.org</a>
OMIM	Knowledgebase	Mendelian disease genes	<a href="#">Best Used For</a> : Disease-gene relationships	Free search/licensing for bulk	Restricted/flagged	URL: <a href="https://www.omim.org">https://www.omim.org</a>





G. Transcriptomics

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
GEO	Functional genomics repository	Expression/omics studies	Best Used For: Processed expression datasets + metadata	Free	Corrected/retained	URL: <a href="https://www.ncbi.nlm.nih.gov/geo">https://www.ncbi.nlm.nih.gov/geo</a>
BioStudies ArrayExpress Collection	Functional genomics repository	Expression/omics studies	Best Used For: Current location for ArrayExpress data	Free	Corrected	URL: <a href="https://www.ebi.ac.uk/biostudies/arrayexpress">https://www.ebi.ac.uk/biostudies/arrayexpress</a>
Expression Atlas	Curated expression atlas	Gene expression	Best Used For: Baseline/differential expression summaries	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/gxa">https://www.ebi.ac.uk/gxa</a>
GTEx	Dataset portal	Human tissue expression	Best Used For: Tissue-specific expression and eQTLs	Free/controlled raw	Full-depth retained	URL: <a href="https://gtexportal.org">https://gtexportal.org</a>

H. Raw sequencing

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
SRA	Raw read repository	Sequencing reads	Best Used For: FASTQ/SRA raw reads	Free	Corrected distinction	URL: <a href="https://www.ncbi.nlm.nih.gov/sra">https://www.ncbi.nlm.nih.gov/sra</a>
ENA Read Archive	Raw read repository	Sequencing reads	Best Used For: European raw read access	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/ena">https://www.ebi.ac.uk/ena</a>
DRA	Raw read repository	Sequencing reads	Best Used For: DDBJ raw read archive	Free	Index entry	URL: <a href="https://ddbj.nig.ac.jp/resource/sra-submission-e.html">https://ddbj.nig.ac.jp/resource/sra-submission-e.html</a>

I. Protein sequence/function

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
UniProt	Protein knowledgebase	Proteins	Best Used For: Protein sequences and functional annotation	Free	Corrected classification	URL: <a href="https://www.uniprot.org">https://www.uniprot.org</a>
Swiss-Prot	Manually reviewed protein KB	Proteins	Best Used For: High-confidence curated protein entries	Free	Full-depth retained	URL: <a href="https://www.uniprot.org">https://www.uniprot.org</a>
TrEMBL	Computational protein KB	Proteins	Best Used For: Large unreviewed protein coverage	Free	Full-depth retained	URL: <a href="https://www.uniprot.org">https://www.uniprot.org</a>
NCBI Protein	Protein database	Proteins	Best Used For: NCBI protein records	Free	Index entry	URL: <a href="https://www.ncbi.nlm.nih.gov/protein">https://www.ncbi.nlm.nih.gov/protein</a>

J. Protein domains

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
InterPro	Integrated domain/family DB	Protein domains/families	Best Used For: Protein family/domain annotation	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/interpro">https://www.ebi.ac.uk/interpro</a>
Pfam	Protein family database	Protein domains	Best Used For: HMM-based domain families	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/interpro/entry/pfam/">https://www.ebi.ac.uk/interpro/entry/pfam/</a>
CDD	Conserved domain database	Protein domains	Best Used For: NCBI domain annotation	Free	Full-depth retained	URL: <a href="https://www.ncbi.nlm.nih.gov/cdd">https://www.ncbi.nlm.nih.gov/cdd</a>
PROSITE	Motif/domain DB	Protein motifs	Best Used For: Patterns/profiles of protein families	Free	Full-depth retained	URL: <a href="https://prosite.expasy.org">https://prosite.expasy.org</a>
SMART	Domain database/tool	Protein domains	Best Used For: Domain architecture analysis	Free	Index entry	URL: <a href="http://smart.embl.de">http://smart.embl.de</a>
SUPERFAMILY	Structural domain classification	Protein domains	Best Used For: SCOP-based superfamilies	Free	Index entry	URL: <a href="https://supfam.org">https://supfam.org</a>

K. Structures

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
RCSB PDB	Experimental structure archive/interface	Macromolecular structures	Best Used For: Experimental protein/nucleic acid structures	Free	Corrected	URL: <a href="https://www.rcsb.org">https://www.rcsb.org</a>
PDBe	PDB regional interface	Macromolecular structures	Best Used For: European PDB interface	Free	Corrected	URL: <a href="https://www.ebi.ac.uk/pdbe/">https://www.ebi.ac.uk/pdbe/</a>
PDBj	PDB regional interface	Macromolecular structures	Best Used For: Japanese PDB interface	Free	Corrected	URL: <a href="https://pdbj.org">https://pdbj.org</a>
wwPDB	Global archive partnership	Macromolecular structures	Best Used For: PDB archive governance	Free	Corrected	URL: <a href="https://www.wwpdb.org">https://www.wwpdb.org</a>



Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
AlphaFold DB	Predicted structure DB	Predicted protein structures	<b>Best Used For:</b> AI-predicted structures, not experimental	Free	Corrected	URL: <a href="https://alphafold.ebi.ac.uk">https://alphafold.ebi.ac.uk</a>
SWISS-MODEL Repository	Homology model repository	Protein models	<b>Best Used For:</b> Template-based protein models	Free	Index entry	URL: <a href="https://swissmodel.expasy.org/repository">https://swissmodel.expasy.org/repository</a>
SCOP	Structure classification	Protein folds	<b>Best Used For:</b> Structural classification	Free	Index entry	URL: <a href="https://scop.mrc-lmb.cam.ac.uk">https://scop.mrc-lmb.cam.ac.uk</a>
CATH	Structure classification	Protein folds	<b>Best Used For:</b> Structural classification	Free	Index entry	URL: <a href="https://www.cathdb.info">https://www.cathdb.info</a>

L. Variants

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
dbSNP	Variant archive	Genetic variation	<b>Best Used For:</b> rsIDs and variant records	Free	Full-depth retained	URL: <a href="https://www.ncbi.nlm.nih.gov/snp">https://www.ncbi.nlm.nih.gov/snp</a>
ClinVar	Clinical variant database	Human variants	<b>Best Used For:</b> Clinical significance interpretations	Free	Full-depth retained	URL: <a href="https://www.ncbi.nlm.nih.gov/clinvar">https://www.ncbi.nlm.nih.gov/clinvar</a>
gnomAD	Population variant database	Human variants	<b>Best Used For:</b> Allele frequencies	Free	Full-depth retained	URL: <a href="https://gnomad.broadinstitute.org">https://gnomad.broadinstitute.org</a>
1000 Genomes	Population variant dataset	Human variants	<b>Best Used For:</b> Population variation reference	Free	Index entry	URL: <a href="https://www.internationalgenome.org">https://www.internationalgenome.org</a>
Ensembl VEP	Variant effect tool	Variant annotation	<b>Best Used For:</b> Variant consequence prediction	Free	Corrected/retained	URL: <a href="https://www.ensembl.org/vep">https://www.ensembl.org/vep</a>
LOVD	Variant database platform	Disease variants	<b>Best Used For:</b> Locus-specific variant data	Free/mixed	Index entry	URL: <a href="https://www.lovd.nl">https://www.lovd.nl</a>
HGMD	Mutation database	Human disease mutations	<b>Best Used For:</b> Published disease mutations	Access: Subscription for full	Restricted/uncertain	URL: <a href="https://www.hgmd.cf.ac.uk">https://www.hgmd.cf.ac.uk</a>

M. Disease/clinical genomics

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
OMIM	Disease-gene knowledgebase	Mendelian diseases	<b>Best Used For:</b> Gene-phenotype relationships	Free search/licensed bulk	Restricted/flagged	URL: <a href="https://www.omim.org">https://www.omim.org</a>
Orphanet	Rare disease knowledgebase	Rare diseases	<b>Best Used For:</b> Rare disease information	Free	Full-depth retained	URL: <a href="https://www.orpha.net">https://www.orpha.net</a>
ClinGen	Clinical genomics knowledgebase	Gene-disease validity	<b>Best Used For:</b> Clinical validity and actionability	Free	Full-depth retained	URL: <a href="https://clinicalgenome.org">https://clinicalgenome.org</a>
DisGeNET	Disease-gene association DB	Gene-disease associations	<b>Best Used For:</b> Integrated disease associations	Free/licensing varies	Full-depth retained	URL: <a href="https://www.disgenet.org">https://www.disgenet.org</a>
MalaCards	Integrated disease portal	Diseases	<b>Best Used For:</b> Disease summaries/cross-refs	Free/restricted	Index entry	URL: <a href="https://www.malacards.org">https://www.malacards.org</a>

N. Pathways

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
KEGG	Pathway/knowledgebase/tool suite	Pathways/metabolism	<b>Best Used For:</b> Pathway maps and orthology	Free web/licensed bulk	Restricted/flagged	URL: <a href="https://www.kegg.jp">https://www.kegg.jp</a>
Reactome	Curated pathway knowledgebase	Pathways	<b>Best Used For:</b> Human biological pathways	Free	Full-depth retained	URL: <a href="https://reactome.org">https://reactome.org</a>
BioCyc	Pathway/genome DB collection	Pathways/genomes	<b>Best Used For:</b> Organism-specific pathways	Access: Mixed/subscription for some	QC: Restricted	URL: <a href="https://biocyc.org">https://biocyc.org</a>
WikiPathways	Community pathway DB	Pathways	<b>Best Used For:</b> Community-curated pathways	Free	Index entry	URL: <a href="https://www.wikipathways.org">https://www.wikipathways.org</a>
GSEA/MSigDB	Gene set database/tool	Pathways/gene sets	<b>Best Used For:</b> Gene set enrichment	Free/mixed license	Index entry	URL: <a href="https://www.gsea-msigdb.org">https://www.gsea-msigdb.org</a>

O. PPIs

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
STRING	Interaction database	Protein interactions	<b>Best Used For:</b> Known/predicted PPI networks	Free academic	Full-depth retained	URL: <a href="https://string-db.org">https://string-db.org</a>
BioGRID	Interaction database	Molecular interactions	<b>Best Used For:</b> Curated genetic/protein interactions	Free academic	Full-depth retained	URL: <a href="https://thebiogrid.org">https://thebiogrid.org</a>





Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
IntAct	Molecular interaction DB	Protein interactions	<a href="#">Best Used For:</a> Curated interaction evidence	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/intact">https://www.ebi.ac.uk/intact</a>
DIP	Interaction database	Protein interactions	<a href="#">Best Used For:</a> Historical PPI resource	Access: Uncertain	QC: Uncertain/deprecated check	URL: <a href="https://dip.doe-mbi.ucla.edu">https://dip.doe-mbi.ucla.edu</a>
MINT	Interaction database	Protein interactions	<a href="#">Best Used For:</a> Historical curated PPI resource	Access: Uncertain	QC: Uncertain/deprecated check	URL: <a href="https://mint.bio.uniroma2.it">https://mint.bio.uniroma2.it</a>

P. Drugs/compounds

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
DrugBank	Drug knowledgebase	Drugs/targets	<a href="#">Best Used For:</a> Drug-target and pharmacology data	Access: Licensing restrictions	Restricted/flagged	URL: <a href="https://go.drugbank.com">https://go.drugbank.com</a>
ChEMBL	Bioactivity database	Compounds/targets	<a href="#">Best Used For:</a> Drug discovery bioactivity data	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/chembl">https://www.ebi.ac.uk/chembl</a>
PubChem	Chemical database	Compounds/bioassays	<a href="#">Best Used For:</a> Chemical structures and assays	Free	Full-depth retained	URL: <a href="https://pubchem.ncbi.nlm.nih.gov">https://pubchem.ncbi.nlm.nih.gov</a>
BindingDB	Binding affinity database	Compounds/targets	<a href="#">Best Used For:</a> Binding affinity data	Free	Full-depth retained	URL: <a href="https://www.bindingdb.org">https://www.bindingdb.org</a>
TTD	Therapeutic target DB	Targets/drugs	<a href="#">Best Used For:</a> Drug targets and therapeutics	Free academic	Index entry	URL: <a href="https://db.idrblab.net/ttd/">https://db.idrblab.net/ttd/</a>

Q. Ontologies

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
Gene Ontology	Ontology/knowledgebase	Gene function	<a href="#">Best Used For:</a> GO terms and annotations	Free	Full-depth retained	URL: <a href="https://geneontology.org">https://geneontology.org</a>
Sequence Ontology	Ontology	Sequence features	<a href="#">Best Used For:</a> Standardized sequence feature terms	Free	Index entry	URL: <a href="http://www.sequenceontology.org">http://www.sequenceontology.org</a>
Human Phenotype Ontology	Ontology	Phenotypes	<a href="#">Best Used For:</a> Human phenotype terms	Free	Full-depth retained	URL: <a href="https://hpo.jax.org">https://hpo.jax.org</a>
Disease Ontology	Ontology	Diseases	<a href="#">Best Used For:</a> Disease terminology	Free	Index entry	URL: <a href="https://disease-ontology.org">https://disease-ontology.org</a>
Uberon	Ontology	Anatomy	<a href="#">Best Used For:</a> Cross-species anatomy ontology	Free	Index entry	URL: <a href="https://uberon.github.io">https://uberon.github.io</a>
MeSH	Controlled vocabulary	Literature/biomedical terms	<a href="#">Best Used For:</a> PubMed indexing vocabulary	Free	Index entry	URL: <a href="https://www.nlm.nih.gov/mesh/meshhome.html">https://www.nlm.nih.gov/mesh/meshhome.html</a>

R. Epigenomics

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
ENCODE	Functional genomics consortium/portal	Regulatory genomics	<a href="#">Best Used For:</a> Regulatory elements and assays	Free	Full-depth retained	URL: <a href="https://www.encodeproject.org">https://www.encodeproject.org</a>
Roadmap Epigenomics	Epigenomics dataset portal	Epigenomics	<a href="#">Best Used For:</a> Human epigenome maps	Free	Index entry	URL: <a href="https://egg2.wustl.edu/roadmap/web_portal/">https://egg2.wustl.edu/roadmap/web_portal/</a>
Cistrome DB	ChIP-seq/chromatin DB	Regulatory genomics	<a href="#">Best Used For:</a> TF/chromatin profiles	Free	Index entry	URL: <a href="http://cistrome.org/db">http://cistrome.org/db</a>

S. Single-cell

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
Human Cell Atlas	Consortium/atlas	Single-cell/spatial omics	<a href="#">Best Used For:</a> Human cell reference maps	Free	Full-depth retained	URL: <a href="https://www.humancellatlas.org">https://www.humancellatlas.org</a>
CELLxGENE	Single-cell data portal	Single-cell data	<a href="#">Best Used For:</a> Explore/download scRNA-seq atlases	Free	Full-depth retained	URL: <a href="https://cellxgene.cziscience.com">https://cellxgene.cziscience.com</a>
Single Cell Expression Atlas	Expression atlas	Single-cell expression	<a href="#">Best Used For:</a> Curated single-cell expression	Free	Index entry	URL: <a href="https://www.ebi.ac.uk/gxa/sc/home">https://www.ebi.ac.uk/gxa/sc/home</a>
PanglaoDB	Single-cell marker DB	Cell markers	<a href="#">Best Used For:</a> Cell type marker genes	Free	Index entry	URL: <a href="https://panglaodb.se">https://panglaodb.se</a>

T. Microbiome



Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
MGNify	Metagenomics resource	Microbiome/metagenomics	<a href="#">Best Used For</a> : Metagenomic dataset analysis	Free	Full-depth retained	URL: <a href="https://www.ebi.ac.uk/metagenomics">https://www.ebi.ac.uk/metagenomics</a>
SILVA	rRNA database	Microbial taxonomy/rRNA	<a href="#">Best Used For</a> : rRNA reference sequences	Free academic	Full-depth retained	URL: <a href="https://www.arb-silva.de">https://www.arb-silva.de</a>
RDP	rRNA database/classifier	Microbial taxonomy/rRNA	<a href="#">Best Used For</a> : Ribosomal database project	Access: Uncertain	QC: Uncertain/deprecated check	URL: <a href="https://rdp.cme.msu.edu">https://rdp.cme.msu.edu</a>
Greengenes2	16S reference database	Microbial taxonomy/16S	<a href="#">Best Used For</a> : Updated replacement for Greengenes	Free	QC: Replacement noted	URL: <a href="https://greengenes2.ucsd.edu">https://greengenes2.ucsd.edu</a>
GTDB	Genome taxonomy DB	Microbial taxonomy	<a href="#">Best Used For</a> : Genome-based taxonomy	Free	Full-depth retained	URL: <a href="https://gtdb.ecogenomic.org">https://gtdb.ecogenomic.org</a>

U. Taxonomy

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
NCBI Taxonomy	Taxonomy database	Organisms	<a href="#">Best Used For</a> : NCBI organism taxonomy IDs	Free	Full-depth retained	URL: <a href="https://www.ncbi.nlm.nih.gov/taxonomy">https://www.ncbi.nlm.nih.gov/taxonomy</a>
Catalogue of Life	Species catalogue	Taxonomy	<a href="#">Best Used For</a> : Species checklist	Free	Index entry	URL: <a href="https://www.catalogueoflife.org">https://www.catalogueoflife.org</a>

V. AMPs/peptides

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
DRAMP	AMP database	Antimicrobial peptides	<a href="#">Best Used For</a> : Natural/synthetic/patented AMP records	Free	Full-depth retained	URL: <a href="https://dramp.cpu-bioinfor.org">https://dramp.cpu-bioinfor.org</a>
APD/APD6	AMP database	Antimicrobial peptides	<a href="#">Best Used For</a> : Curated AMP data; APD6 current	Free	Corrected	URL: <a href="https://aps.unmc.edu">https://aps.unmc.edu</a>
dbAMP	AMP database/tool	Antimicrobial peptides	<a href="#">Best Used For</a> : AMP sequences/prediction; verify current status	Free/uncertain	QC: Requires verification	URL: <a href="https://awi.cuhk.edu.cn/dbAMP">https://awi.cuhk.edu.cn/dbAMP</a>
CAMP4	AMP database/tool	Antimicrobial peptides	<a href="#">Best Used For</a> : Current CAMP release with AMP data/tools	Free	Corrected	URL: <a href="https://camp.bicnirrh.res.in">https://camp.bicnirrh.res.in</a>
DBAASP	AMP database	Antimicrobial peptides	<a href="#">Best Used For</a> : Structure/activity/assay context	Free/mixed	Full-depth retained	URL: <a href="https://dbaasp.org">https://dbaasp.org</a>

W. Cancer genomics

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
TCGA/GDC	Cancer genomics portal	Cancer multi-omics	<a href="#">Best Used For</a> : TCGA and GDC cancer data	Free/controlled raw	Full-depth retained	URL: <a href="https://portal.gdc.cancer.gov">https://portal.gdc.cancer.gov</a>
cBioPortal	Cancer genomics visualization portal	Cancer genomics	<a href="#">Best Used For</a> : Interactive cancer genomics analysis	Free	Full-depth retained	URL: <a href="https://www.cbioportal.org">https://www.cbioportal.org</a>
COSMIC	Cancer mutation database	Somatic mutations	<a href="#">Best Used For</a> : Somatic cancer mutations	Access: Licensing restrictions	Restricted/flagged	URL: <a href="https://cancer.sanger.ac.uk/cosmic">https://cancer.sanger.ac.uk/cosmic</a>
ICGC	Cancer genome consortium portal	Cancer genomics	<a href="#">Best Used For</a> : International cancer genome data	Free/mixed	Index entry	URL: <a href="https://dcc.icgc.org">https://dcc.icgc.org</a>
OncoKB	Precision oncology KB	Cancer variants/drugs	<a href="#">Best Used For</a> : Actionable cancer alterations	Free academic/licensed	Restricted/flagged	URL: <a href="https://www.oncokb.org">https://www.oncokb.org</a>

X. Model organisms

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
FlyBase	Model organism DB	Drosophila	<a href="#">Best Used For</a> : Fruit fly genetics/genomics	Free	Full-depth retained	<a href="https://flybase.org">https://flybase.org</a>
WormBase	Model organism DB	C. elegans	<a href="#">Best Used For</a> : Worm genetics/genomics	Free	Full-depth retained	<a href="https://wormbase.org">https://wormbase.org</a>
MGI	Model organism DB	Mouse	<a href="#">Best Used For</a> : Mouse genetics/genomics	Free	Full-depth retained	<a href="https://www.informatics.jax.org">https://www.informatics.jax.org</a>
ZFIN	Model organism DB	Zebrafish	<a href="#">Best Used For</a> : Zebrafish genetics/genomics	Free	Full-depth retained	<a href="https://zfin.org">https://zfin.org</a>
SGD	Model organism DB	Yeast	<a href="#">Best Used For</a> : S. cerevisiae genetics/genomics	Free	Full-depth retained	<a href="https://www.yeastgenome.org">https://www.yeastgenome.org</a>
TAIR	Model organism DB	Arabidopsis	<a href="#">Best Used For</a> : Plant genetics/genomics	Access: Mixed/subscription	Restricted	<a href="https://www.arabidopsis.org">https://www.arabidopsis.org</a>



Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
PomBase	Model organism DB	Fission yeast	Best Used For: S. pombe genetics/genomics	Free	Index entry	<a href="https://www.pombase.org">https://www.pombase.org</a>

Y. FAIR/data standards

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
FAIRsharing	Standards/database registry	Data standards/repositories	FAIR standards and database registry	Free	Full-depth retained	<a href="https://fairsharing.org">https://fairsharing.org</a>
BioSamples	Sample metadata repository	Biological samples	Sample metadata records	Free	Index entry	<a href="https://www.ebi.ac.uk/biosamples">https://www.ebi.ac.uk/biosamples</a>
BioProject	Project metadata DB	Study metadata	Project/study accession records	Free	Index entry	<a href="https://www.ncbi.nlm.nih.gov/bioproject">https://www.ncbi.nlm.nih.gov/bioproject</a>
BioStudies	Study-linked data repository	Study data	One-stop study data archive	Free	Corrected for ArrayExpress	<a href="https://www.ebi.ac.uk/biostudies">https://www.ebi.ac.uk/biostudies</a>
Zenodo	General repository	Research outputs	Datasets/software/preprints	Free	Index entry	<a href="https://zenodo.org">https://zenodo.org</a>

Z. Directories/catalogs

Database/Resource	Resource Type	Domain	Best Used For	Access	QC	URL
NAR Database Collection	Database directory	Database catalog	Curated list of molecular biology databases	Free	Full-depth retained	<a href="https://academic.oup.com/nar/pages/molecular_biology_database_collection">https://academic.oup.com/nar/pages/molecular_biology_database_collection</a>
Database Commons	Database catalog	Database catalog	Global biological database registry	Free	Full-depth retained	<a href="https://ngdc.cncb.ac.cn/databasecommons">https://ngdc.cncb.ac.cn/databasecommons</a>
bio.tools	Tool/database registry	Bioinformatics tools/services	Bioinformatics tool registry	Free	Full-depth retained	<a href="https://bio.tools">https://bio.tools</a>
FAIRsharing Databases	Database/standard registry	FAIR resources	Databases, standards, policies	Free	Full-depth retained	<a href="https://fairsharing.org/databases">https://fairsharing.org/databases</a>
NCBI Handbook	Book/manual	NCBI resources	Free NCBI reference book	Free	Index entry	<a href="https://www.ncbi.nlm.nih.gov/books/NBK143764">https://www.ncbi.nlm.nih.gov/books/NBK143764</a>
GenScript Bioinformatics Tools	Tool collection	Practical bioinformatics tools	Calculators/tools, not database catalog	Free/commercial site	Classified as tools	<a href="https://www.genscript.com/tools.html">https://www.genscript.com/tools.html</a>

## Appendix B: Raw Links for Copy/Paste

Raw official URLs are grouped by category. These are intentionally not embedded hyperlinks so they can be copied into LinkedIn comments, notes, reference managers, or teaching handouts.

### A — General Portals

<b>NCBI:</b> <a href="https://www.ncbi.nlm.nih.gov">https://www.ncbi.nlm.nih.gov</a>	<b>EMBL-EBI:</b> <a href="https://www.ebi.ac.uk">https://www.ebi.ac.uk</a>	<b>DDBJ:</b> <a href="https://www.ddbj.nig.ac.jp">https://www.ddbj.nig.ac.jp</a>
<b>ExPASy:</b> <a href="https://www.expasy.org">https://www.expasy.org</a>	<b>Ensembl:</b> <a href="https://www.ensembl.org">https://www.ensembl.org</a>	

### B — Literature

<b>PubMed:</b> <a href="https://pubmed.ncbi.nlm.nih.gov">https://pubmed.ncbi.nlm.nih.gov</a>	<b>PubMed Central:</b> <a href="https://www.ncbi.nlm.nih.gov/pmc">https://www.ncbi.nlm.nih.gov/pmc</a>	<b>Europe PMC:</b> <a href="https://europepmc.org">https://europepmc.org</a>
<b>Google Scholar:</b> <a href="https://scholar.google.com">https://scholar.google.com</a>	<b>Semantic Scholar:</b> <a href="https://www.semanticscholar.org">https://www.semanticscholar.org</a>	<b>Scopus (subscription):</b> <a href="https://www.scopus.com">https://www.scopus.com</a>
<b>Web of Science (subscription):</b> <a href="https://www.webofscience.com">https://www.webofscience.com</a>		

### C — Nucleotide Sequences

<b>GenBank:</b> <a href="https://www.ncbi.nlm.nih.gov/genbank">https://www.ncbi.nlm.nih.gov/genbank</a>	<b>ENA:</b> <a href="https://www.ebi.ac.uk/ena">https://www.ebi.ac.uk/ena</a>	<b>DDBJ:</b> <a href="https://www.ddbj.nig.ac.jp">https://www.ddbj.nig.ac.jp</a>
<b>RefSeq:</b> <a href="https://www.ncbi.nlm.nih.gov/refseq">https://www.ncbi.nlm.nih.gov/refseq</a>	<b>NCBI Nucleotide:</b> <a href="https://www.ncbi.nlm.nih.gov/nucleotide">https://www.ncbi.nlm.nih.gov/nucleotide</a>	

### D — Similarity Tools

<b>NCBI BLAST:</b> <a href="https://blast.ncbi.nlm.nih.gov">https://blast.ncbi.nlm.nih.gov</a>	<b>EBI BLAST:</b> <a href="https://www.ebi.ac.uk/Tools/sss/ncbiblast">https://www.ebi.ac.uk/Tools/sss/ncbiblast</a>	<b>HMMER:</b> <a href="https://www.ebi.ac.uk/Tools/hmmer">https://www.ebi.ac.uk/Tools/hmmer</a>
<b>DIAMOND:</b> <a href="https://github.com/bbuchfink/diamond">https://github.com/bbuchfink/diamond</a>	<b>FASTA (EBI):</b> <a href="https://www.ebi.ac.uk/Tools/sss/fasta">https://www.ebi.ac.uk/Tools/sss/fasta</a>	

### E — Genome Browsers

<b>Ensembl:</b> <a href="https://www.ensembl.org">https://www.ensembl.org</a>	<b>UCSC Genome Browser:</b> <a href="https://genome.ucsc.edu">https://genome.ucsc.edu</a>	<b>NCBI GDV:</b> <a href="https://www.ncbi.nlm.nih.gov/genome/gdv">https://www.ncbi.nlm.nih.gov/genome/gdv</a>
<b>JBrowse:</b> <a href="https://jbrowse.org">https://jbrowse.org</a>	<b>IGV:</b> <a href="https://igv.org">https://igv.org</a>	

### F — Gene Databases

<b>NCBI Gene:</b> <a href="https://www.ncbi.nlm.nih.gov/gene">https://www.ncbi.nlm.nih.gov/gene</a>	<b>GeneCards:</b> <a href="https://www.genecards.org">https://www.genecards.org</a>	<b>HGNC:</b> <a href="https://www.genenames.org">https://www.genenames.org</a>
<b>OMIM:</b> <a href="https://www.omim.org">https://www.omim.org</a>	<b>Ensembl Genes:</b> <a href="https://www.ensembl.org">https://www.ensembl.org</a>	

**G — Transcriptomics**

<b>GEO:</b> <a href="https://www.ncbi.nlm.nih.gov/geo">https://www.ncbi.nlm.nih.gov/geo</a>	<b>ArrayExpress/BioStudies:</b> <a href="https://www.ebi.ac.uk/biostudies/arrayexpress">https://www.ebi.ac.uk/biostudies/arrayexpress</a>	<b>Expression Atlas:</b> <a href="https://www.ebi.ac.uk/gxa">https://www.ebi.ac.uk/gxa</a>
<b>GTEX:</b> <a href="https://gtexportal.org">https://gtexportal.org</a>	<b>SRA:</b> <a href="https://www.ncbi.nlm.nih.gov/sra">https://www.ncbi.nlm.nih.gov/sra</a>	

**H — Raw Sequencing**

<b>SRA (NCBI):</b> <a href="https://www.ncbi.nlm.nih.gov/sra">https://www.ncbi.nlm.nih.gov/sra</a>	<b>ENA Read Archive:</b> <a href="https://www.ebi.ac.uk/ena">https://www.ebi.ac.uk/ena</a>	<b>DRA (DDBJ):</b> <a href="https://www.ddbj.nig.ac.jp">https://www.ddbj.nig.ac.jp</a>
---	---	---

**I — Protein Sequence/Function**

<b>UniProt:</b> <a href="https://www.uniprot.org">https://www.uniprot.org</a>	<b>Swiss-Prot (reviewed):</b> <a href="https://www.uniprot.org/uniprotkb?query=reviewed:true">https://www.uniprot.org/uniprotkb?query=reviewed:true</a>	<b>InterPro:</b> <a href="https://www.ebi.ac.uk/interpro">https://www.ebi.ac.uk/interpro</a>
<b>PROSITE:</b> <a href="https://prosite.expasy.org">https://prosite.expasy.org</a>	<b>SMART:</b> <a href="http://smart.embl.de">http://smart.embl.de</a>	

**J — Protein Domains**

<b>Pfam (via InterPro):</b> <a href="https://www.ebi.ac.uk/interpro/entry/pfam/">https://www.ebi.ac.uk/interpro/entry/pfam/</a>	<b>CDD:</b> <a href="https://www.ncbi.nlm.nih.gov/cdd">https://www.ncbi.nlm.nih.gov/cdd</a>	<b>SUPERFAMILY:</b> <a href="https://supfam.org">https://supfam.org</a>
<b>PRINTS (legacy):</b> <a href="http://130.88.97.239/PRINTS/index.php">http://130.88.97.239/PRINTS/index.php</a>		

**K — Protein Structures**

<b>RCSB PDB:</b> <a href="https://www.rcsb.org">https://www.rcsb.org</a>	<b>PDBe:</b> <a href="https://www.ebi.ac.uk/pdbe">https://www.ebi.ac.uk/pdbe</a>	<b>PDBj:</b> <a href="https://pd bj.org">https://pd bj.org</a>
<b>AlphaFold DB:</b> <a href="https://alphafold.ebi.ac.uk">https://alphafold.ebi.ac.uk</a>	<b>SWISS-MODEL Repository:</b> <a href="https://swissmodel.expasy.org/repository">https://swissmodel.expasy.org/repository</a>	

**L — Variants**

<b>dbSNP:</b> <a href="https://www.ncbi.nlm.nih.gov/snp">https://www.ncbi.nlm.nih.gov/snp</a>	<b>ClinVar:</b> <a href="https://www.ncbi.nlm.nih.gov/clinvar">https://www.ncbi.nlm.nih.gov/clinvar</a>	<b>gnomAD:</b> <a href="https://gnomad.broadinstitute.org">https://gnomad.broadinstitute.org</a>
<b>COSMIC:</b> <a href="https://cancer.sanger.ac.uk/cosmic">https://cancer.sanger.ac.uk/cosmic</a>	<b>Ensembl VEP:</b> <a href="https://www.ensembl.org/vep">https://www.ensembl.org/vep</a>	<b>ClinGen:</b> <a href="https://clinicalgenome.org">https://clinicalgenome.org</a>
<b>LOVD:</b> <a href="https://www.lovd.nl">https://www.lovd.nl</a>		

**M — Disease/Clinical**

<b>OMIM:</b> <a href="https://www.omim.org">https://www.omim.org</a>	<b>Orphanet:</b> <a href="https://www.orpha.net">https://www.orpha.net</a>	<b>DisGeNET:</b> <a href="https://www.disgenet.org">https://www.disgenet.org</a>
---	---	---

**GWAS Catalog:**<https://www.ebi.ac.uk/gwas>**MalaCards:**<https://www.malacards.org>**ClinGen:**<https://clinicalgenome.org>**N — Pathways****KEGG:**<https://www.kegg.jp>**Reactome:**<https://reactome.org>**WikiPathways:**<https://www.wikipathways.org>**BioCyc:**<https://biocyc.org>**STRING:**<https://string-db.org>**O — Protein Interactions****STRING:**<https://string-db.org>**BioGRID:**<https://thebiogrid.org>**IntAct:**<https://www.ebi.ac.uk/intact>**DIP (legacy):**<https://dip.doe-mbi.ucla.edu>**MINT (legacy):**<https://mint.bio.uniroma2.it>**P — Drugs/Compounds****DrugBank (tiered):**<https://go.drugbank.com>**ChEMBL:**<https://www.ebi.ac.uk/chembl>**PubChem:**<https://pubchem.ncbi.nlm.nih.gov>**BindingDB:**<https://www.bindingdb.org>**TTD:**<https://db.idrblab.net/ttd>**Q — Ontologies****Gene Ontology (GO):**<https://geneontology.org>**HPO:**<https://hpo.jax.org>**Disease Ontology:**<https://disease-ontology.org>**Uberon:**<https://obophenotype.github.io/uberon>**MeSH:**<https://www.nlm.nih.gov/mesh>**Sequence Ontology:**<http://www.sequenceontology.org>**R — Epigenomics****ENCODE:**<https://www.encodeproject.org>**Roadmap Epigenomics:**<https://www.roadmapepigenomics.org>**Cistrome:**<https://cistrome.org>**JASPAR:**<https://jaspar.elixir.no>**ChIP-Atlas:**<https://chip-atlas.org>**S — Single-cell/Spatial****Human Cell Atlas:**<https://www.humancellatlas.org>**CellxGene:**<https://cellxgene.cziscience.com>**PanglaoDB:**<https://panglaoDB.se>**Single Cell Expression Atlas:**<https://www.ebi.ac.uk/gxa/sc>**T — Microbiome****MGnify:****SILVA:****GTDB:**

<https://www.ebi.ac.uk/metagenomics><https://www.arb-silva.de><https://gtdb.ecogenomic.org>

RDP (limited updates):

<https://rdp.cme.msu.edu>

MG-RAST:

<https://www.mg-rast.org>

## U — Taxonomy

NCBI Taxonomy:

<https://www.ncbi.nlm.nih.gov/taxonomy>

UniProt Taxonomy:

<https://www.uniprot.org/taxonomy>

Catalogue of Life:

<https://www.catalogueoflife.org>

## V — AMPs/Peptides

DRAMP:

<https://dramp.cpu-bioinfor.org>

APD/APD6:

<https://aps.unmc.edu>

CAMPR4:

<https://camp.bicnirrh.res.in>

DBAASP:

<https://dbaasp.org>

dbAMP (verify status):

<https://awi.cuhk.edu.cn/dbAMP>

## W — Cancer Genomics

TCGA / GDC Portal:

<https://portal.gdc.cancer.gov>

cBioPortal:

<https://www.cbioportal.org>

COSMIC:

<https://cancer.sanger.ac.uk/cosmic>

ICGC:

<https://dcc.icgc.org>

OncoKB (tiered):

<https://www.oncokb.org>

## X — Model Organisms

FlyBase (Drosophila):

<https://flybase.org>

WormBase (C. elegans):

<https://wormbase.org>

MGI (Mouse):

<https://www.informatics.jax.org>

ZFIN (Zebrafish):

<https://zfin.org>

SGD (S. cerevisiae):

<https://www.yeastgenome.org>

TAIR (Arabidopsis):

<https://www.arabidopsis.org>

PomBase (S. pombe):

<https://www.pombase.org>

Alliance of Genome Resources:

<https://www.alliancegenome.org>

## Y — FAIR/Data Standards

FAIRsharing:

<https://fairsharing.org>

BioSamples:

<https://www.ebi.ac.uk/biosamples>

BioProject:

<https://www.ncbi.nlm.nih.gov/bioproject>

BioStudies:

<https://www.ebi.ac.uk/biostudies>

Zenodo:

<https://zenodo.org>

## Z — Directories/Catalogs

NAR Database Collection:

[https://academic.oup.com/nar/pages/molecular\\_biology\\_database\\_collection](https://academic.oup.com/nar/pages/molecular_biology_database_collection)

bio.tools:

<https://bio.tools>

Database Commons:

<https://ngdc.cncb.ac.cn/databasecommons>

FAIRsharing:

<https://fairsharing.org>



## Appendix C: Deprecated, Limited-Update, and Restricted Resources

The following resources have deprecated, limited, or uncertain status as of 2026. Always verify the status of any database before using it in a new analysis.

Resource	URL	Category	Status / Note	Current Status
DIP (Database of Interacting Proteins)	<a href="https://dip.doe-mbi.ucla.edu">https://dip.doe-mbi.ucla.edu</a>	PPI	Limited updates since ~2017; use BioGRID or IntAct instead	Uncertain
MINT (Molecular INTERaction database)	<a href="https://mint.bio.uniroma2.it">https://mint.bio.uniroma2.it</a>	PPI	Limited updates in recent years; data accessible through IntAct	Uncertain
RDP (Ribosomal Database Project)	<a href="https://rdp.cme.msu.edu">https://rdp.cme.msu.edu</a>	rRNA taxonomy	Limited updates since ~2014; use SILVA 138.1 instead	Active but outdated
Greengenes (original)	<a href="https://greengenes.secondgenome.com">https://greengenes.secondgenome.com</a>	16S taxonomy	Retired; use Greengenes2 ( <a href="https://greengenes2.ucsd.edu">https://greengenes2.ucsd.edu</a> ) instead	Retired
Pfam standalone website	<a href="https://pfam.xfam.org">https://pfam.xfam.org</a>	Protein domains	Retired 2022; now integrated into InterPro at <a href="https://www.ebi.ac.uk/interpro">https://www.ebi.ac.uk/interpro</a>	Retired — redirect
ArrayExpress (standalone)	<a href="https://www.ebi.ac.uk/arrayexpress">https://www.ebi.ac.uk/arrayexpress</a>	Expression data	Integrated into BioStudies; URL redirects to BioStudies	Redirected
dbAMP	<a href="https://awi.cuhk.edu.cn/dbAMP">https://awi.cuhk.edu.cn/dbAMP</a>	AMP database	Uncertain operational status; verify before use	Uncertain
Roadmap Epigenomics (new data)	<a href="https://www.roadmapepigenomics.org">https://www.roadmapepigenomics.org</a>	Epigenomics	Data collection COMPLETE; no new data being generated (existing data still available)	Static — no new data
HGMD (free tier)	<a href="https://www.hgmd.cf.ac.uk">https://www.hgmd.cf.ac.uk</a>	Disease mutations	Full database requires subscription; free tier severely limited	Subscription required
NCBI UniGene	N/A	Gene expression	Archived; no longer maintained; use GEO or Expression Atlas	Archived



Resource	URL	Category	Status / Note	Current Status
PRINTS (standalone)	<a href="http://130.88.97.239/PRINTS">http://130.88.97.239/PRINTS</a>	Protein fingerprints	Not actively maintained since ~2012; accessible through InterPro	Legacy/InterPro only
SUPERFAMILY (standalone)	<a href="https://supfam.org">https://supfam.org</a>	Structural domains	Reduced maintenance; accessible through InterPro (SUPERFAMILY member)	Limited updates

**General rule:** Before using any database not covered in the main atlas sections, check:

- (1) When was it last updated?
- (2) Is the website accessible?
- (3) Is there a recommended replacement? Always verify operational status before investing time in an analysis.

## Appendix D: Glossary

Accession number: A stable identifier assigned to a database record, such as a sequence, study, structure, variant, or sample.

Term	Definition
<b>Accession number</b>	A stable, unique identifier assigned to a biological database record (sequence, structure, variant, sample, study). Examples: NM_000546.6 (RefSeq mRNA), P04637 (UniProt protein), GSE12345 (GEO series).
<b>API (Application Programming Interface)</b>	A programmatic interface allowing software to query or retrieve data from a database without using the web browser. Most major databases provide REST APIs returning JSON or XML.
<b>ASV (Amplicon Sequence Variant)</b>	An exact sequence variant derived from amplicon sequencing (e.g., 16S rRNA), representing a unique biological sequence without clustering. Preferred over OTUs for precision.
<b>Benchmark dataset</b>	A curated collection of biological data with known, validated properties used to evaluate computational methods. Should only be used for evaluation, NOT for training machine learning models.
<b>BLAST E-value</b>	The Expected Value — the number of alignments with a given score expected by chance in a database of the given size. Lower E-values indicate more significant hits. $E < 1e-5$ typically indicates homology.
<b>Curation</b>	The human or computational process of reviewing, annotating, validating, and standardizing database records. Manually curated data is more reliable but covers fewer sequences.
<b>FAIR principles</b>	Findable, Accessible, Interoperable, Reusable — a framework for scientific data management published by Wilkinson et al. (2016) in Scientific Data.
<b>FTP (File Transfer Protocol)</b>	A network protocol used for bulk downloading of large database files. Most major databases provide FTP access for complete database downloads.
<b>Germline variant</b>	A genetic variant present in all cells of an organism and heritable. Detected in blood/saliva for clinical genomics. See: ClinVar, gnomAD, dbSNP.
<b>HGVS nomenclature</b>	Human Genome Variation Society nomenclature — standardized notation for describing sequence variants. Example: NM_000546.6:c.817C>T for a TP53 variant at the mRNA level.
<b>InChIKey</b>	A fixed-length (27-character) hash of the International Chemical Identifier (InChI). Used as a universal chemical identifier across databases. Example: BSYNRYMUTXBXSQ-UHFFFAOYSA-N (aspirin).
<b>MAG (Metagenome-Assembled Genome)</b>	A genome sequence assembled from metagenomic data, representing a single organism. Quality varies; high-quality MAGs have $\geq 90\%$ completeness and $\leq 5\%$ contamination.
<b>MIC (Minimum Inhibitory Concentration)</b>	The lowest concentration of an antimicrobial agent that prevents visible growth of a microorganism. Not directly comparable across different assay conditions.

Term	Definition
<b>OTU (Operational Taxonomic Unit)</b>	A cluster of similar sequences (typically $\geq 97\%$ identity for 16S rRNA) used to represent microbial taxa. Now largely replaced by ASVs for better precision.
<b>pLDDT</b>	Predicted Local Distance Difference Test — AlphaFold2's per-residue confidence metric (0–100). $>90$ = very high confidence; $<50$ = likely disordered. Stored in PDB B-factor column.
<b>Primary database</b>	A database storing original, experimentally determined data submitted directly by researchers. Minimal post-submission curation. Examples: GenBank, PDB, GEO, SRA.
<b>PWM (Position Weight Matrix)</b>	A matrix representing the probability of each nucleotide at each position in a transcription factor binding site. Used by JASPAR and other TF binding databases.
<b>REST API</b>	Representational State Transfer API — web-based API using HTTP requests (GET, POST) to retrieve data. Most modern bioinformatics databases provide REST APIs returning JSON or XML.
<b>rsID</b>	Reference SNP cluster ID — a stable identifier assigned by dbSNP to known genetic variants. Example: rs1234567. Used as a universal variant identifier across databases.
<b>Secondary database</b>	A database built by integrating, re-annotating, and curating data from primary sources. Higher reliability but may lag primary databases in coverage. Examples: UniProt/Swiss-Prot, Pfam, KEGG.
<b>SMILES</b>	Simplified Molecular Input Line Entry System — a text representation of chemical structures. Used by PubChem, ChEMBL, and other chemical databases.
<b>Somatic mutation</b>	A genetic mutation acquired during an organism's lifetime, present only in specific cells (e.g., tumor cells). Not heritable. See: COSMIC, TCGA, cBioPortal.
<b>SRA Toolkit</b>	A suite of command-line tools for downloading and converting raw sequencing data from NCBI SRA. Key tools: prefetch, fasterq-dump, sam-dump.
<b>TF binding motif</b>	The DNA sequence pattern recognized by a transcription factor. Represented as a PWM in databases like JASPAR. Used to predict TF binding sites in regulatory regions.
<b>VUS</b>	Variant of Uncertain Significance — a ClinVar classification meaning there is insufficient evidence to classify the variant as pathogenic or benign. NOT evidence of pathogenicity.
<b>Versioned accession</b>	An accession number with a version suffix indicating that the sequence or record has been updated. Example: NM_000546.6 (version 6 of the TP53 mRNA). Always record the full versioned accession.

## About This Atlas

The Bioinformatics Databases and Data Resources Reference Atlas is a comprehensive reference guide for bioinformatics databases, tools, and resources. It is designed for students, researchers, clinicians, and bioinformaticians who need to navigate the complex landscape of biological data resources.

## How to Cite This Atlas

When citing this atlas in publications, please use the following format:

Mohamed Mostafa Mohamed. (2026). Bioinformatics Databases and Data Resources Reference Atlas: A Practical Guide to Biological Data Resources, Repositories, Knowledgebases, Ontologies, Tools, Their Uses, Strengths, Limitations, and Scientific Applications (First Edition – Integrated Master Edition) [Computer software]. Self-published. <https://doi.org/10.5281/zenodo.20533722>

## Recommended citation note:

Because database URLs, access policies, APIs, and release versions change frequently, users should also cite the specific databases and database releases used in their analyses.

## Acknowledgments

This atlas draws on information from hundreds of primary database publications, documentation pages, and community resources. We acknowledge the work of the many database developers, curators, and communities who maintain these essential resources for the scientific community.

## Disclaimer

This atlas is provided for educational and reference purposes. Database URLs, access policies, and content change continuously. Always verify current information directly from the database before beginning a project. The authors make no warranty regarding the accuracy or completeness of the information provided.

All database entries are marked with a verification date of May 2026. Information may have changed since this date.