

5. Information databases in experimental biology

NCBI - National Center for Biotechnology Information



Entrez is a molecular biology database system that provides **integrated access** to **nucleotide and protein sequence data**, **gene-centered and genomic mapping information**, **3D structure data**, **PubMed MEDLINE**, and more.

The system is produced by the **National Center for Biotechnology Information (NCBI)** and is available via the Internet.

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

All Databases Search

NCBI Home
 Resource List (A-Z)
 All Resources
 Chemicals & Bioassays
 Data & Software
 DNA & RNA
 Domains & Structures
 Genes & Expression
 Genetics & Medicine
 Genomes & Maps
 Homology
 Literature
 Proteins
 Sequence Analysis
 Taxonomy
 Training & Tutorials
 Variation

Welcome to NCBI
 The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
 Deposit data or manuscripts into NCBI databases

Download
 Transfer NCBI data to your computer

Learn
 Find help documents, attend a class or watch a tutorial

Develop
 Use NCBI APIs and code libraries to build applications

Analyze
 Identify an NCBI tool for your data analysis task

Research
 Explore NCBI research and collaborative projects

COVID-19 Information
[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

Popular Resources
 PubMed
 Bookshelf
 PubMed Central
 BLAST
 Nucleotide
 Genome
 SNP
 Gene
 Protein
 PubChem

NCBI News & Blog
 View Global Distribution of Microbial AMR Genes with the New MicroBIGG-E Map
06 Dec 2023
 Graphical display provides new way to explore data! NCBI's Pathogen Detection
 Coming Soon! Updates to the ClinVar Website
28 Nov 2023
 In order to support the inclusion of submitted somatic variation data, we are
 Faster and Focused Searches with BLAST+ 2.15.0
21 Nov 2023
 New version now available! Do you use NCBI's standalone BLAST tool?
[More...](#)

5. Information databases in experimental biology

Collecting sequences? For what is it good for?

INSDC

[About](#) [Policy](#) [Advisors](#) [Submitting Standards](#) [News & Announcements](#) [Documents](#)



The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between [DDBJ](#), [EMBL-EBI](#) and [NCBI](#).

INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.



NCBI



5. Information databases in experimental biology

NCBI Gene bank - the core of the sequence database

Nucleic Acids Research, 2024, **52**, D134–D137
<https://doi.org/10.1093/nar/gkad903>
 Advance access publication date: 27 October 2023
Database issue



GenBank 2024 Update

Eric W. Sayers ^{ID}*, Mark Cavanaugh, Karen Clark, Kim D. Pruitt, Stephen T. Sherry, Linda Yankie and Ilene Karsch-Mizrachi ^{ID}

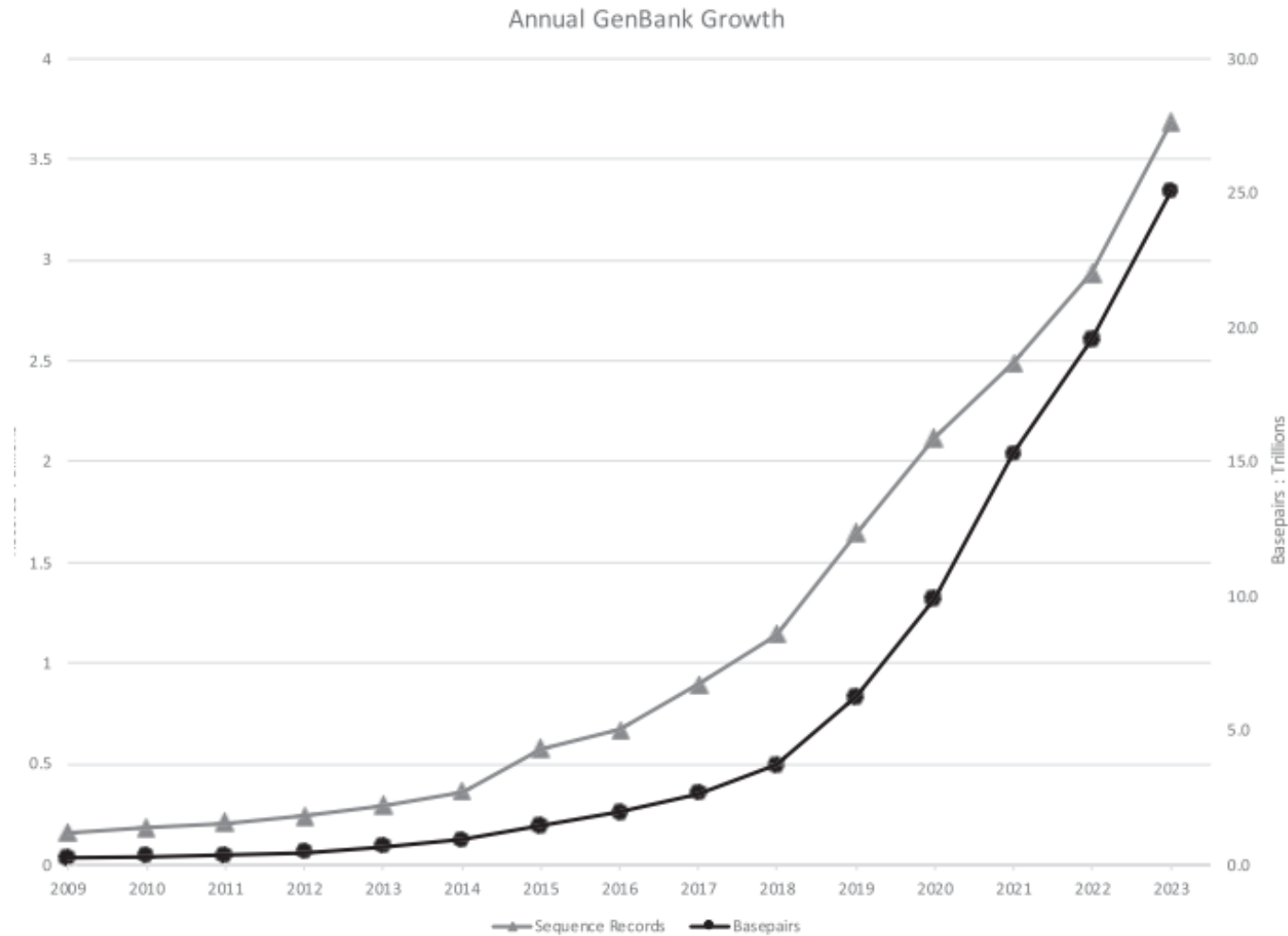
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

GenBank® (<https://www.ncbi.nlm.nih.gov/genbank/>) is a comprehensive, public database that contains **25 trillion base pairs from over 3.7 billion nucleotide sequences for 557 000 formally described species**. Daily data exchange with the **European Nucleotide Archive (ENA)** and the **DNA Data Bank of Japan (DDBJ)** ensures worldwide coverage. Recent updates include policies for including spatio-temporal metadata, clarified documentation for GenBank data processing, enhanced foreign contamination screening tools, new processes in the Submission Portal, migration of Entrez Genome and Assembly displays into NCBI Datasets, and the impending retirement of tbl2asn, replaced by table2asn.

5. Information databases in experimental biology

NCBI - National Center for Biotechnology Information



5. Information databases in experimental biology

NCBI - National Center for Biotechnology Information

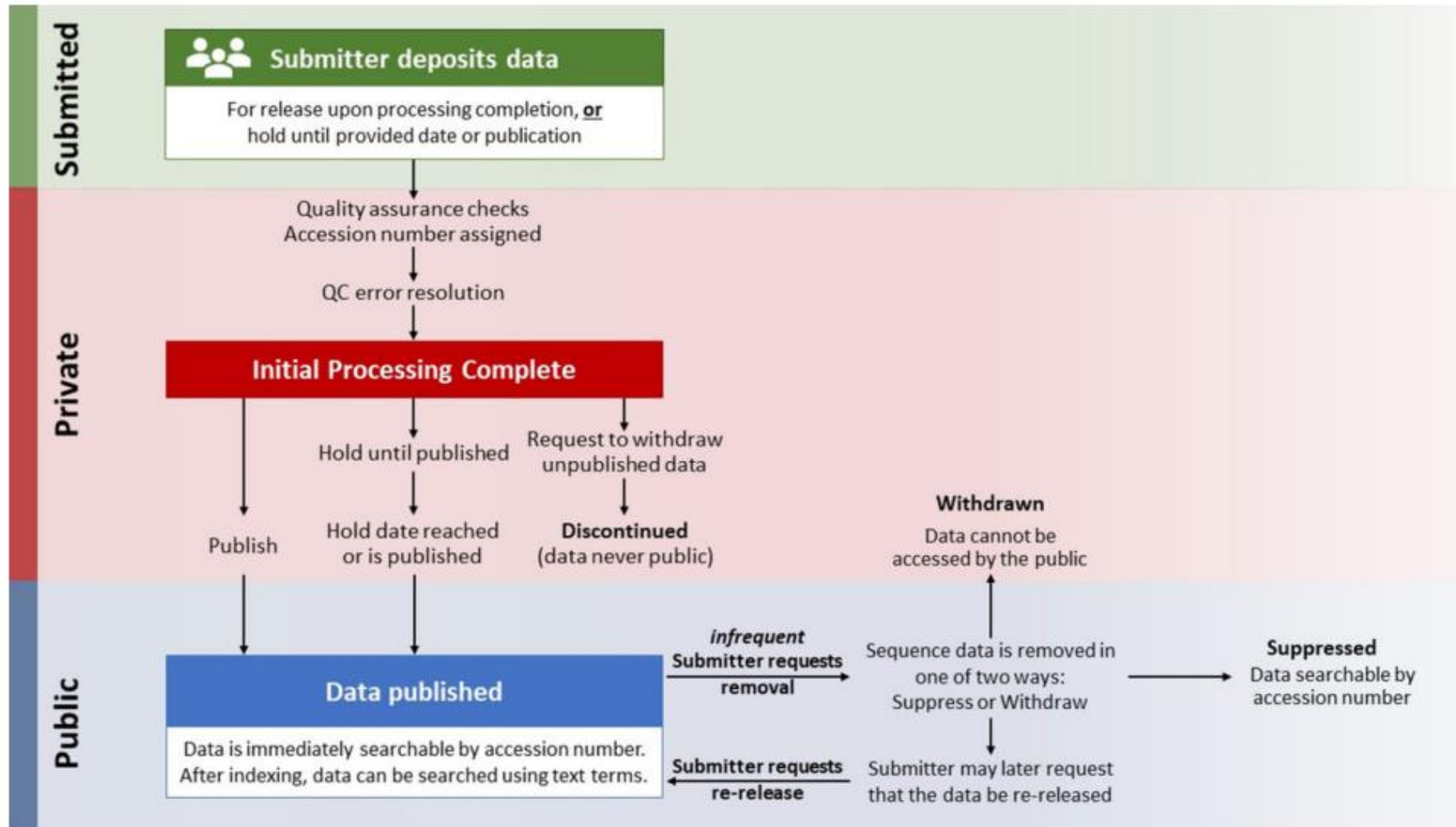
Table 1. GenBank divisions

Division	Description	Records	Base pairs
WGS	Whole genome shotgun data	2 631 493 489	22 294 446 104 543
TSA	Transcriptome shotgun data	703 290 753	660 355 737 229
PLN	Plants	7 267 977	607 785 846 175
INV	Invertebrates	11 706 120	562 054 124 309
VRL	Viruses	12 117 023	253 624 539 927
BCT	Bacteria	2 588 641	204 717 329 672
VRT	Other vertebrates	3 725 098	134 590 232 569
ROD	Rodents	646 856	96 273 380 697
MAM	Other mammals	718 185	71 142 498 026
TLS	Targeted loci studies	124 421 006	48 289 699 026
EST	Expressed sequence tags	78 009 188	43 476 079 939
PAT	Patent sequences	57 387 889	32 777 953 076
HTG	High-throughput genomic	177 352	27 825 565 592
GSS	Genome survey sequences	41 103 736	26 390 262 337
PRI	Primates	1 144 158	15 944 087 492
ENV	Environmental samples	10 165 304	10 294 243 239
SYN	Synthetic	308 231	8 276 291 009
PHG	Phages	35 373	1 319 896 723
HTC	High-throughput cDNA	651 488	741 162 888
STS	Sequence tagged sites	1 346 988	640 923 137
UNA	Unannotated	760	4 530 817
TOTAL	All GenBank sequences	3 688 305 615	25 100 970 488 422

Release 257 (August 2023).

5. Information databases in experimental biology

NCBI - National Center for Biotechnology Information



5. Information databases in experimental biology

Nucleic acid Research Journal list of databases

The most comprehensive listing of databases used by biologists. DNA, RNA, nucleic acids modifications, proteins, protein modifications, low molecular weight compounds, gene expression databases, various biomolecule interaction databases, as well as number of tools for prediction of structure-function characteristics of biomolecules

OXFORD
ACADEMIC

Journals

You are here: [NAR Journal Home](#) » [Database Summary Paper Categories](#)

NAR Database Summary Paper Category List

[Nucleotide Sequence Databases](#)
[RNA sequence databases](#)
[Protein sequence databases](#)
[Structure Databases](#)
[Genomics Databases \(non-vertebrate\)](#)
[Metabolic and Signaling Pathways](#)
[Human and other Vertebrate Genomes](#)
[Human Genes and Diseases](#)
[Microarray Data and other Gene Expression Databases](#)
[Proteomics Resources](#)
[Other Molecular Biology Databases](#)
[Organelle databases](#)
[Plant databases](#)
[Immunological databases](#)
[Cell biology](#)

[► Compilation Paper](#)
[► Category List](#)
[► Alphabetical List](#)
[► Category/Paper List](#)
[► Search Summary Papers](#)

Oxford University Press is not responsible for the content of external internet sites

[► Compilation Paper](#)
[► Category List](#)
[► Alphabetical List](#)
[► Category/Paper List](#)
[► Search Summary Papers](#)



Volume 52, Issue D1
5 January 2024
[Cover image](#)
ISSN 0305-1048
EISSN 1362-4962

EDITORIAL

MAJOR MULTI-DATABASE
RESOURCES

NUCLEIC ACID SEQUENCE, STRUCTURE AND REGULATION

PROTEIN SEQUENCE AND
STRUCTURE, MOTIFS AND
DOMAINS

METABOLIC AND SIGNALLING
PATHWAYS, ENZYMES

VIRUSES, BACTERIA, PROTOZOA
AND FUNGI

HUMAN GENOME, MODEL
ORGANISMS, COMPARATIVE
GENOMICS

GENOMIC VARIATION, DISEASES
AND DRUGS

PLANTS

OTHER DATABASES

Volume 52, Issue D1, 5 January 2024

Page 1 of 2

1 2 Next

EDITORIAL

The 2024 *Nucleic Acids Research* database issue and the online molecular biology database collection [🔗](#)

Daniel J Rigden and Xosé M Fernández

Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D1–D9,
<https://doi.org/10.1093/nar/gkad1173>

[Abstract](#) ▼ [View article](#)

MAJOR MULTI-DATABASE RESOURCES

EMBL's European Bioinformatics Institute (EMBL-EBI) in 2023 [🔗](#)
Matthew Thakur and others

Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D10–D17,
<https://doi.org/10.1093/nar/gkad1088>



[Abstracts](#) ▼ [View article](#)

Database Resources of the National Genomics Data Center, China
National Center for Bioinformation in 2024 [🔗](#)

5. Information databases in experimental biology

All model organisms have their own scientific community-curated databases

In plant experimental biology, good examples is [TAIR](#) (The Arabidopsis Information Resource) curating the information resources for *Arabidopsis thaliana*

The screenshot shows the TAIR website interface. At the top, there is a search bar with the text "Enter search text" and a dropdown menu for "Gene". Below the search bar is a navigation menu with options: "Advanced Search", "Browse", "Tools", "Portals", "Download", "Submit", "News", and "Stocks".

The main content area features a large green banner for the "NAASC Arabidopsis Community Awards". The banner includes the NAASC logo (The North American Arabidopsis Steering Committee) and the text: "Open to all in the global Arabidopsis Community! Nominate your colleagues, trainees, or yourself to recognize excellence in the Arabidopsis community!".

Key details from the banner:

- Nominations***: Due Jan. 31, 2025, 200-500 words.
- Full packages for Round 2****: Due April 30, 2025, CV, Statement, 3 support letters.
- NOMINATIONS INVITED IN 3 AWARD CATEGORIES:**
 - P. N. Benfey Arabidopsis Community Lifetime Achievement
 - Dissemination of Arabidopsis Knowledge (early & later career awards)

Additional information: "Full details including eligibility & selection criteria: bit.ly/3DtjNEi".

Footnote: "**Self-nominations accepted. **Nominators for candidates advancing to Round 2 will be notified by Feb. 28, 2025 to compile a full package".

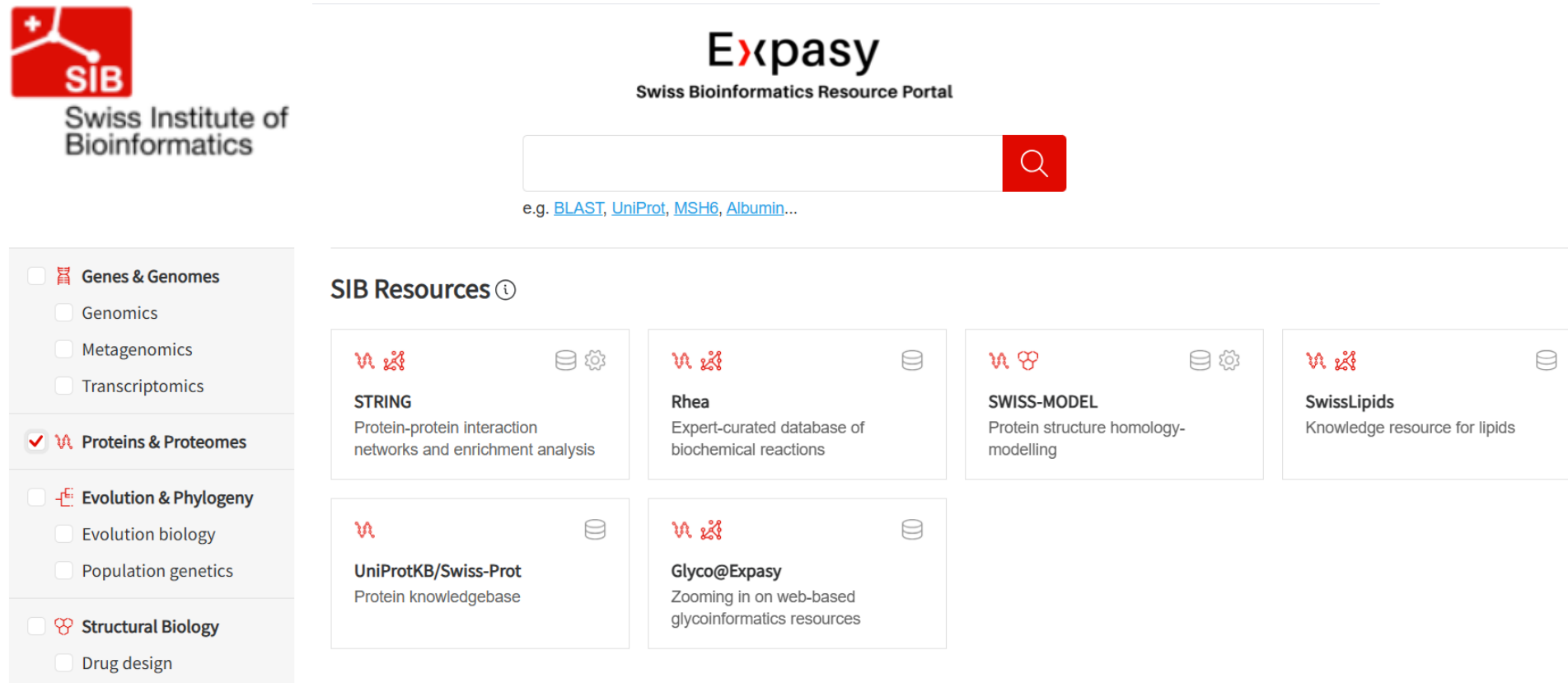
On the right side of the page, there is a "Breaking News" section with two items:

- NAASC Community Award Nominations Open** [Dec 20, 2024]: Nominations for NAASC community awards are open and being accepted until January 31, 2025.
- 25th Anniversary of TAIR** [Nov 22, 2024]: TAIR officially turned 25 on October 1, 2024! Read our [blog post](#) covering the highlights from the last 25 years.

5. Information databases in experimental biology


Expasy - Swiss Bioinformatics Resource Portal

The focus of this resource portal is on proteins, which are covered in [UniProtKB](#)/Swiss-Prot well-curated database. It contains [numerous tools](#), which experimental biologists use every day



SIB
Swiss Institute of Bioinformatics

Expasy
Swiss Bioinformatics Resource Portal

Search bar: 

e.g. [BLAST](#), [UniProt](#), [MSH6](#), [Albumin](#)...

SIB Resources ⓘ

- STRING**
Protein-protein interaction networks and enrichment analysis
- Rhea**
Expert-curated database of biochemical reactions
- SWISS-MODEL**
Protein structure homology-modelling
- SwissLipids**
Knowledge resource for lipids
- UniProtKB/Swiss-Prot**
Protein knowledgebase
- Glyco@Expasy**
Zooming in on web-based glycoinformatics resources

Navigation Menu:

- Genes & Genomes
 - Genomics
 - Metagenomics
 - Transcriptomics
- Proteins & Proteomes
- Evolution & Phylogeny
 - Evolution biology
 - Population genetics
- Structural Biology
 - Drug design

5. Information databases in experimental biology

BAR – great bio-analytic resource page for plant biologists

tools for functional genomics and proteomics in plants, including ePlant gene expression visualization database

The image shows a screenshot of the ePlant database interface on the left and a data visualization diagram on the right.

ePlant Interface (Left):

- Logo: BAR ePlant Arabidopsis thaliana
- Search bar: Enter a gene name (Example: ABI3 or AT5G60200)
- Buttons: Expression Angler, Mutant Phenotype Selector
- Navigation icons: Home, Info, Layers, etc.
- Status: No genes currently loaded
- Footer: About Help Comments

Data Visualization Tools (Left of Diagram):

- World eFP
- Plant eFP
- Tissue & Experiment eFP
- Cell eFP
- Chromosome viewer
- Interaction viewer
- Molecule viewer
- Sequence Browser

Data Visualization Tools for Multiple Levels of Plant Data (Right):

This diagram illustrates the flow of information from environmental factors to phenotypic responses. It shows the following components and their relationships:

- environment** (cloud and sun icon) influences **natural variation**, **DNA sequence (gene & promoter)**, and **ncRNA & conserved regions outside genes**.
- natural variation** leads to **DNA sequence (gene & promoter)**.
- DNA sequence (gene & promoter)** leads to **RNA transcript (sequence, abundance & alternative splicing)** via **transcription**.
- RNA transcript (sequence, abundance & alternative splicing)** leads to **protein sequence** via **translation**.
- protein sequence** leads to **3D structure** via **folding**.
- 3D structure** leads to **subcellular localization** and **spatio-temporal distribution / perturbation response**.
- 3D structure** leads to **metabolism**.
- metabolism** leads to **primary metabolism** and **secondary metabolism**.
- primary metabolism** leads to **phenotype / response**.
- secondary metabolism** leads to **phenotype / response**.
- environment** influences **signaling & signal transduction cascades**.
- RNA transcript (sequence, abundance & alternative splicing)** influences **signaling & signal transduction cascades**.
- protein sequence** influences **signaling & signal transduction cascades**.
- 3D structure** influences **signaling & signal transduction cascades**.
- signaling & signal transduction cascades** influences **protein networks**.
- signaling & signal transduction cascades** influences **methylation**.
- methylation** influences **DNA sequence (gene & promoter)**.
- protein networks** influences **phenotype / response**.

Legend:

- Influences
- leads to
- "central dogma" of molecular biology

5. Information databases in experimental biology

Protein data bank RCSB PDB

It contains both experimentally determined and computed 3D structures of proteins

RCSB PDB Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ About ▾ Careers COVID-19 Help Contact us

RCSB **PDB** PROTEIN DATA BANK 229,183 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

|Enter search term(s), Entry ID(s), or sequence| Include CSM ?

Advanced Search | Browse Annotations

RCSB PDB-101 wwPDB EMDataResource NAKB wwPDB Foundation PDB-Dev

Access Computed Structure Models (CSMs) of available model organisms Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

Explore NEW Features

PDB-101 Training Resources

December **Molecule of the Month**

Flagellar Motor