Biological Databases Overview

Prabal Ghosh[1][0009−0004−3449−5811]

Universite Cote d'Azur, Sophia Antipolis, France
prabal5ghosh@gmail.com

1 TCGA Database

Summary: The Cancer Genome Atlas (TCGA) has molecularly characterized over 20,000 primary cancer and matched normal samples across 33 cancer types. It has generated more than 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data, contributing to advancements in cancer diagnosis, treatment, and prevention.

Version: 42.0 (30 January 2025)

Types of Data: Genomics (SNV, CNV), transcriptomics (transcriptome profiling), epigenomics (DNA methylation), biospecimen, clinical, proteomic, somatic structural variation, and structural variation.

Data Statistics:

- 86 Projects
- 69 Primary Sites
- 44,736 Cases
- 1,121,816 Files
- 22,534 Genes
- 2,940,240 Mutations

General Information: Users can filter projects based on primary sites, disease types, and other criteria. After selecting a project, a summary is displayed, including case numbers and available data types. Projects can be saved as cohorts, and multiple datasets can be added to a cart for further analysis.

2 GTEx Database

Summary: The Genotype-Tissue Expression (GTEx) Portal is a publicly available resource designed for studying gene expression and regulation across different tissues, individuals, developmental stages, and species.

Version: V10

Types of Data: Gene expression and Quantitative Trait Loci (QTL).

Data Statistics:

- 54 Tissues
- 946 Donors
- 19,788 Samples

Donor Information: Includes sex, race, age, cause of death, and tissue counts per donor.

General Information: The number of samples per tissue can be sorted and viewed for analysis.

3 GEO Database

Summary: The Gene Expression Omnibus (GEO) is an international public repository that archives and freely shares microarray, next-generation sequencing, and other high-throughput functional genomics data submitted by researchers.

General Information: GEO is often used alongside PubMed. Researchers can locate relevant papers and search for GEO accession numbers to access dataset details. From a GEO entry, users can also retrieve SRA database accessions and search for related datasets using SRA Explorer.

4 ProteomeExchange Database

Summary: ProteomeXchange is a consortium for sharing mass spectrometry-based proteomics data. It provides a centralized platform for submitting, accessing, and distributing proteomics datasets from various research groups and laboratories.

Types of Data: Proteomics data from multiple species, including humans and rats.

General Information: Users can explore interactive plots displaying dataset distributions across species and spectrometer types. The database also indicates which hosting repository contains a selected dataset, allowing users to access detailed study information from the corresponding website.

5 Metabolomics Workbench

Summary: The Metabolomics Workbench Metabolite Database provides structural and annotation data for biologically relevant metabolites.

Types of Data: Metabolomics data from studies on cells, tissues, and organisms, including both small- and large-scale experiments.

Data Statistics:

- 3,623 Studies
- Over 167,000 Entries

General Information: Users can select studies and metabolites, visualize data through various plots, and conduct statistical analyses.