ACEMFS FUT Minna Bioinformatics Workshop

Introduction to Biological Databases and Resources

Itunoluwa Isewon PhD
Covenant University

Why Databases Matter?

What are Biological Databases?

- Vast, organized digital libraries storing biological data.
- Include DNA, RNA, protein sequences, gene expression, metabolic pathways, and more.

Why Biological Databases for Fungal/Mycotoxin Research?

- **Accelerates discovery:** No need to re-sequence common genes.
- **Provides context:** Compare your data with millions of existing entries.
- **Enables 'Omics' research:** Allows for genomics, transcriptomics, and proteomics studies.
- **Essential for pathway analysis:** Identify genes involved in mycotoxin biosynthesis.
- **Crucial for data integration:** Connect different layers of biological information.

GENES → ENZYMES → PATHWAYS → METABOLITES → OUTCOMES

Brief History & Types of Databases

- **Early era:** Nucleotide/protein archives (GenBank/ENA/DDBJ; UniProt)
- **Curation era:** InterPro, RefSeq, Swiss-Prot, PDB
- **Systems era:** KEGG, Reactome, BioCyc/MetaCyc
- **Domain-specialized:** FungiDB, MycoBank, MycoCosm; toxin/chemical knowledge bases
- **Types:**
    - **Primary** (raw submissions: GenBank, PRIDE, GEO/SRA)
    - **Secondary** (processed/curated: RefSeq, UniProtKB/Swiss-Prot, InterPro, Ensembl)
    - **Tertiary/Knowledge Bases** (integrated pathways: KEGG, Reactome, BioCyc)
    - **Taxonomic/Nomenclature**: MycoBank, Index Fungorum

FAIR & Reproducibility Essentials

- **FAIR**: Findable, Accessible, Interoperable, Reusable Rich metadata (collection site, substrate, growth conditions, toxin assay)
- Use standard IDs (NCBI TaxID, UniProt Accession, ChEBI ID, PubChem CID)
- Cite datasets (PRIDE/GEO accessions) and software versions

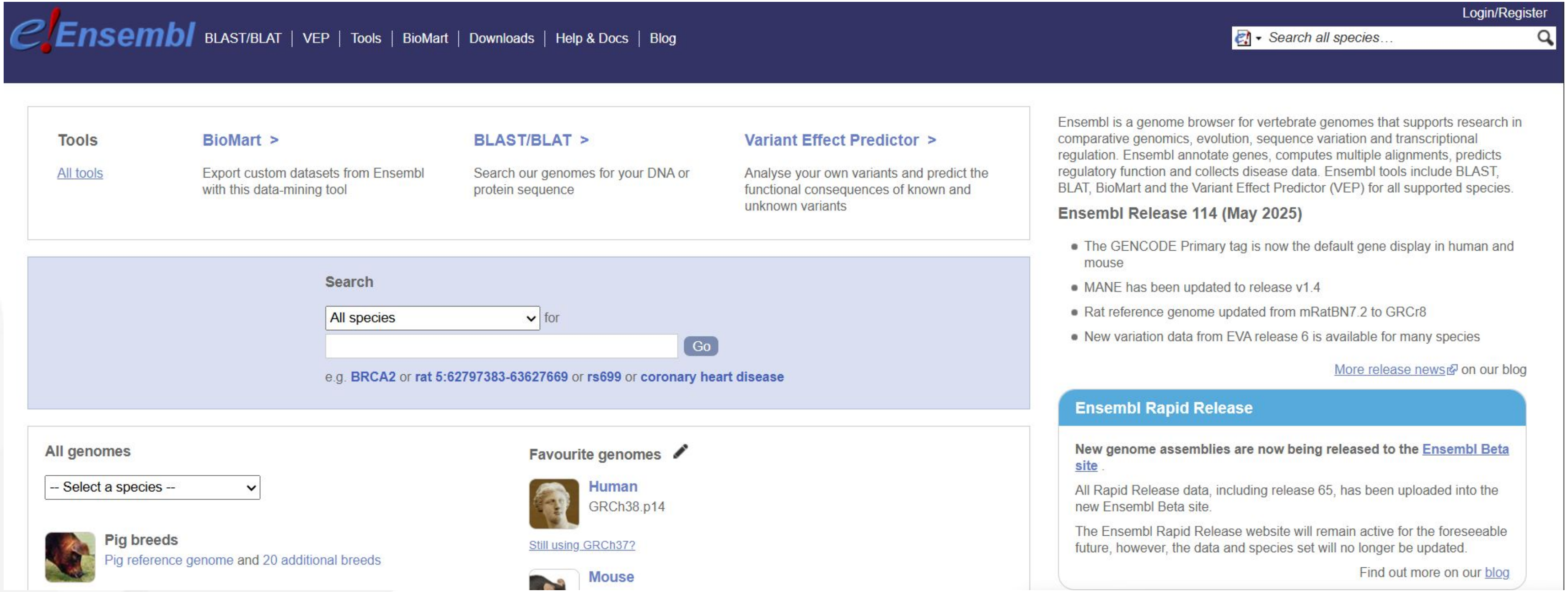
Core Database Categories

Genomics Databases

- **What they hold:** DNA and RNA sequences, genome assemblies, and annotations.
- **Key Examples:**
  - NCBI GenBank/RefSeq — primary nucleotide archive & curated references
  - ENA (EMBL-EBI) — European mirror; strong programmatic access
  - Ensembl Fungi — gene models, comparative genomics, orthology
  - JGI MycoCosm — fungal genome portal; many assemblies & annotations
  - FungiDB — integrated functional genomics for fungi & oomycetes

Genomics Databases

- **What to Retrieve & How**
  - Reference genomes, alternate assemblies, gene models (GFF3/GTF), CDS/proteins (FASTA), gene sequences, variants (VCF)
  - **Identifiers**: BioProject (PRJNA/PRJEB), BioSample (SAMN), Assembly (GCA/GCF), Gene (GeneID), Transcript/Protein (XM/XP/NP), TaxID
  - **Tools**: Web UI, FTP, APIs, EDirect/ENA Browser Tools


Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
Search all species…

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

All genomes

-- Select a species --

Pig breeds

Pig reference genome and 20 additional breeds

Favourite genomes

Human
GRCh38.p14

Still using GRCh37?

Mouse

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 114 (May 2025)

- The GENCODE Primary tag is now the default gene display in human and mouse
- MANE has been updated to release v1.4
- Rat reference genome updated from mRatBN7.2 to GRCr8
- New variation data from EVA release 6 is available for many species

More release news on our blog

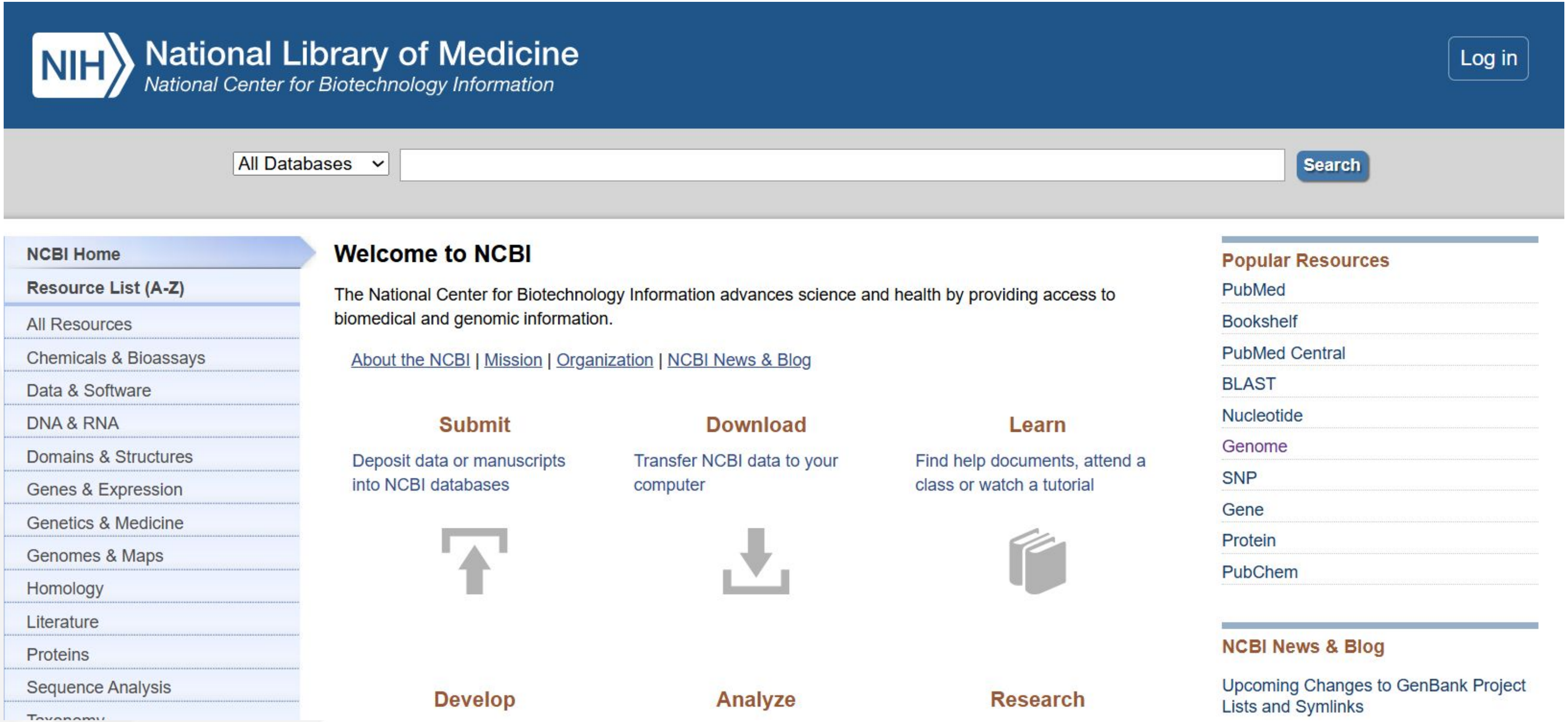
Ensembl Rapid Release

New genome assemblies are now being released to the Ensembl Beta site .

All Rapid Release data, including release 65, has been uploaded into the new Ensembl Beta site.

The Ensembl Rapid Release website will remain active for the foreseeable future, however, the data and species set will no longer be updated.

Find out more on our blog



National Library of Medicine

National Center for Biotechnology Information

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

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

Analyze

Research

Popular Resources

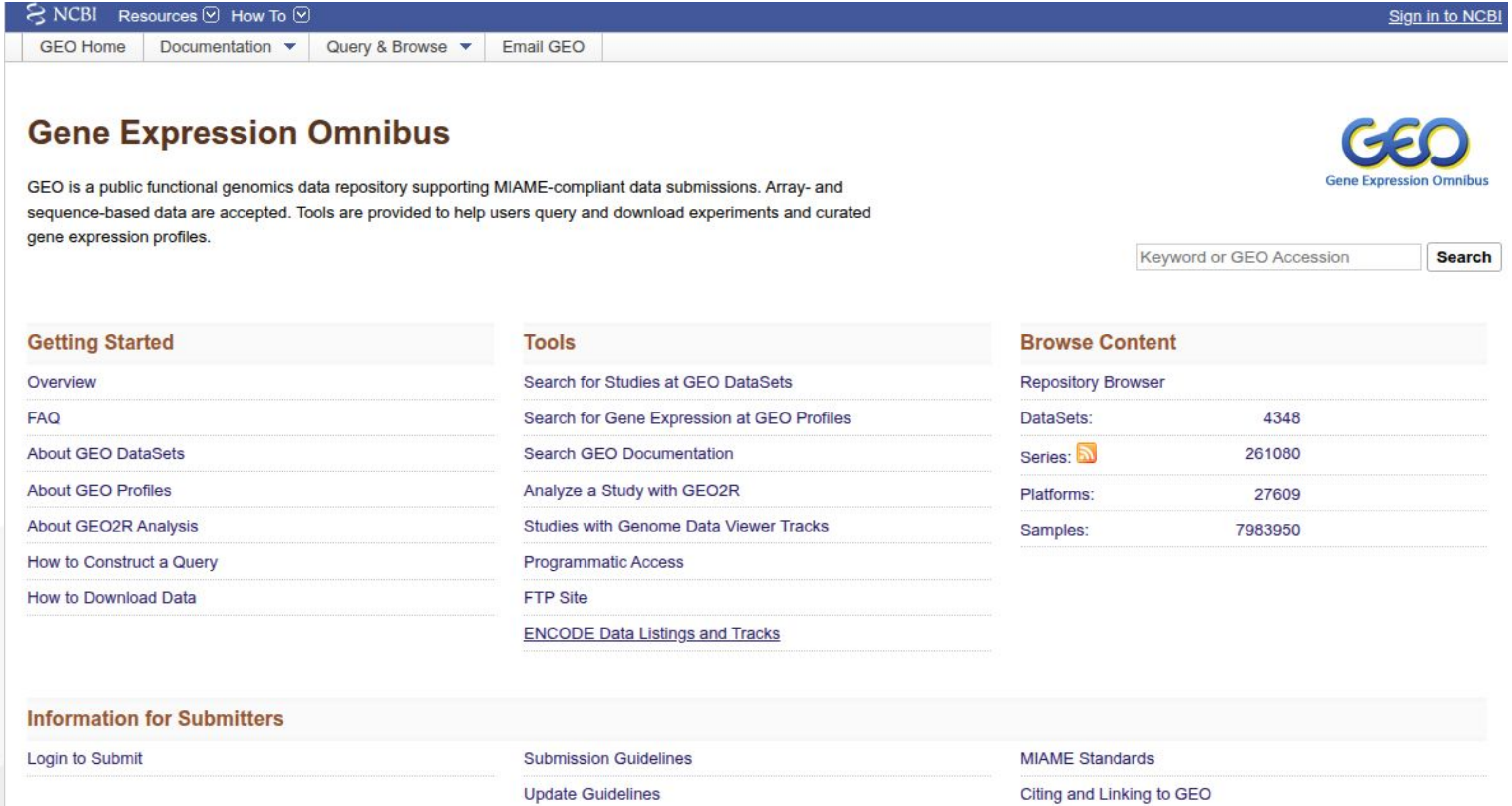
- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Upcoming Changes to GenBank Project Lists and Symlinks

Transcriptomics Databases

- **What they hold:** Gene expression data (e.g., RNA-Seq, microarray).
- **Key Examples:**
  - GEO (Gene Expression Omnibus) — curated studies and series
  - SRA — raw sequence reads (RNA-seq, amplicon, metatranscriptomes)
  - Expression Atlas (EMBL-EBI) — baseline & differential expression across conditions



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site
- ENCODE Data Listings and Tracks

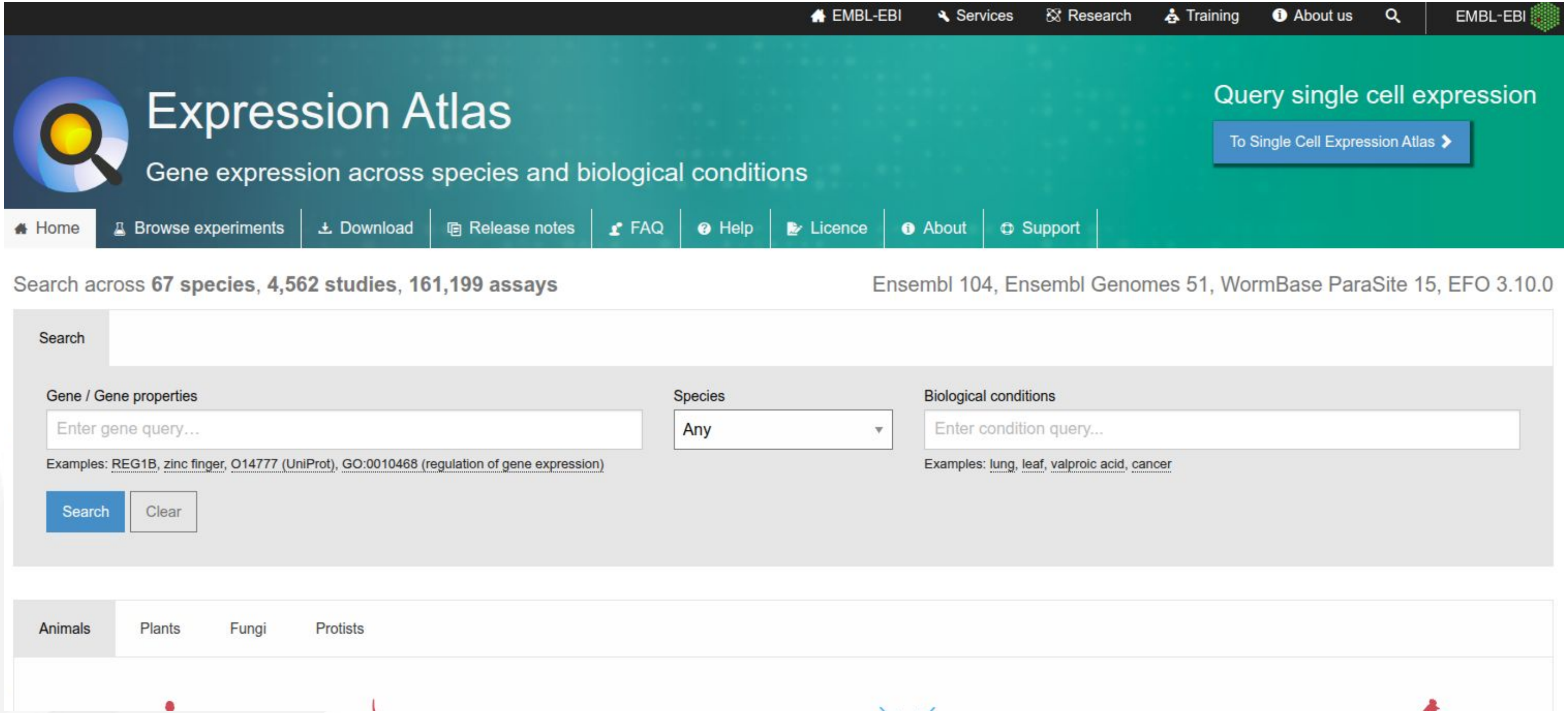
Browse Content

Repository Browser

DataSets:	4348
Series:	261080
Platforms:	27609
Samples:	7983950

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO



Expression Atlas

Gene expression across species and biological conditions

Query single cell expression

Search across **67 species**, **4,562 studies**, **161,199 assays**

Ensembl 104, Ensembl Genomes 51, WormBase ParaSite 15, EFO 3.10.0

Search

Gene / Gene properties: Enter gene query…

Examples: REG1B, zinc finger, O14777 (UniProt), GO:0010468 (regulation of gene expression)

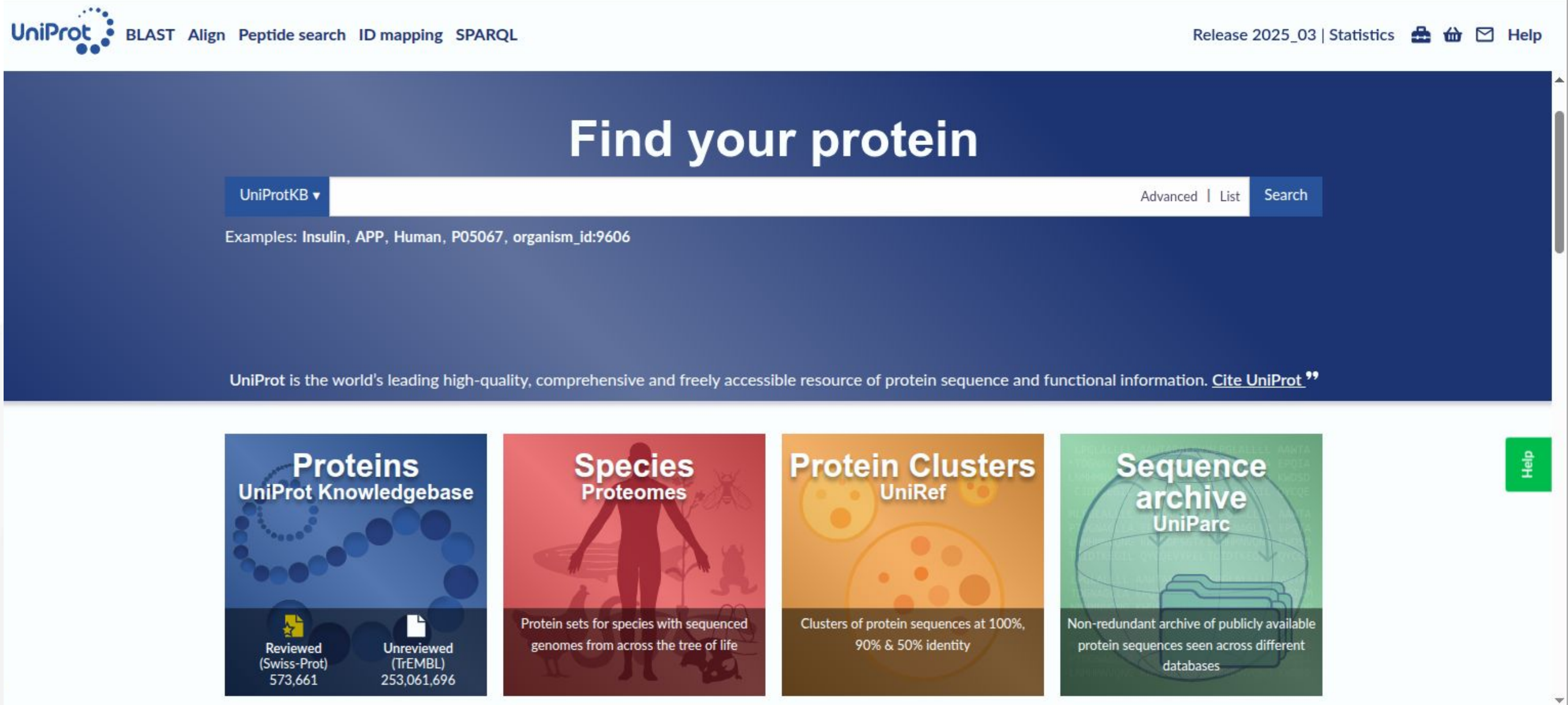
Species: Any

Biological conditions: Enter condition query...

Examples: lung, leaf, valproic acid, cancer

Proteomics Databases

- **What they hold:** Protein sequences, structures, functions, and post-translational modifications.
- **Key Examples:**
    - PDB (Protein Data Bank) — focuses on 3D structures of proteins and nucleic acids.
    - UniProtKB — gold standard for protein sequences & functional annotations; Swiss-Prot (reviewed) vs TrEMBL (unreviewed)
    - PRIDE (PRoteomics IDEntifications Database) — proteomics repository; search by species/tissue/keyword
    - PeptideAtlas (optional), MassIVE/GNPS (metabolomics/proteomics)



Find your protein

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt

Proteins
UniProt Knowledgebase

Reviewed (Swiss-Prot) 573,661

Unreviewed (TrEMBL) 253,061,696

Species
Proteomes

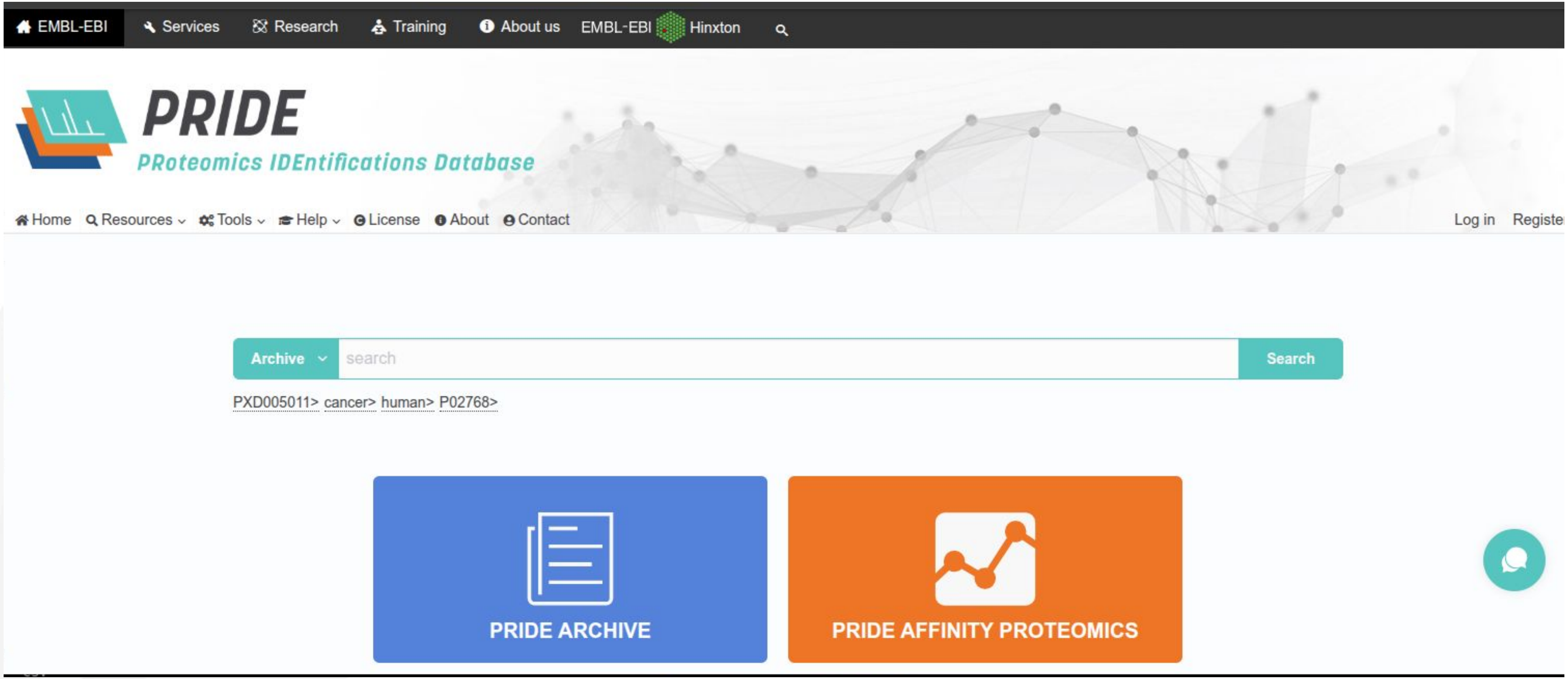
Protein sets for species with sequenced genomes from across the tree of life

Protein Clusters
UniRef

Clusters of protein sequences at 100%, 90% & 50% identity

Sequence archive
UniParc

Non-redundant archive of publicly available protein sequences seen across different databases



EMBL-EBI Services Research Training About us EMBL-EBI Hinxton

PRIDE

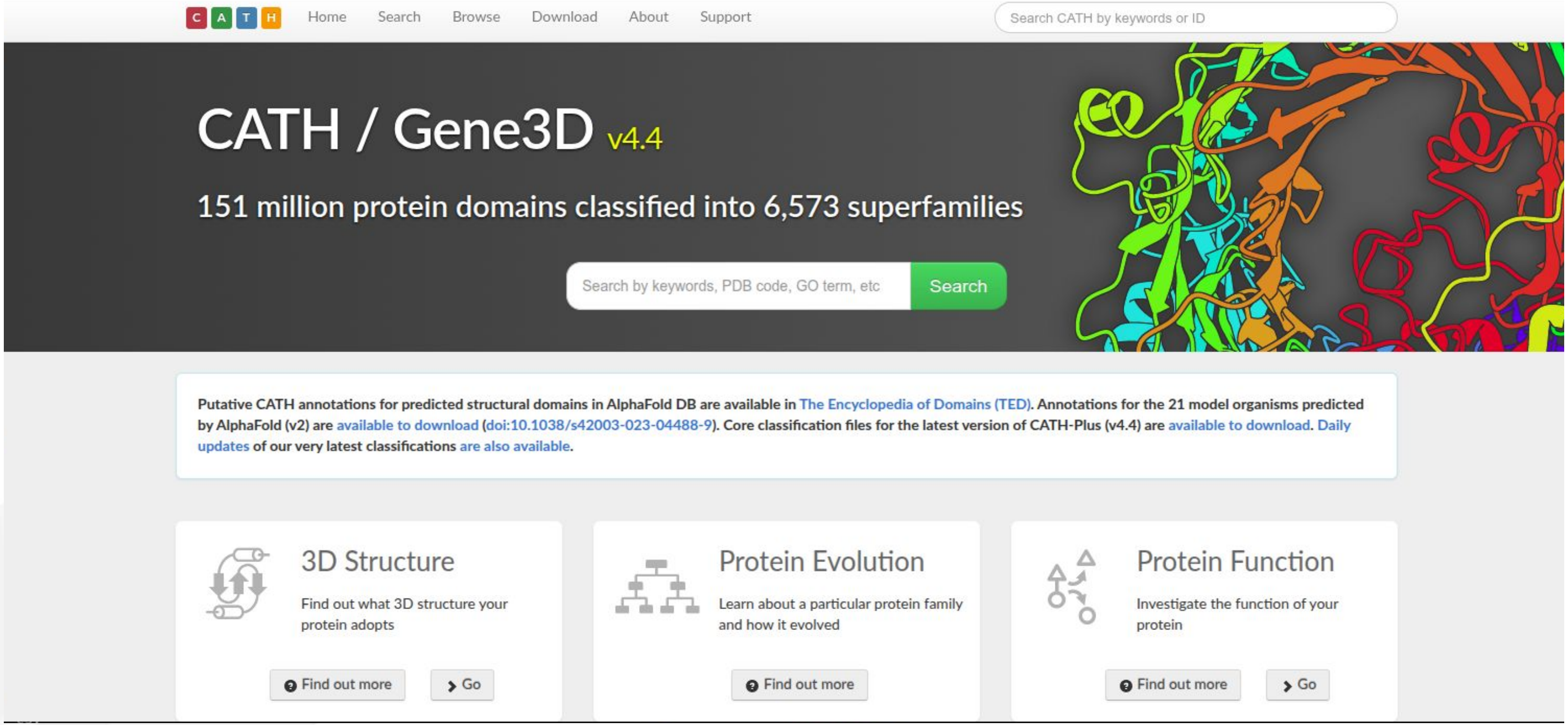
PRoteomics IDEntifications Database

Home Resources Tools Help License About Contact Log in Registe

PXD005011> cancer> human> P02768>

Structural Databases

- **What they hold:** Three-dimensional (3D) structures of macromolecules, primarily proteins and nucleic acids.
- **Key Examples:**
  - PDB (RCSB: Research Collaboratory for Structural Bioinformatics) — experimentally determined 3D structures
  - SCOP (Structural Classification of Proteins) — protein domains based on their structural and evolutionary relationships.
  - CATH (Class, Architecture, Topology, Homologous superfamily) — hierarchical classification of protein domain structures.
  - SWISS-MODEL Repository — homology models
  - AlphaFold DB — predicted structures with confidence scores (pLDDT)


C A T H
Home
Search
Browse
Download
About
Support
Search CATH by keywords or ID

CATH / Gene3D v4.4

151 million protein domains classified into 6,573 superfamilies

Search by keywords, PDB code, GO term, etc
Search

Putative CATH annotations for predicted structural domains in AlphaFold DB are available in The Encyclopedia of Domains (TED). Annotations for the 21 model organisms predicted by AlphaFold (v2) are available to download (doi:10.1038/s42003-023-04488-9). Core classification files for the latest version of CATH-Plus (v4.4) are available to download. Daily updates of our very latest classifications are also available.

3D Structure

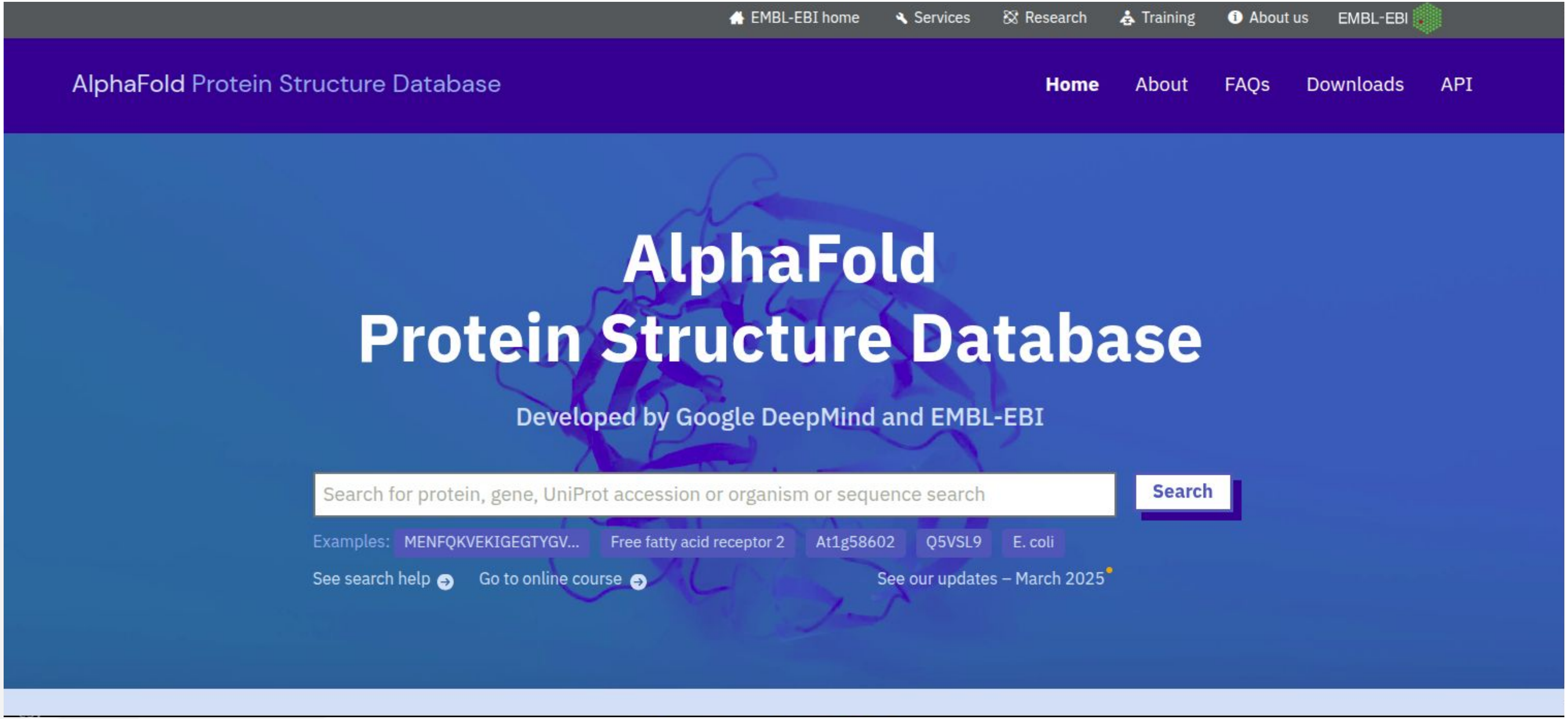
Find out what 3D structure your protein adopts

Protein Evolution

Learn about a particular protein family and how it evolved

Protein Function

Investigate the function of your protein



EMBL-EBI home · Services · Research · Training · About us · EMBL-EBI

AlphaFold Protein Structure Database

Home · About · FAQs · Downloads · API

AlphaFold Protein Structure Database

Developed by Google DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search

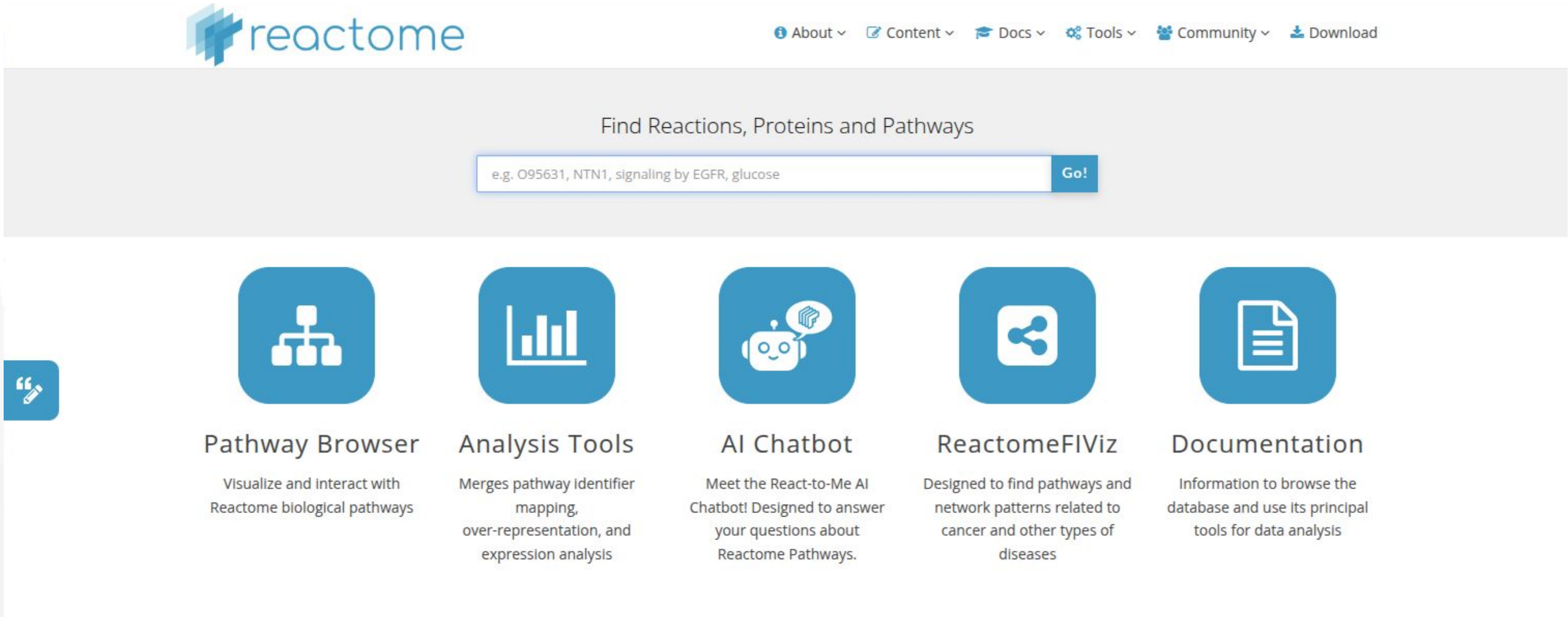
Search

Examples: MENFQKVEKIGEGTYGV... · Free fatty acid receptor 2 · At1g58602 · Q5VSL9 · E. coli

See search help · Go to online course · See our updates – March 2025

Metabolic & Pathway Databases

- **What they hold:** Biochemical pathways, metabolic reactions, and molecular interactions.
- **Key Examples:**
  - KEGG (Kyoto Encyclopedia of Genes and Genomes) — pathways, modules, orthologs; KEGG Mapper & REST API
  - Reactome — curated pathways; orthology projections to fungi (where available)
  - MetaCyc/BioCyc — curated metabolic pathways; species-specific Pathway/Genome Databases (PGDBs)
  - FungiDB — pathway & GO enrichment; links genes to pathways



Find Reactions, Proteins and Pathways

e.g. O95631, NTN1, signaling by EGFR, glucose

Pathway Browser

Visualize and interact with Reactome biological pathways

Analysis Tools

Merges pathway identifier mapping, over-representation, and expression analysis

AI Chatbot

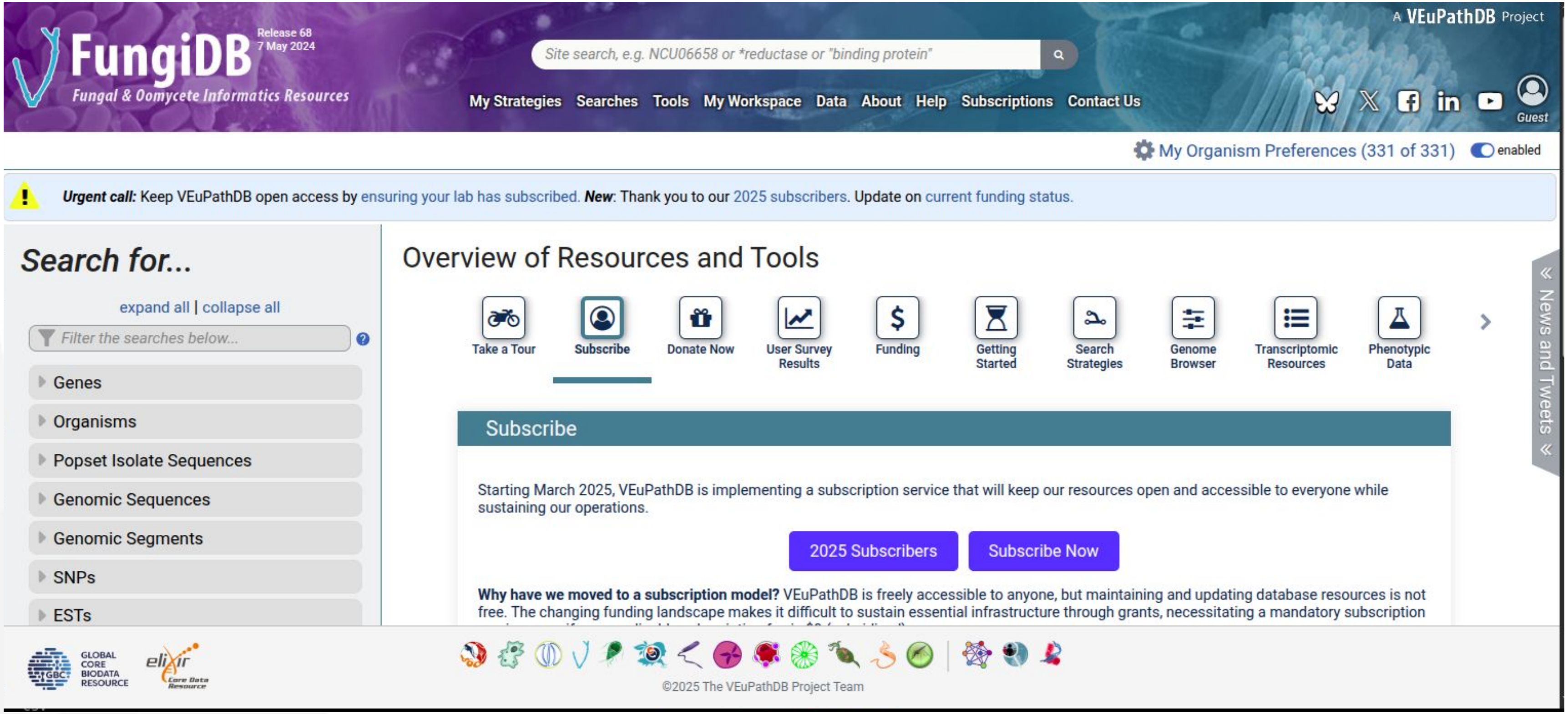
Meet the React-to-Me AI Chatbot! Designed to answer your questions about Reactome Pathways.

ReactomeFIViz

Designed to find pathways and network patterns related to cancer and other types of diseases

Documentation

Information to browse the database and use its principal tools for data analysis



Urgent call: Keep VEuPathDB open access by ensuring your lab has subscribed. ***New***: Thank you to our 2025 subscribers. Update on current funding status.

Overview of Resources and Tools

Subscribe

Starting March 2025, VEuPathDB is implementing a subscription service that will keep our resources open and accessible to everyone while sustaining our operations.

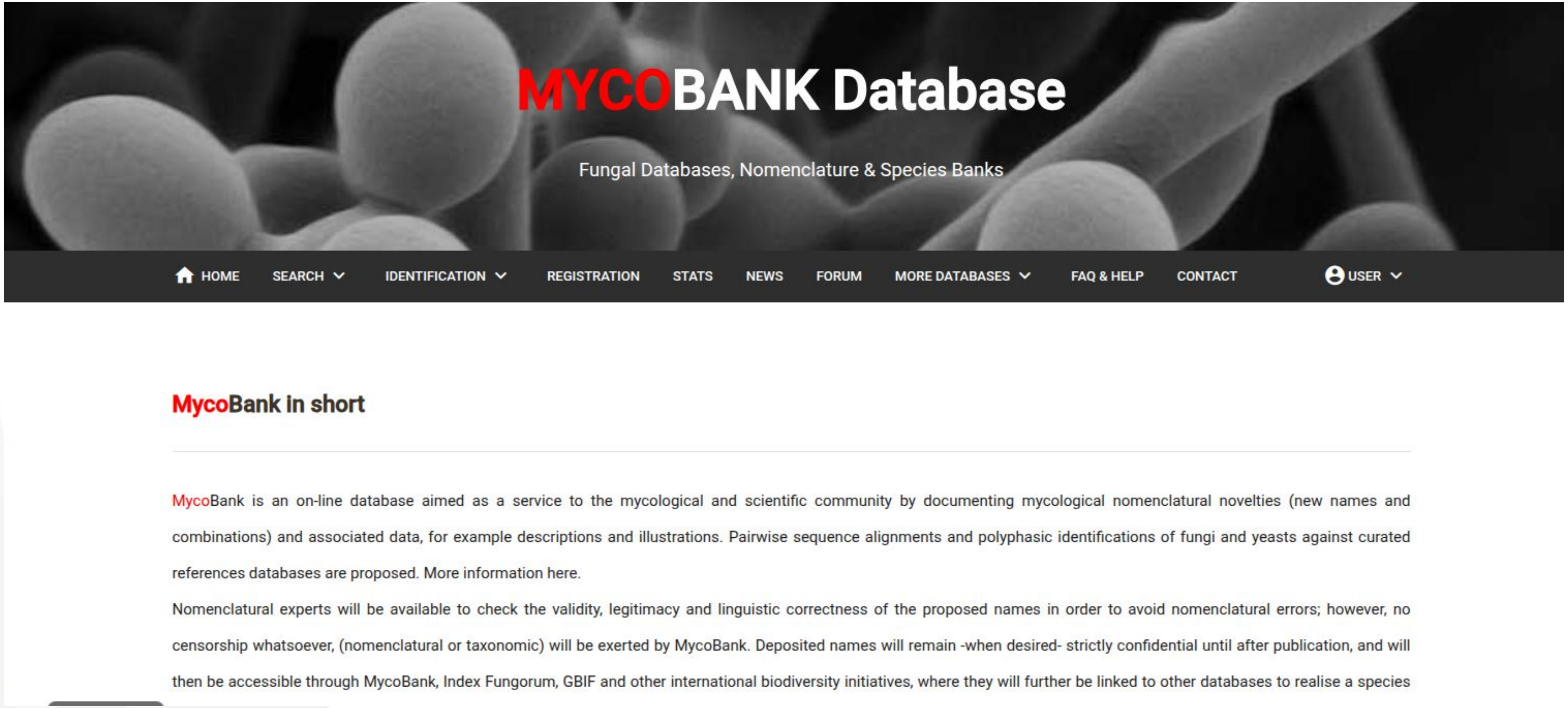
Why have we moved to a subscription model? VEuPathDB is freely accessible to anyone, but maintaining and updating database resources is not free. The changing funding landscape makes it difficult to sustain essential infrastructure through grants, necessitating a mandatory subscription

Specialized Fungal & Toxin Resources

- **What they hold:** Highly curated data specific to fungi and their toxins.
- **Key Examples:**
  - MycoBank — taxonomy/nomenclature, type material, literature links
  - Index Fungorum — complementary taxonomic resource
  - FungiDB — genomes, gene pages, tools (orthologs, pathways, expression)
  - MycoCosm — many fungal genomes; community annotation
  - MIBiG — curated biosynthetic gene clusters (BGCs) with metabolites

Specialized Fungal & Toxin Resources

- **Key Examples:**
  - **antiSMASH DB** — predicted BGCs across genomes
  - **ChEBI (Chemical Entities of Biological Interest)** — chemical ontology with mycotoxins (e.g., aflatoxin B1, ochratoxin A)
  - **PubChem** — compound records, bioactivity, vendor info
  - **T3DB** — toxin knowledge bases (includes mycotoxins)
  - **EPA CompTox Dashboard** — toxicity data and identifiers
  - **MetaboLights** — metabolomics studies (LC-MS/GC-MS) including mycotoxins



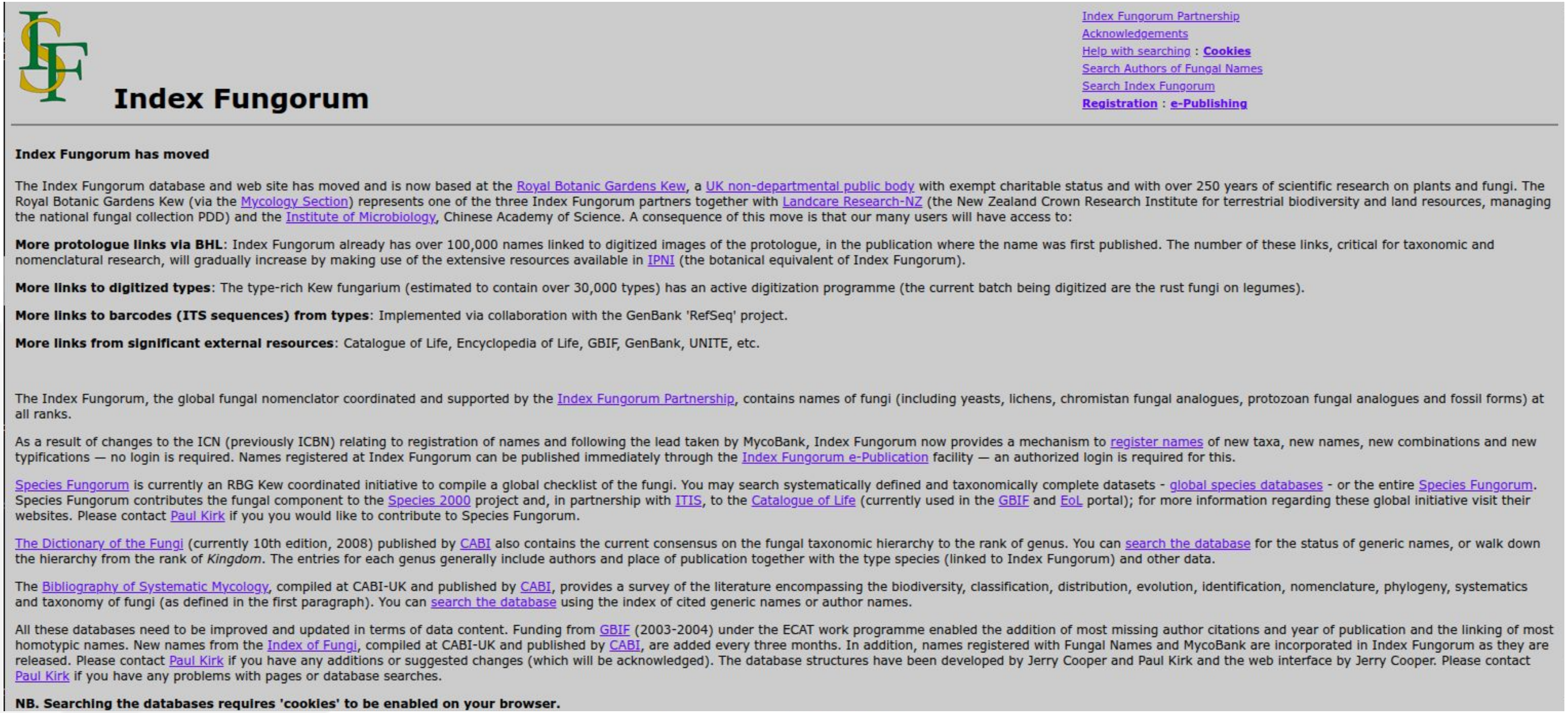
MYCOBANK Database

Fungal Databases, Nomenclature & Species Banks

MycoBank in short

MycoBank is an on-line database aimed as a service to the mycological and scientific community by documenting mycological nomenclatural novelties (new names and combinations) and associated data, for example descriptions and illustrations. Pairwise sequence alignments and polyphasic identifications of fungi and yeasts against curated references databases are proposed. More information here.

Nomenclatural experts will be available to check the validity, legitimacy and linguistic correctness of the proposed names in order to avoid nomenclatural errors; however, no censorship whatsoever, (nomenclatural or taxonomic) will be exerted by MycoBank. Deposited names will remain -when desired- strictly confidential until after publication, and will then be accessible through MycoBank, Index Fungorum, GBIF and other international biodiversity initiatives, where they will further be linked to other databases to realise a species



Index Fungorum

Index Fungorum Partnership
Acknowledgements
Help with searching : **Cookies**
Search Authors of Fungal Names
Search Index Fungorum
Registration : **e-Publishing**

Index Fungorum has moved

The Index Fungorum database and web site has moved and is now based at the Royal Botanic Gardens Kew, a UK non-departmental public body with exempt charitable status and with over 250 years of scientific research on plants and fungi. The Royal Botanic Gardens Kew (via the Mycology Section) represents one of the three Index Fungorum partners together with Landcare Research-NZ (the New Zealand Crown Research Institute for terrestrial biodiversity and land resources, managing the national fungal collection PDD) and the Institute of Microbiology, Chinese Academy of Science. A consequence of this move is that our many users will have access to:

More protologue links via BHL: Index Fungorum already has over 100,000 names linked to digitized images of the protologue, in the publication where the name was first published. The number of these links, critical for taxonomic and nomenclatural research, will gradually increase by making use of the extensive resources available in IPNI (the botanical equivalent of Index Fungorum).

More links to digitized types: The type-rich Kew fungarium (estimated to contain over 30,000 types) has an active digitization programme (the current batch being digitized are the rust fungi on legumes).

More links to barcodes (ITS sequences) from types: Implemented via collaboration with the GenBank 'RefSeq' project.

More links from significant external resources: Catalogue of Life, Encyclopedia of Life, GBIF, GenBank, UNITE, etc.

The Index Fungorum, the global fungal nomenclator coordinated and supported by the Index Fungorum Partnership, contains names of fungi (including yeasts, lichens, chromistan fungal analogues, protozoan fungal analogues and fossil forms) at all ranks.

As a result of changes to the ICN (previously ICBN) relating to registration of names and following the lead taken by MycoBank, Index Fungorum now provides a mechanism to register names of new taxa, new names, new combinations and new typifications — no login is required. Names registered at Index Fungorum can be published immediately through the Index Fungorum e-Publication facility — an authorized login is required for this.

Species Fungorum is currently an RBG Kew coordinated initiative to compile a global checklist of the fungi. You may search systematically defined and taxonomically complete datasets - global species databases - or the entire Species Fungorum. Species Fungorum contributes the fungal component to the Species 2000 project and, in partnership with ITIS, to the Catalogue of Life (currently used in the GBIF and EoL portal); for more information regarding these global initiative visit their websites. Please contact Paul Kirk if you you would like to contribute to Species Fungorum.

The Dictionary of the Fungi (currently 10th edition, 2008) published by CABI also contains the current consensus on the fungal taxonomic hierarchy to the rank of genus. You can search the database for the status of generic names, or walk down the hierarchy from the rank of *Kingdom*. The entries for each genus generally include authors and place of publication together with the type species (linked to Index Fungorum) and other data.

The Bibliography of Systematic Mycology, compiled at CABI-UK and published by CABI, provides a survey of the literature encompassing the biodiversity, classification, distribution, evolution, identification, nomenclature, phylogeny, systematics and taxonomy of fungi (as defined in the first paragraph). You can search the database using the index of cited generic names or author names.

All these databases need to be improved and updated in terms of data content. Funding from GBIF (2003-2004) under the ECAT work programme enabled the addition of most missing author citations and year of publication and the linking of most homotypic names. New names from the Index of Fungi, compiled at CABI-UK and published by CABI, are added every three months. In addition, names registered with Fungal Names and MycoBank are incorporated in Index Fungorum as they are released. Please contact Paul Kirk if you have any additions or suggested changes (which will be acknowledged). The database structures have been developed by Jerry Cooper and Paul Kirk and the web interface by Jerry Cooper. Please contact Paul Kirk if you have any problems with pages or database searches.

NB. Searching the databases requires 'cookies' to be enabled on your browser.

Workflow Platforms & Tools

- **What They Hold:** Workflow platforms contain a collection of interconnected tools, pipelines, and scripts that automate complex data analysis.
- **Why use them:** To connect different database queries and perform complex analyses without extensive coding.
- **Key Examples**
    - **Galaxy:** An accessible web-based platform for computational biology. It offers a user-friendly interface to build and run complex bioinformatics workflows.
    - **EMBL-EBI Tools:** A comprehensive suite of online tools for biological data analysis. They are designed to be interconnected, allowing for seamless data transfer and analysis across databases.
    - **GenePattern:** Provides access to hundreds of tools for genomics analysis. It is particularly strong in creating and sharing reproducible research pipelines.

Workflow Platforms & Tools

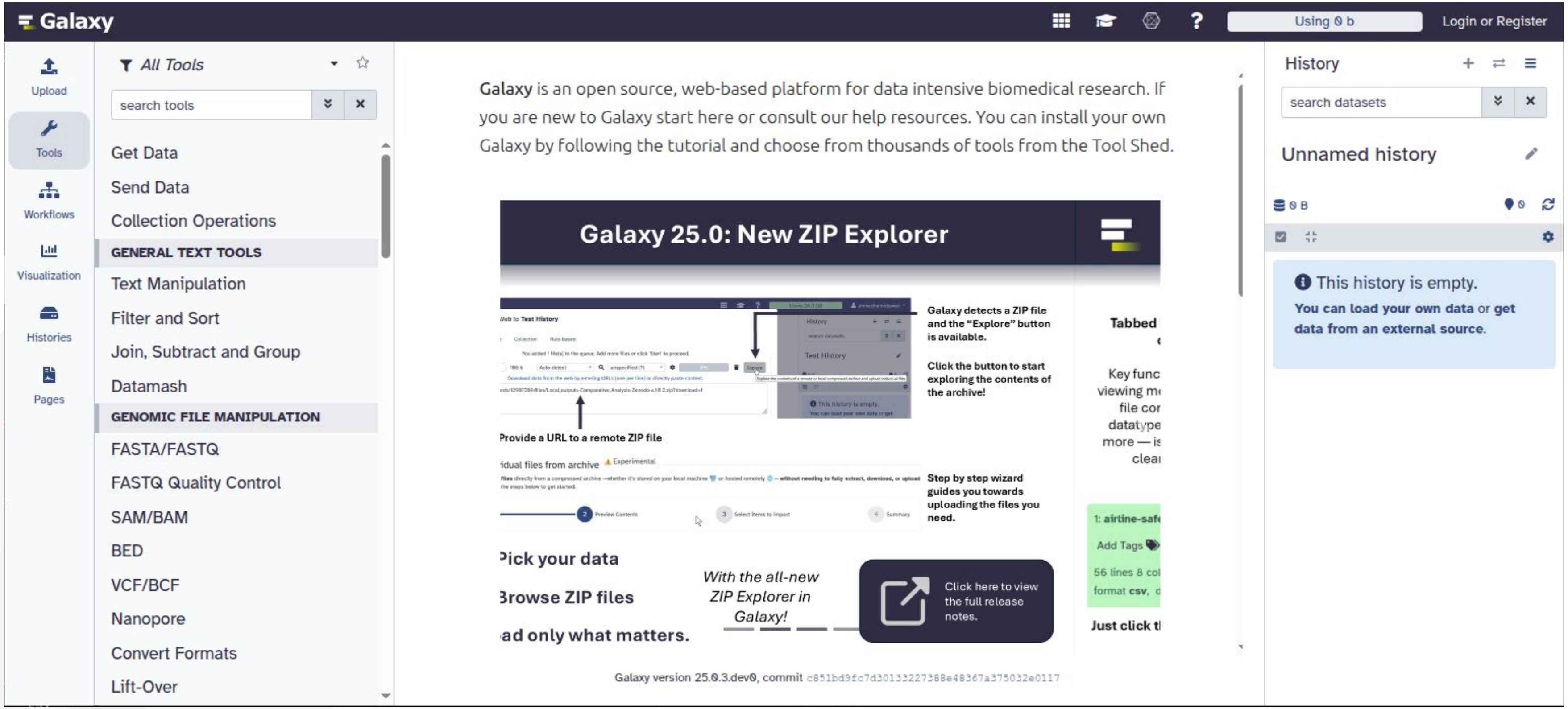
- **Key Examples**
  - **Cytoscape**: A powerful platform for visualizing and analyzing biological networks. It is essential for making sense of complex interactions from genomics and proteomics data.
  - **BioPython**: A Python package for biological computation. It provides tools for manipulating and analyzing biological data, making it a valuable resource for researchers who can program.
  - **KBase**: A comprehensive platform that integrates data, tools, and analysis workflows for systems biology research. It aims to accelerate scientific discoveries by providing advanced computational resources.

Workflow Platforms & Tools

- **Key Examples**
  - **Geneious**: A powerful bioinformatics software platform. It offers a user-friendly interface with tools for sequence analysis, molecular cloning, and phylogenetics.
  - **Conda**/**Bioconda**: A package and environment management system. It simplifies software installation and dependency management for scientific computing.
  - **Bioinformatics Toolkit**: A collection of web-based tools for various bioinformatics analyses. It offers a suite of tools for sequence analysis, structure prediction, and more.
  - **Bioconductor**: An open-source software project that provides tools for the analysis of high-throughput genomic data. It offers a comprehensive collection of R packages for bioinformatics.

Workflow Platforms & Tools

- **Key Examples**
  - **Taverna**: A workflow management system for scientific data analysis. It allows researchers to create, execute, and share complex workflows.
  - **Snakemake**: A workflow management system and Python-based scripting language for creating reproducible data analysis pipelines. It simplifies defining rules and dependencies.
  - **Docker**: A platform for running applications in containers. It is used to package analysis tools and dependencies, ensuring consistent execution across different environments.
  - **Nextflow**: A data-driven workflow management system. It simplifies creating and executing scalable and reproducible computational pipelines, with support for container technologies like Docker.



This history is empty.

You can load your own data or get data from an external source.

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Galaxy 25.0: New ZIP Explorer

Galaxy detects a ZIP file and the "Explore" button is available.

Click the button to start exploring the contents of the archive!

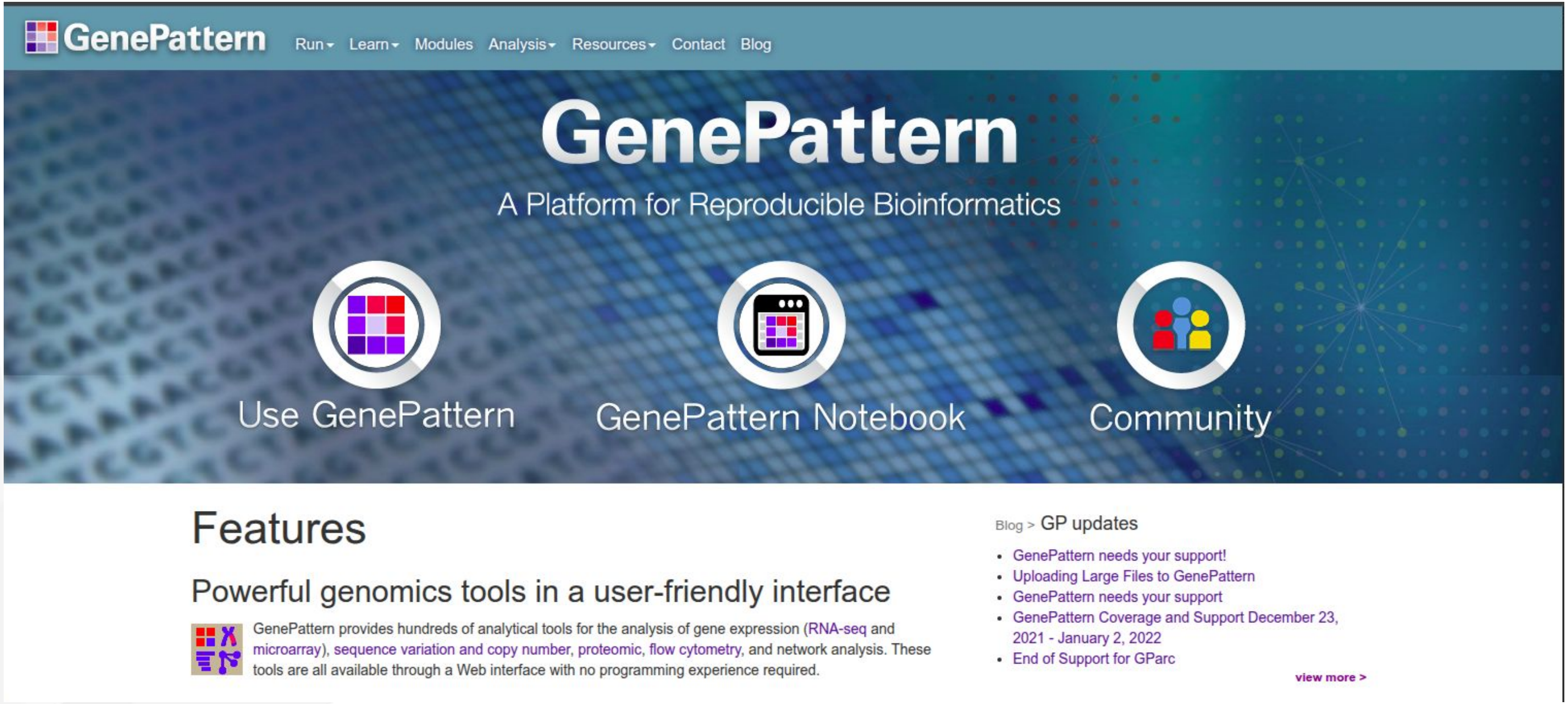
Provide a URL to a remote ZIP file

Step by step wizard guides you towards uploading the files you need.

With the all-new ZIP Explorer in Galaxy!

Click here to view the full release notes.

Galaxy version 25.0.3.dev0, commit c851bd9fc7d30133227388e48367a375032e0117



GenePattern

A Platform for Reproducible Bioinformatics

Features

Powerful genomics tools in a user-friendly interface

GenePattern provides hundreds of analytical tools for the analysis of gene expression (RNA-seq and microarray), sequence variation and copy number, proteomic, flow cytometry, and network analysis. These tools are all available through a Web interface with no programming experience required.

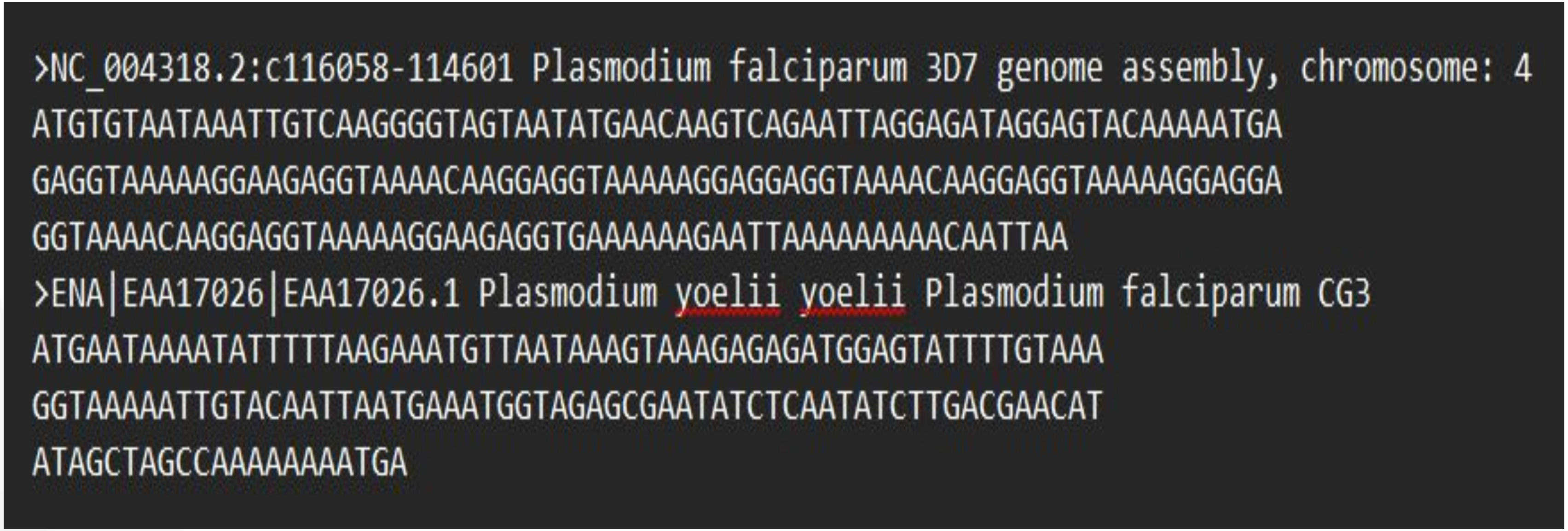
Blog > GP updates

- GenePattern needs your support!
- Uploading Large Files to GenePattern
- GenePattern needs your support
- GenePattern Coverage and Support December 23, 2021 - January 2, 2022
- End of Support for GParc

view more >

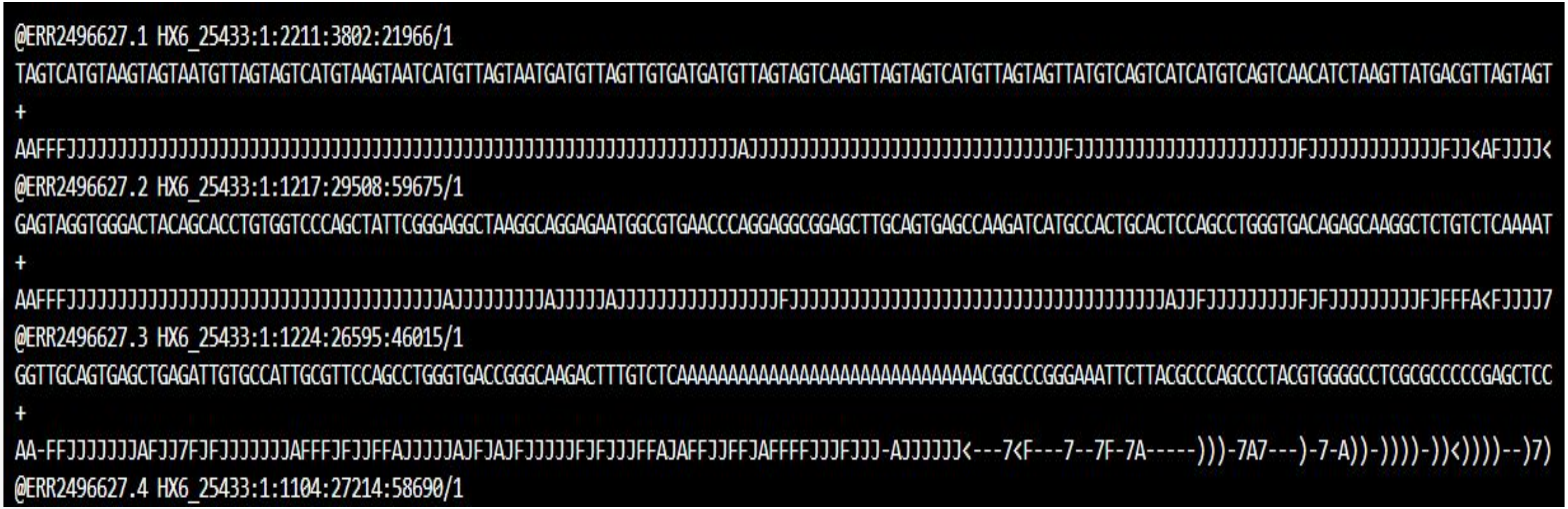
Sequence Data Formats

- **FASTA (.fa, .fasta):** Used for storing DNA, RNA, or protein sequences. It has two parts:
  - A header line starting with > that contains the sequence name and information.
  - The sequence itself on the following lines.
- **FASTQ (.fq, .fastq):** Used for raw sequencing reads. It stores both the sequence and a quality score for each base. It has four lines per sequence:
  - A header line starting with @.
  - The raw sequence.
  - A separator line, usually just a +.
  - A string of quality characters, one for each base in the sequence.



```
>NC_004318.2:c116058-114601 Plasmodium falciparum 3D7 genome assembly, chromosome: 4
ATGTGTAATAAATTGTCAAGGGGTAGTAATATGAACAAGTCAGAATTAGGAGATAGGAGTACAAAAATGA
GAGGTAAAAAGGAAGAGGTAAAACAAGGAGGTAAAAAGGAGGAGGTAAAACAAGGAGGTAAAAAGGAGGA
GGTAAAACAAGGAGGTAAAAAGGAAGAGGTGAAAAAAGAATTAAAAAAAAAACAATTAA
>ENA|EAA17026|EAA17026.1 Plasmodium yoelii yoelii Plasmodium falciparum CG3
ATGAATAAAATATTTTTAAGAAATGTTAATAAAGTAAAGAGAGATGGAGTATTTTGTAAA
GGTAAAAATTGTACAATTAATGAAATGGTAGAGCGAATATCTCAATATCTTGACGAACAT
ATAGCTAGCCAAAAAAAAATGA
```

FASTA Format



```
@ERR2496627.1 HX6_25433:1:2211:3802:21966/1
TAGTCATGTAAGTAGTAATGTTAGTAGTCATGTAAGTAATCATGTTAGTAATGATGTTAGTTGTGATGATGTTAGTAGTCAAGTTAGTAGTCATGTTAGTAGTTATGTCAGTCATCATGTCAGTCAACATCTAAGTTATGACGTTAGTAGT
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJAJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJFJJ<AFJJJJ<
@ERR2496627.2 HX6_25433:1:1217:29508:59675/1
GAGTAGGTGGGACTACAGCACCTGTGGTCCCAGCTATTCGGGAGGCTAAGGCAGGAGAATGGCGTGAACCCAGGAGGCGGAGCTTGCAGTGAGCCAAGATCATGCCACTGCACTCCAGCCTGGGTGACAGAGCAAGGCTCTGTCTCAAAAT
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJAJJJJJJJJJAJJJJJAJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJAJJFJJJJJJJJJFJFJJJJJJJJJFJFFFA<FJJJJ7
@ERR2496627.3 HX6_25433:1:1224:26595:46015/1
GGTTGCAGTGAGCTGAGATTGTGCCATTGCGTTCCAGCCTGGGTGACCGGGCAAGACTTTGTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACGGCCCGGGAAATTCTTACGCCCAGCCCTACGTGGGGCCTCGCGCCCCCGAGCTCC
+
AA-FFJJJJJJJAFJJ7FJFJJJJJJJAFFFJFJJFFAJJJJJAJFJAJFJJJJJFJFJJJFFAJAFFJJFFJAFFFFJJJFJJJ-AJJJJJJ<---7<F---7--7F-7A-----)))-7A7---)-7-A))-))))-))<))))--)7)
@ERR2496627.4 HX6_25433:1:1104:27214:58690/1
```

FASTQ Format

Common File Formats in Bioinformatics

Format (Extension)	Description	Structure	Primary Use	Example / Key Tools
FASTA (.fasta, .fa, .fna)	Text-based format for nucleotide or protein sequences.	A > header line followed by lines of sequence data.	Storing genome, transcriptome, and protein sequences.	>gene1 Homo sapiens ATGCGTAAGT...
FASTQ (.fastq, .fq)	Stores sequence data and corresponding quality scores.	Four lines per entry: identifier (@), sequence, separator (+), and quality scores.	Storing raw Next-Generation Sequencing (NGS) reads.	@SEQ_ID GATTT... +!"*((...
SAM/BAM (.sam, .bam)	Sequence Alignment Map (text) and its binary equivalent (BAM).	A header section followed by alignment lines, each with 11+ tab-delimited fields.	Storing alignment data of reads against a reference genome.	Tools: BWA, Bowtie, STAR

Common File Formats in Bioinformatics

Format (Extension)	Description	Structure	Primary Use	Example / Key Tools
VCF (.vcf)	Variant Call Format for storing gene sequence variations.	Metadata lines (##), a header line (#), and data lines for each variant.	Storing SNPs, INDELs, and other genetic variations.	#CHROM POS ID REF ALT...
GFF/GTF (.gff, .gff3, .gtf)	General Feature/Transfer Format for genome annotations.	Nine tab-separated columns describing genomic features (e.g., gene, exon).	Storing gene models and transcript annotations.	Tools: Genome browsers, RNA-seq analysis
BED (.bed)	Browser Extensible Data format for defining genomic regions.	A simple, tab-delimited format with at least three columns: chromosome, start, and end.	Representing genomic intervals like peaks or annotations.	Tools: UCSC Genome Browser, BEDTools

Common File Formats in Bioinformatics

Format (Extension)	Description	Structure	Primary Use	Example / Key Tools
PDB (.pdb)	Protein Data Bank format for 3D macromolecule structures.	Contains atomic coordinates, secondary structure information, and other structural data.	Protein modeling, molecular docking, and structural biology.	Tools: PyMOL, Chimera, RCSB
CLUSTAL/ALN (.aln)	A common text format for multiple sequence alignments.	Header followed by aligned sequences, often with sequence names and alignment blocks.	Visualizing and analyzing multiple sequence alignments.	Tools: ClustalW, MUSCLE, MAFFT

Practical Demo

- **Demo 1 (NCBI GenBank):**
  - **URL:** https://www.ncbi.nlm.nih.gov/genbank/
  - **Search Term:** aflD[gene] AND Aspergillus flavus[organism]
- **Demo 2 (KEGG):**
  - **URL:** https://www.genome.jp/kegg/pathway.html
  - **Search Term:** Aflatoxin biosynthesis or map00254
- **Demo 3 (PDB):**
  - **URL:** https://www.rcsb.org/
  - **Search Term:** Aflatoxin biosynthesis or a specific enzyme name related to mycotoxins.

Exercises

Hands-On

Scenario: You've just performed an RNA-Seq experiment on *Aspergillus fumigatus* grown in two conditions: one with a high-sugar medium (A) and one with a low-sugar medium (B). You find that the gene with the NCBI accession number **XM_001481541** is highly up-regulated in condition B.

Tasks:

1. **Access:** Go to NCBI, find the record for **XM_001481541**.
2. **Analyze:** What is the gene's name and its type?
3. **Integrate:** Use this information to hypothesize why this gene might be more active when the fungus is grown on a low-sugar medium.
4. **Connect:** Is there a known pathway this gene is part of? Use a database like UniProt or FungiDB to check.

Exercises

Exercise 1: Retrieve & Annotate a Toxin Gene.

- **Task:** Find the gene sequence for a key enzyme in the patulin biosynthetic pathway in *Penicillium expansum* using NCBI.
- **Deliverable:** A FASTA file of the sequence and a summary of its function.

Exercise 2: Map a Mycotoxin Biosynthetic Pathway.

- **Task:** Using KEGG, find and visualize the entire pathway for ochratoxin A biosynthesis.
- **Deliverable:** A screenshot of the pathway with key enzymes highlighted.

Exercise 3: Analyze Expression Data.

- **Task:** Find an RNA-Seq dataset in GEO related to a mycotoxin-producing fungus. Identify a gene that is significantly upregulated when the fungus is grown on a particular substrate.
- **Deliverable:** The accession number and a brief description of the gene and its expression pattern.