



# Biomedical Databases

GSND 5340Q, BMDA

W. Evan Johnson, Ph.D.

Professor, Division of Infectious Disease

Director, Center for Data Science

Rutgers University – New Jersey Medical School

2025-05-07

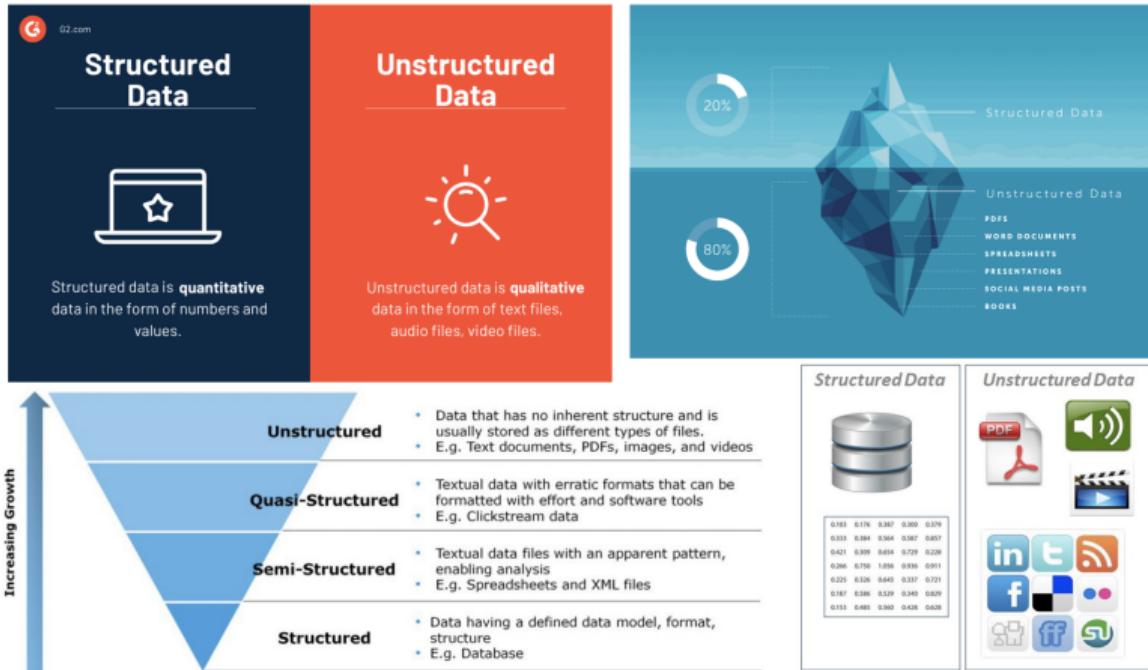
## Biological Databases and resources

# BIG DATA



Big Data has fundamentally changed how we look at science and business. Along with advances in analytic methods, they are providing unparalleled insights into our physical world and society

# Structured vs. Unstructured data



# Biology Is A Data Science

- ▶ Hundreds of thousand of species
- ▶ Million of articles in scientific literature
- ▶ Genetic Information
  - ▶ Gene names
  - ▶ Phenotype of mutants
  - ▶ Location of genes/mutations on chromosomes
  - ▶ Linkage (relationships between genes)

# Data and Metadata

- ▶ Data are “concrete” objects
  - ▶ e.g. number, tweet, nucleotide sequence
- ▶ Metadata describes properties of data
  - ▶ e.g. object is a number, each tweet has an author
- ▶ Database structure may contain metadata
  - ▶ Type of object (integer, float, string, etc)
  - ▶ Size of object (strings at most 4 characters long)
  - ▶ Relationships between data (chromosomes have zero or more genes)

# What is a Database?

- ▶ A data collection that needs to be :
  - ▶ Organized
  - ▶ Searchable
  - ▶ Up-to-date
- ▶ Challenge:
  - ▶ Change “meaningless” data into useful, accessible information

# A spreadsheet can be a Database

A spreadsheet contains:

- ▶ Rectangular data
- ▶ Structured
- ▶ No metadata

Search tools:

- ▶ Excel
- ▶ grep
- ▶ python/R

SNP ID	SNPSeq ID	Gene	+primer	-primer	Hap A	Hap B	Hap C
D1Mit160_1	10.MMHAP6 7FLD1.seq	lymphocyte antigen 84	AAGGTAAAA GGCATCAG CACAGCC	TCAACCTGG AGTCAGAGG CT	C	—	A
M-05554_1	12.MMHAP3 1FLD3.seq	procollagen, type III, alpha	TGCAGAA GCTGAAGTC TA	TTTGAGGT GTTAATGGTT CT	C	—	A
M-05554_2	X60184	complement component factor i	ACTTCCAGC CCTGGCTCT	ATATGCCACC AAGAAGCA	A	C	—
M-09947_3	AF067835	caspase 8	TCACAGAGG GAAACATGA AG	CTCCACATTG AACCAAAGC A	G	C	T
M-11415_1	U02023	insulin-like growth factor binding protein	GGGAAAAGC CTGAAAGAA GC	AGCTGAAAC CGGACATCA AT	T	G	—
D1Mit284_3	J05234	nucleolin	TGTTGGAAC CGACTTCTTC A	AAGAGTCAA AGAATTATG GAATGA	G	T	T

# A filesystem can be a Database

Hierarchical data:

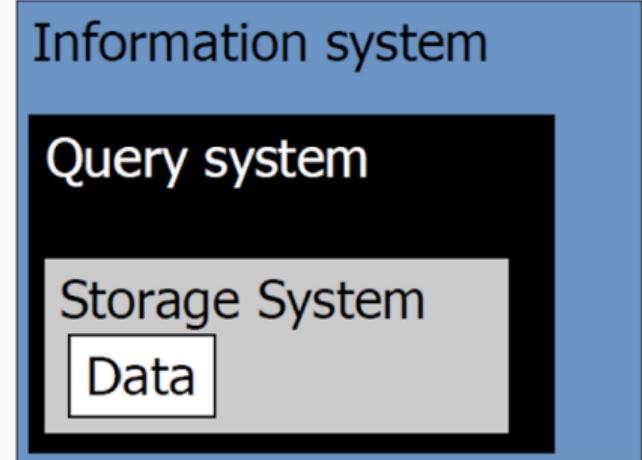
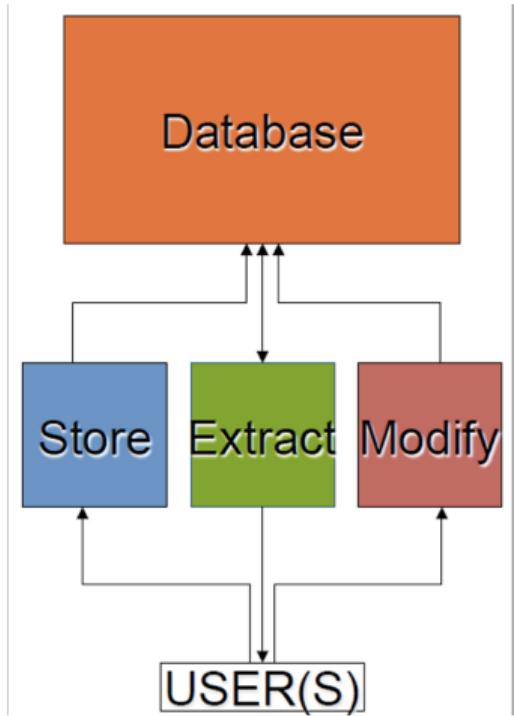
- ▶ Some metadata:  
File, symlink, etc
- ▶ Unstructured

Search tools:

- ▶ ls
- ▶ find
- ▶ grep

```
--(./projectnb)-
--(2019-02-21:16:27)-- tree bubhub | head -n 40
bubhub
├── bamcmp.simg
└── bubhub-conda
    ├── contributors.txt
    ├── package_recipes
    │   ├── args
    │   │   ├── bld.bat
    │   │   └── build.sh
    │   ├── bash_kernel
    │   ├── blast
    │   ├── bx-python
    │   │   ├── build.sh
    │   │   └── bx-python
    │       └── build
    │           ├── bdist.linux-x86_64
    │           ├── lib.linux-x86_64-3.6
    │           └── bx
    │               ├── align
    │               │   ├── axt.py
    │               │   ├── _core.cpython-36m-x86_64-linux-gnu.so
    │               │   ├── core.py
    │               │   ├── _epo.cpython-36m-x86_64-linux-gnu.so
    │               │   ├── epo.py
    │               │   ├── epo_tests.py
    │               │   ├── __init__.py
    │               │   ├── lav.py
    │               │   ├── lav_tests.py
    │               │   ├── maf.py
    │               │   ├── maf_tests.py
    │               │   ├── score.py
    │               │   ├── score_tests.py
    │               └── sitemask
    │                   ├── core.py
    │                   ├── _cpg.cpython-36m-x86_64-linux-gnu.so
    │                   ├── cpg.py
    │                   ├── __init__.py
    │                   ├── quality.py
    │                   └── sitemask_tests.py
    └── tools
```

# Organization and Types of Databases



# Organization and Types of Databases

Every database has tools that: Store, Extract, Modify

## **Flat file databases (flat DBMS)**

- Simple, restrictive, table

## **Hierarchical databases**

- Simple, restrictive, tables

## **Relational databases (RDBMS)**

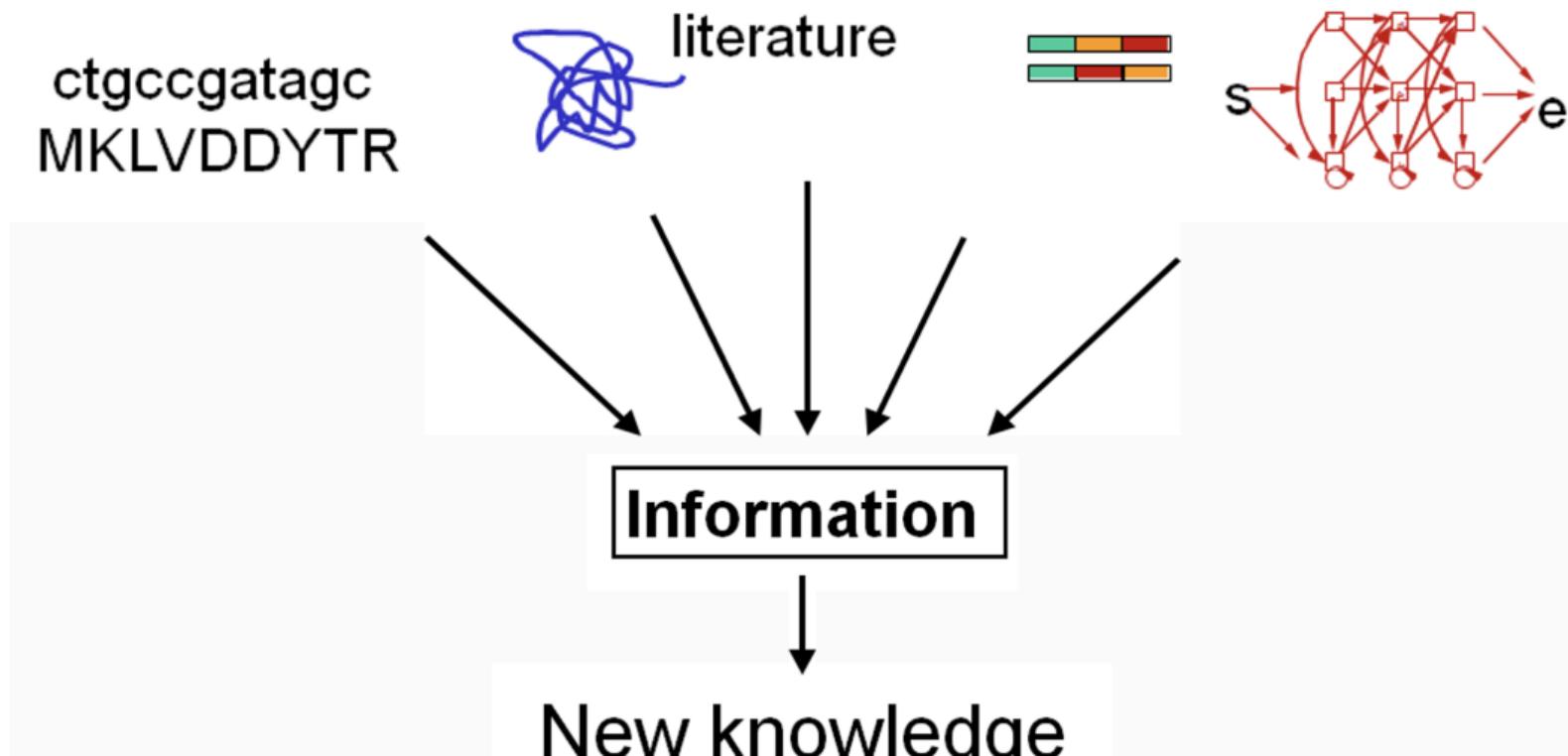
- Complex, versatile, tables

## **Object-oriented databases (ODBMS)**

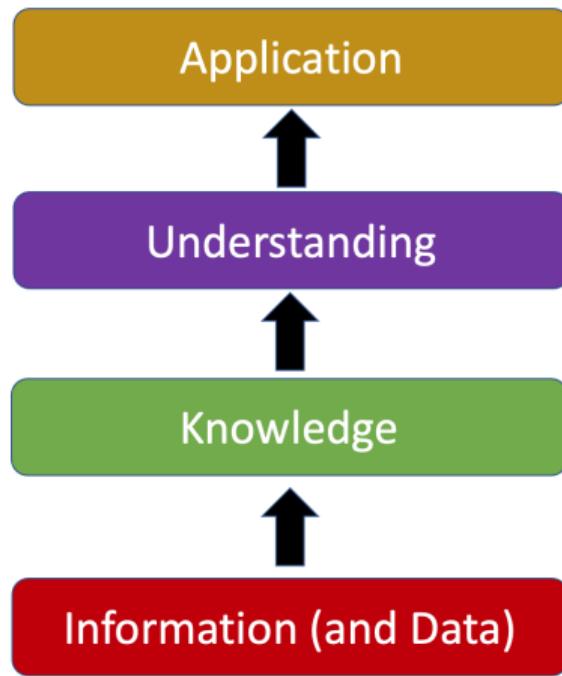
- Data warehouses and distributed databases

## **Unstructured databases (object store DBs)**

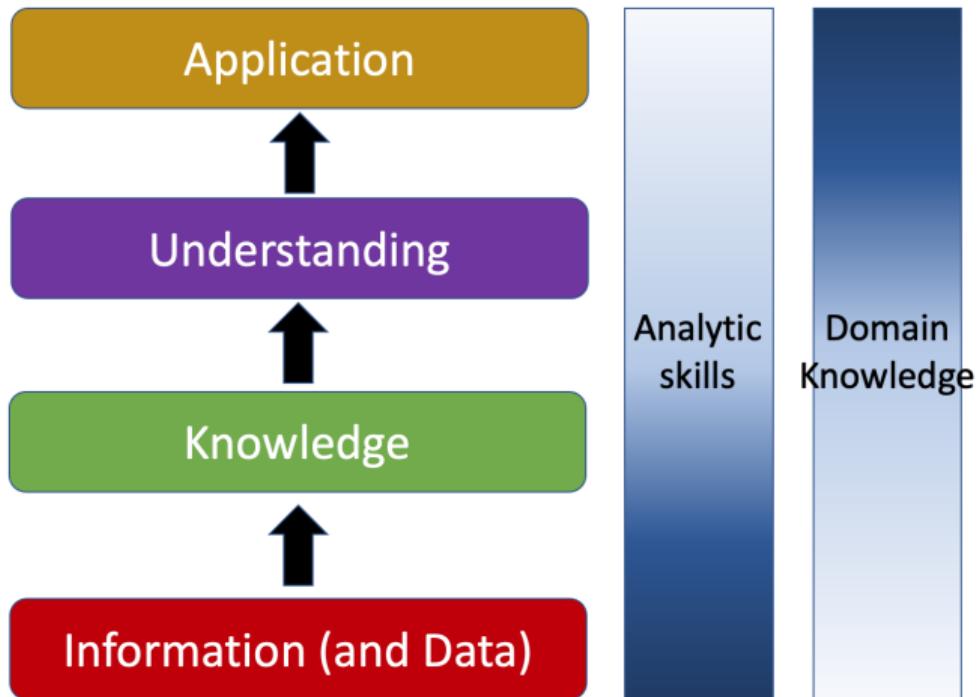
# Where do the data come from?



# Using biomedical data



# Using biomedical data



# Types of Biological Data

Primary data types (observed properties):

- ▶ Molecular Sequence: nucleic or amino acids
- ▶ Quantity: DNA, RNA, Protein, cell count, metabolites
- ▶ Locality: membrane, nucleus, epithelium
- ▶ Structure: 3D conformation, proximity, size

Secondary data types (inferred properties):

- ▶ Molecular Function
- ▶ Dynamics
- ▶ Relation: multiplicity, distribution, binding
- ▶ Association: co-occurrence, correlation
- ▶ Predicted: computational models

# Types of Biological Databases

Primary Databases:

- ▶ Original submissions by experimentalists
- ▶ Content controlled by the submitter
- ▶ Examples: GenBank, Trace, SRA, SNP, GEO

Secondary databases:

- ▶ Results of analysis of primary databases
- ▶ Aggregate of many databases
- ▶ Content controlled by third party (NCBI)
- ▶ Examples: NCBI Protein, Refseq, RefSNP, GEO datasets, UniGene, Homologene, Structure, Conserved Domain

# The NCBI hosts more than 50 different tools and databases

National Center for Biotechnology Information (NCBI):

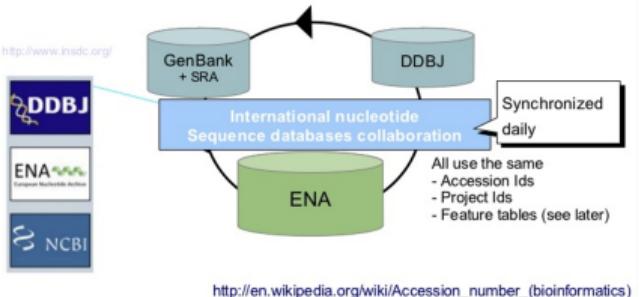
- ▶ **PubMed** - search for citations/papers
- ▶ **BLAST** - search for sequences
- ▶ **Nucleotide** - all nucleotide sequences (DNA, etc)
- ▶ **Genome** - published genomes
- ▶ **Protein** - amino acid sequences
- ▶ **SNP (dbSNP)** - all human SNPs
- ▶ **dbGAP** - Controlled access datasets

# International Sequence Database Collaboration

**Primary sequence dbs are synchronised and every sequence receives a unique identifier**

All database maintainers assign and share a unique **accession number** (AC) to each sequence – besides their own ID number – (info at NCBI). Sequences can get updated, and the accession number is extended with a version number, e.g..1 (see SVA)

Example of acc number: BC010109.2



- ▶ International Sequence Database Collaboration
- ▶ National Centre for Biotechnology Information (NCBI)
- ▶ European Nucleotide Archive (ENA)
- ▶ DNA Data Bank of Japan (DDBJ)

# PubMed is a search engine for literature

- ▶ Citation/publication databases
- ▶ Medline: <https://www.nlm.nih.gov/bsd/pmresources.html>
  - ▶ NLM journal citation database.
  - ▶ Includes citations 5,600 scholarly journals
- ▶ PubMed <https://www.ncbi.nlm.nih.gov/pubmed/>
  - ▶ Includes MEDLINE
  - ▶ journals/manuscripts deposited in PMC
  - ▶ NCBI Bookshelf

# Searching PubMed with MeSH terms

**MeSH (Medical Subject Headings)** is the NLM controlled vocabulary used for indexing articles for PubMed.

- ▶ the U.S. National Library of Medicine's controlled vocabulary
- ▶ arranged in a hierarchical manner called the MeSH Tree Structures updated annually



The screenshot shows the PubMed search interface. The search query '(microRNA[Title]) AND bastola[Author]' has been entered. The results page displays a single article titled 'Contribution of bioinformatics prediction in microRNA-based cancer therapeutics.' by Banwait JK<sup>1</sup>, Bastola DR<sup>2</sup>. The article is from *Adv Drug Deliv Rev*, 2015 Jan;81:94-103. It includes links to Elsevier full-text and PMC full-text. The right sidebar provides options to save items and add to favorites.

Format: Abstract ▾

Send to ▾

Full text links

ELSEVIER FULL-TEXT ARTICLE    PMC Full text FREE

Save items

Add to Favorites ▾

Center for Data Science

# Google Scholar

Google Scholar is another alternative for finding publications:



**W. Evan Johnson** 

Professor of Medicine  
Verified email at rutgers.edu

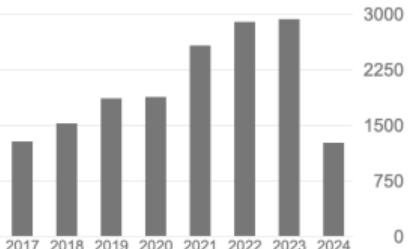
Data Science Computational Biology Bioinformatics Metagenomics Tuberculosis



	CITED BY	YEAR
<input type="checkbox"/> <a href="#">Adjusting batch effects in microarray expression data using empirical Bayes methods</a> WE Johnson, C Li, A Rabinovic <i>Biostatistics</i> 8 (1), 118-127	7202	2007
<input type="checkbox"/> <a href="#">The sva package for removing batch effects and other unwanted variation in high-throughput experiments</a> JT Leek, WE Johnson, HS Parker, AE Jaffe, JD Storey <i>Bioinformatics</i> 28 (6), 882-883	4502	2012
<input type="checkbox"/> <a href="#">Tackling the widespread and critical impact of batch effects in high-throughput data</a> JT Leek, RB Scharpf, HC Bravo, D Simcha, B Langmead, WE Johnson, ... <i>Nature Reviews Genetics</i> 11 (10), 733-739	2091	2010
<input type="checkbox"/> <a href="#">... 5 more</a>	14	2010

**Cited by** [VIEW ALL](#)

	All	Since 2019
Citations	21219	13432
h-index	42	30
i10-index	88	77



**Public access** [VIEW ALL](#)

	0 articles	97 articles
0 articles	0	97
97 articles	97	0



 Center for Data Science

## Collections of DNA/RNA sequences

# GenBank is an annotated collection of all publicly available DNA sequences

GenBank: <https://www.ncbi.nlm.nih.gov/genbank/>

- ▶ Flat file
- ▶ DNA only sequence database
- ▶ Archival in nature: Historical, Redundant
- ▶ Sample GenBank record (accession number U49845)
  - ▶ NCBI: <https://www.ncbi.nlm.nih.gov/nuccore/U49845>
  - ▶ ENA: <https://www.ebi.ac.uk/ena/data/view/U49845>
  - ▶ DDBJ: <http://getentry.ddbj.nig.ac.jp/top-e.html>



# GenBank Flat File

# Header

- Title
  - Taxonomy
  - Citation

## Features (AA seq)

# Ensembl (European Bioinformatics Institute)

- ▶ Comprehensive DNA/RNA sequence and annotation database
- ▶ Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data
- ▶ Analysis tools:
  - ▶ BLAST
  - ▶ BioMart
  - ▶ Variant Effect Predictor

 Ensembl

BLAST/BLAST | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh38.p10) ▾

Location: 13:32,315,474-32,400,266 Gene: BRCA2

Gene-based displays

- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Comparative Genomics
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
  - Ensembl protein families
- Ontologies
  - GO: Biological process
  - GO: Molecular function
  - GO: Cellular component
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
  - Gene expression
  - Regulation
  - External references
  - Supporting evidence
- ID History
  - Gene history

**Gene: BRCA2 ENSG00000139618**

Description: BRCA2, DNA repair associated [Source HGNC Symbol/Accession number: FACC, FAD, FANCD1, BRCC2, FANCD, FAD1, XRCC11]

Synonyms: Chromosome 13:32,315,474-32,400,266 forward strand.

Location: GRCh38 CM000675.2

About this gene: This gene has 7 transcripts (splice variants), 88 orthologues

Transcripts: Hide transcript table

Name	Transcript ID	bp	Protein	Biotype	CCDS
BRCA2-201	ENST00000380152.7	11986	3418aa	Protein coding	CCDS0344.0
BRCA2-202	ENST00000544455.5	10984	3418aa	Protein coding	CCDS0344.0
BRCA2-202	ENST00000470094.1	842	106aa	Nonsense mediated decay	-
BRCA2-203	ENST00000528762.1	495	64aa	Nonsense mediated decay	-
BRCA2-207	ENST00000614259.1	7950	No protein	Processed transcript	-
BRCA2-204	ENST00000530893.6	2011	No protein	Processed transcript	-
BRCA2-205	ENST00000533776.1	523	No protein	Retained intron	-

Show/hide columns (1 hidden)

**Summary** ⓘ

## Databases for reference sequences

# Reference genomes

- ▶ NCBI Genome -  
<https://www.ncbi.nlm.nih.gov/datasets/genome/>
  - ▶ Contains both Genbank and Refseq accessions
- ▶ Ensembl - <https://useast.ensembl.org/index.html>

# microRNAs

- ▶ miRbase: <https://mirbase.org> allows you to search and browse miRNAs from a host of species
- ▶ Enables download of genomic coordinates for miRNAs (.gff) and sequence (.fa)

# rRNAs and tRNAs databases

- ▶ rRNAs
- ▶ Greengenes - <https://greengenes.secondgenome.com>
- ▶ Silva - <https://www.arb-silva.de>
- ▶ tRNAs
  - ▶ gtRNAdb- <http://gtrnadb.ucsc.edu>

# Species-specific databases

For other well studied model organisms, you can often find collections of resources and tools:

- ▶ Flybase - *D. melanogaster*
- ▶ Wormbase - *C. elegans*
- ▶ Zfin - *D. rerio*
- ▶ Influenza (link!)

# RNACentral



- ▶ Developed by EBI and Wellcome Trust
- ▶ Integration of many other up-to-date RNA resources and tools

Click here for an overview of core functions

Existing datasets – sequencing data (SRA or GEO)

# dbGaP: genotype-phenotype interactions in humans

- ▶ Most studies contain PHI and are subject to strict access control and usage regulations
- ▶ Applications required to be granted access to datasets
- ▶ Any data with sensitive health information must be stored, handled, and interacted with according to appropriate regulations

# SRA - Sequencing Read Archive

- ▶ Raw sequencing data and alignment info
- ▶ Metagenomics, environmental samples, biomedical sequencing

Download via:

- ▶ SRA-toolkit
- ▶ FTP links on EMBL ENA

# A quick demonstration of SRA-toolkit and EMBL-ENA

- ▶ SRA-toolkit is the official tool released by NCBI to directly download SRA files
  - ▶ Notorious for being obtuse to use and confusing commands / documentation
- ▶ EMBL-ENA hosts FTP links directly
  - ▶ Most but not all SRA accessions available
  - ▶ Have to use wget, curl, or other methods to download

# A quick demonstration of SRA-toolkit

Download a file for an asthma host microbiome dataset

```
## attach the sratoolkit
module load sratoolkit

## Save accession to download
acc = "SRR1528344"

## Download using fastq-dump
fastq-dump $acc
# option --split-3 is needed for paired end reads

## don't forget to compress the file!
gzip $acc.fastq
```

# A quick demonstration of SRA-toolkit

Download all files for an asthma host microbiome dataset

```
accs=( $( cat SRR_Acc_List.txt ) )
for i in $(seq 0 ${#accs[@]})
do
    fastq-dump ${a[i]};
    gzip ${a[i]}*
done
```

# Batch script (SLURM)

```
#!/bin/bash
#SBATCH --job-name=microbiome_download
#SBATCH --mem=1G
#SBATCH --time=01:00:00

module load sratoolkit
cd $HOME/tmp/

accs=( $( cat SRR_Acc_List.txt ) )
for i in ${seq 0 ${#accs[@]}}
do
    fastq-dump ${a[i]};
    gzip ${a[i]}*
done
```

Save as a file and use sbatch to submit

# Batch array (SLURM)

```
#!/bin/bash
#SBATCH --job-name=microbiome_download
#SBATCH --output=asthma.out
#SBATCH --array=0-27
#SBATCH --cpus-per-task=1
#SBATCH --mem=1G
#SBATCH --time=00:20:00

module load sratoolkit
cd /scratch/$USER

accs=$( $( cat SRR_Acc_List.txt ) )
acc_number=${accs[$SLURM_ARRAY_TASK_ID]}

fastq-dump --gzip $acc_number
```

Save as a file and use sbatch to submit

# Gene Expression Data

## GEO - Genome Expression Omnibus

- ▶ Tied to SRA via Bioproject ID
- ▶ GEO also contains processed data, typically specific to a publication
- ▶ You may find: Alignments (BAM/SAM), visualizations (.bg, .bed) and intermediate results (delimited formats)

# GEOquery example

```
# Load the geoquery library: BiocManager::install("GEOquery")
library(GEOquery)

# Search for a dataset in GEO
geo_search <- getGEO("GSE1297", GSEMatrix = TRUE)

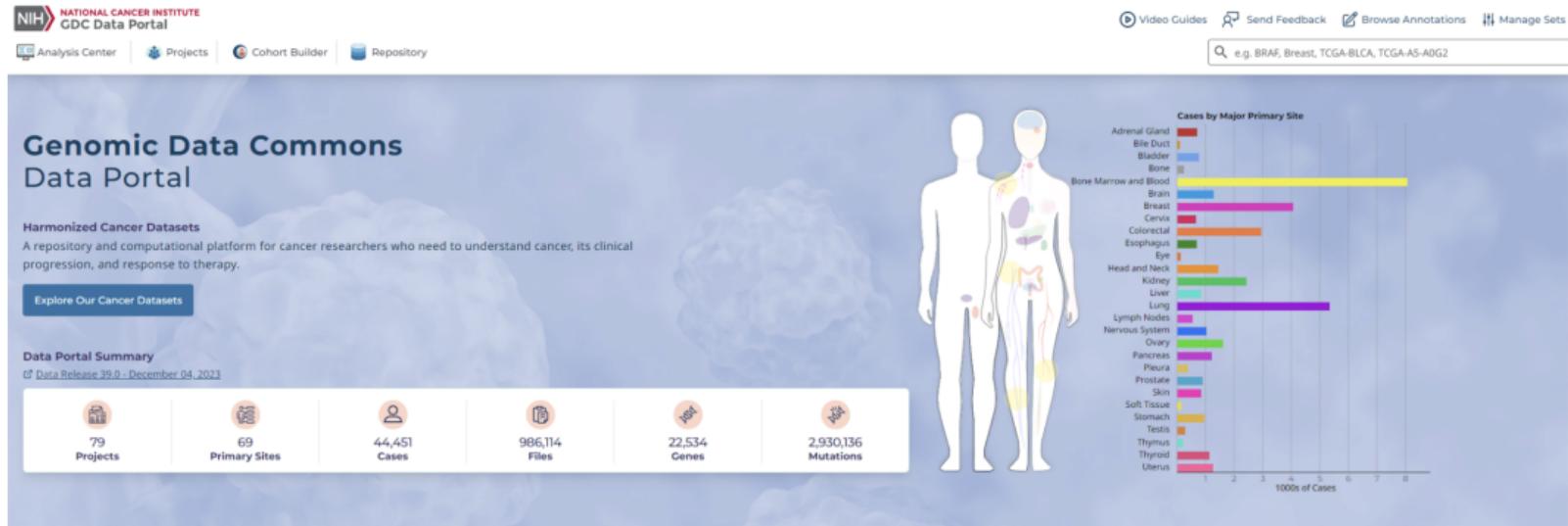
# Display basic information about the dataset
print(geo_search)

# Extract expression data
expression_data <- exprs(geo_search[[1]])

# Display the first few rows of expression data
head(expression_data)
```

## Domain-specific data

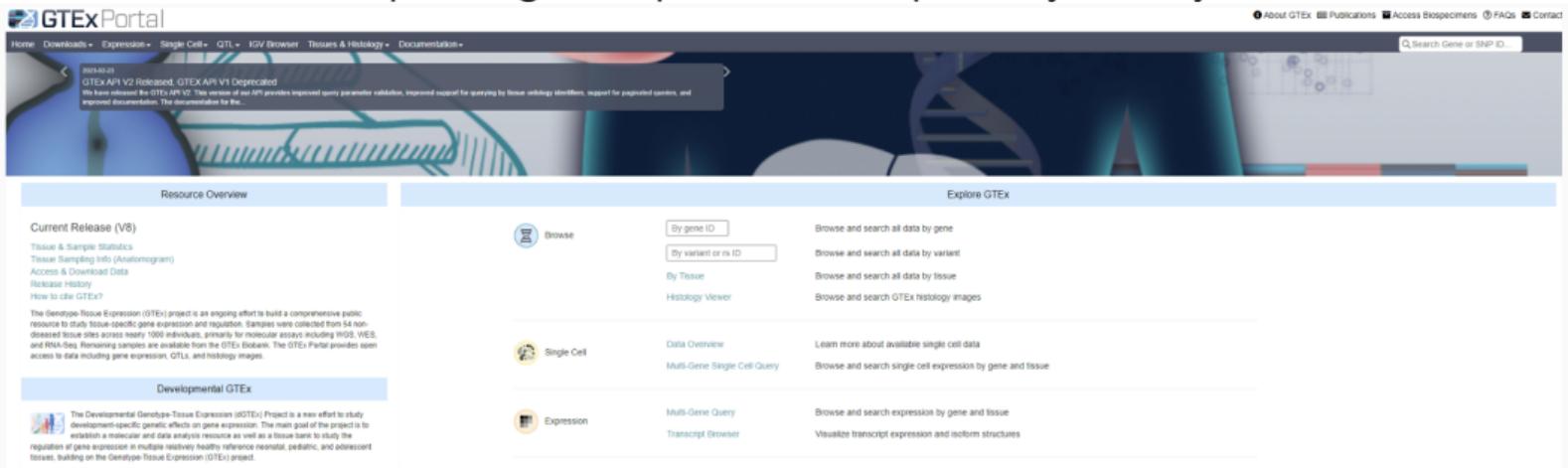
# TCGA / NCI Genomic Data Commons



<https://portal.gdc.cancer.gov>

# Tissue-specific gene expression

GTEx collects tissue-specific gene expression from primarily healthy individuals



The screenshot shows the GTEx Portal homepage. At the top, there is a banner about the release of API V2. Below the banner, there is a search bar and a navigation menu with links like Home, Downloads, Expressions, Single Cell, QTL, IGV Browser, Tissues & Histology, Documentation, About GTEx, Publications, Access Biospecimens, FAQs, and Contact.

The main content area is divided into several sections:

- Resource Overview:** Includes links to Tissue & Sample Statistics, Tissue Sampling Info (Anatomogram), Access Biospecimen Data, Release History, and How to cite GTEx?
- Explore GTEx:** Features a "Browse" section with links to By gene ID, By variant or rs ID, By Tissue, and Histology Viewer. It also includes sections for Single Cell (Data Overview, Multi-Gene Single Cell Query) and Expression (Multi-Gene Query, Transcript Browser).
- Developmental GTEx:** Describes the Developmental Genotype-Tissue Expression (dGTEx) Project, which aims to study tissue-specific genetic effects on gene expression. It mentions that samples are being collected as a tissue bank to study the regulation of gene expression in multiple relatively healthy reference neonatal, pediatric, and adolescent tissues, building on the Genotype-Tissue Expression (GTEx) project.

<https://www.gtexportal.org/home/>

# DNA regulatory elements

ENCODE contains a host of information about DNA regulatory elements



<https://www.encodeproject.org>

# Protein databases

# NCBI Protein

Protein sequence database: <https://www.ncbi.nlm.nih.gov/protein/>

NCBI Resources How To Sign in to NCBI

Protein Protein BRAC Create alert Advanced Help

Species Summary 20 per page Sort by Default order Send to: Filters: Manage Filters

Animals (79,941)  
Plants (14,521)  
Fungi (12,247)  
Protists (4,931)  
Bacteria (7,704)  
Archaea (34)  
Viruses (571)  
Customize ...

Source databases PDB (232)  
RefSeq (73,065)  
UniProtKB / Swiss-Prot (516)  
Customize ...

Genetic compartments Chloroplast (2)  
Mitochondrion (53)  
Plasmid (12)  
Plastid (2)

Sequence length Custom range ...

See [braC branched-chain amino acid ABC transporter substrate-binding protein BraC](#) in the Gene database  
[braC reference sequences Protein \(1\)](#)

See the [results of this search \(87 items\)](#) in our new [Identical Protein Groups](#) database.

Items: 1 to 20 of 120705

<< First < Prev Page 1 of 6036 Next > Last >>

1. [BRAC..partial \[Poeciliopsis prolifica\]](#)  
106 aa protein  
Accession: JA055668.1 GI: 958322777  
BioProject Nucleotide Taxonomy  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

2. [BraC \[Pseudomonas putida BIRD-1\]](#)  
371 aa protein

Results by taxon

Top Organisms [Tree](#)  
[Homo sapiens \(1503\)](#)  
[Plasmodium falciparum \(1448\)](#)  
[Mycobacterium abscessus \(1298\)](#)  
[Mus musculus \(1240\)](#)  
[Rhizobius irregularis \(1203\)](#)  
All other taxa (114013)  
More...

Find related data Database: Select

Find items

Search details

Center for Data Science

# Genpept

## Chain A, Structure Of The Btb (tramtrack And Bric A Brac) Human Gigaxonin

PDB: 2PPI\_A

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: [View](#)

LOCUS 2PPI\_A 144 aa linear PRI 26-OCT-20  
 DEFINITION Chain A, Structure Of The Btb (tramtrack And Bric A Brac) Domain

Human Gigaxonin.

ACCESSION 2PPI\_A

VERSION 2PPI\_A

DBSOURCE pdb: molecule 2PPI, chain 65, release Oct 22, 2017;  
 deposition: Apr 30, 2007;  
 class: Structural Protein;  
 source: Mmdb\_id: [46639](#), Pdb\_id 1: 2PPI;  
 Exp. method: X-Ray Diffraction.

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
 Catarrhini; Hominidae; Homo.

REFERENCE 1 (residues 1 to 144)

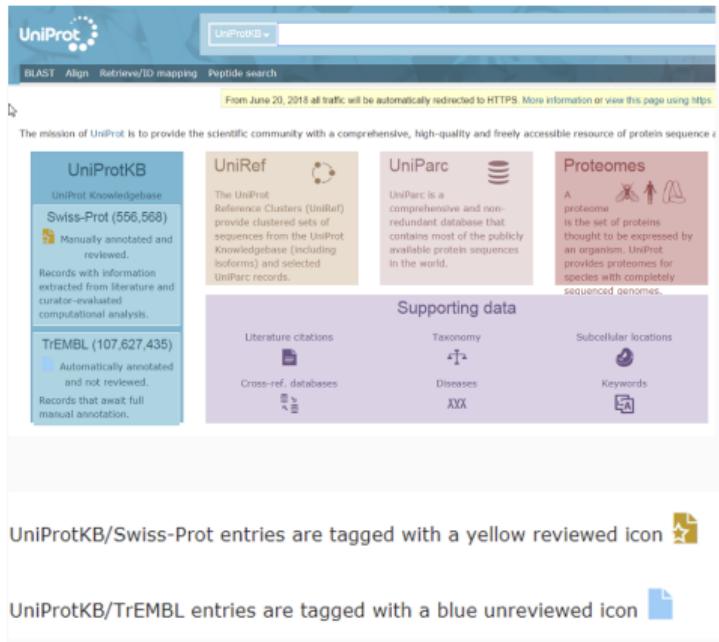
AUTHORS Amos,A., Turnbull,A.P., Tickle,J., Keates,T., Bullock,A., S.P.,  
 Burgess-Brown,N., Debreczeni,J.E., Ugochukwu,E., Umeano,C.,  
 Pike,A.C.W., Papagrigoriou,E., Sundstrom,M., Arrowsmith,C.H.,  
 Weigelt,J., Edwards,A., Von Delft,F. and Knapp,S.

TITLE Structure Of The Btb (Tramtrack And Bric A Brac) Domain Human Gigaxonin  
 JOURNAL Unpublished

Region	//region_name="Domain 1" /region_name="BTB/POZ domain; pfam00651" /db_xref="CDD: <a href="#">279045</a> " 46..96
Region	/region_name="Domain 2" /note="NCBI Domains" 49..144
SecStr	/region_name="BTB" /note="Broad-Complex, Tramtrack and Bric a brac; smart00225" /db_xref="CDD: <a href="#">197585</a> " 49..55
SecStr	/sec_str_type="sheet" /note="strand 1" 56..62
SecStr	/sec_str_type="sheet" /note="strand 2" 63..70
SecStr	/sec_str_type="helix" /note="helix 2" 71..79
SecStr	/sec_str_type="helix" /note="helix 3" 89..94
SecStr	/sec_str_type="sheet" /note="strand 3" 98..109
SecStr	/sec_str_type="helix" /note="helix 4" 120..130
SecStr	/sec_str_type="helix" /note="helix 5"

# Uniprot

- ▶ The Universal Protein Resource
- ▶ Comprehensive resource for protein sequence and annotation data
- ▶ Collaboration between:
  - ▶ EMBL-EBI
  - ▶ Swiss Institute of Bioinformatics
  - ▶ Protein Information Resource
- ▶ Entries in two categories:
  - ▶ Swiss-Prot (experimentally verified)
  - ▶ TrEMBL (computer-annotated)
- ▶ <http://www.uniprot.org/>



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and function.

**UniProtKB**  
UniProt Knowledgebase  
Swiss-Prot (556,568)  
Manually annotated and reviewed.  
Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (107,627,435)  
Automatically annotated and not reviewed.  
Records that await full manual annotation.

**UniRef**  
The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

**UniParc**  
UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

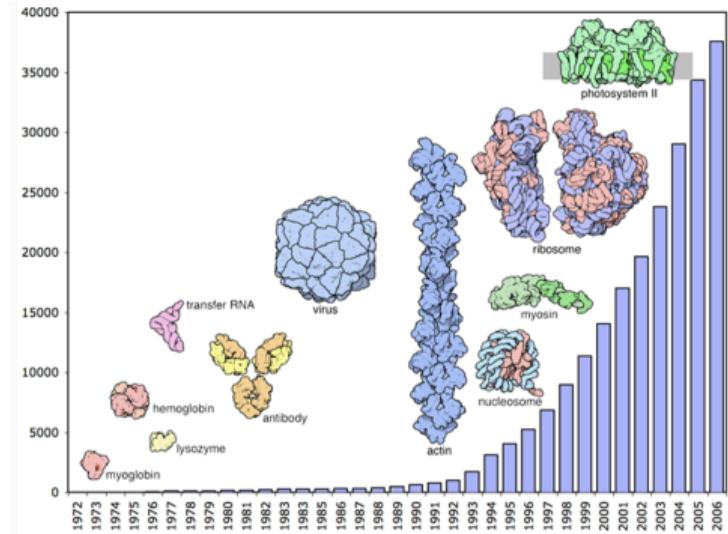
**Supporting data**  
Literature citations, Cross-ref. databases, Taxonomy, Diseases, Subcellular locations, Keywords.

UniProtKB/Swiss-Prot entries are tagged with a yellow reviewed icon 

UniProtKB/TrEMBL entries are tagged with a blue unreviewed icon 

# Protein Structure database - PDB

- ▶ Protein Data Bank (PDB)  
<http://www.rcsb.org/>
- ▶ Dedicated to 3D structure of proteins and peptides
- ▶ ~150,000 predicted and experimental (solved) structures

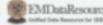


# PDB: kinesin 6

**149174 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education**

Search by PDB ID, author, macromolecule, sequence, or ligands

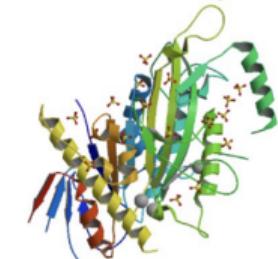
[Advanced Search](#) | [Browse by Annotations](#)

[Structure Summary](#) [3D View](#) [Annotations](#) [Sequence](#) [Sequence Similarity](#) [Structure Similarity](#) [Experiment](#)

**5X3E**



**kinesin 6**

**DOI:** [10.2210/pdb5X3E/pdb](https://doi.org/10.2210/pdb5X3E/pdb)

**Classification:** [MOTOR PROTEIN](#)

**Organism(s):** [Caenorhabditis elegans](#)

**Expression System:** [Escherichia coli-Thermus thermophilus shuttle vector pTRH1](#)

**Deposited:** 2017-02-04 **Released:** 2017-04-19

**Deposition Author(s):** [Chen, Z., Guan, R., Zhang, L.](#)

**Funding Organization(s):** Chinese Key Research Plan-Protein Sciences; National Natural Science Foundation of China; Junior One Thousand Talents program; National Science Foundation of China; 863 Program

**Experimental Data Snapshot**

**Method:** X-RAY DIFFRACTION  
**Resolution:** 2.61 Å  
**R-Value Free:** 0.240  
**R-Value Work:** 0.209

**wwPDB Validation**

Metric	Percentile Ranks	Value
Rfree	0.240	7
Clashscore	7	0
Ramachandran outliers	0	1.1%
Sidechain outliers	1.1%	13.0%
RSRZ outliers	13.0%	0



 Center for Data Science

[3D View: Structure](#) | [Electron Density](#) | [Ligand Interaction](#)

[Standalone Viewers](#)  
[Protein Workshop](#) | [Ligand Explorer](#)

# Protein Family Database

- ▶ <http://pfam.xfam.org/family/piwi>
- ▶ Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using Hidden Markov Models

**Family: Piwi (PF02171)**

Summary Domain organisation

139 architectures 3730 sequences 4 interactions 568 species 103 structures

**Domain organisation**

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

**There are 893 sequences with the following architecture: ArgoN, ArgoL1, PAZ, ArgoL2, ArgoMid, Piwi**

X1WG39\_DANRE [Danio rerio (Zebrafish) (Brachydanio rerio)] Uncharacterized protein {ECO:0000313|Ensembl:ENSDARP00000129194} (858 residues)



[Show all sequences with this architecture.](#)

**There are 678 sequences with the following architecture: Piwi**

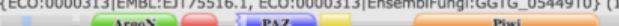
Z4YLE4\_MOUSE [Mus musculus (Mouse)] Piwi-like protein 4 {ECO:0000313|Ensembl:ENSMUSP00000111307} (458 residues)



[Show all sequences with this architecture.](#)

**There are 581 sequences with the following architecture: ArgoN, ArgoL1, PAZ, ArgoL2, Piwi**

J3NVY6\_GAGT3 [Gaeumannomyces graminis var. tritici (strain R3-111a-1) (Wheat and barley take-all root rot fungus)] Uncharacterized protein {ECO:0000313|EMBL:ET75516.1, ECO:0000313|EnsemblFungi:GGTG\_05449T0} (1022 residues)



[Show all sequences with this architecture.](#)

**There are 447 sequences with the following architecture: PAZ, Piwi**

V4B7N4\_LOTGI [Lottia gigantea (Giant owl limpet)] Uncharacterized protein {ECO:0000313|EMBL:ESO84639.1} (791 residues)

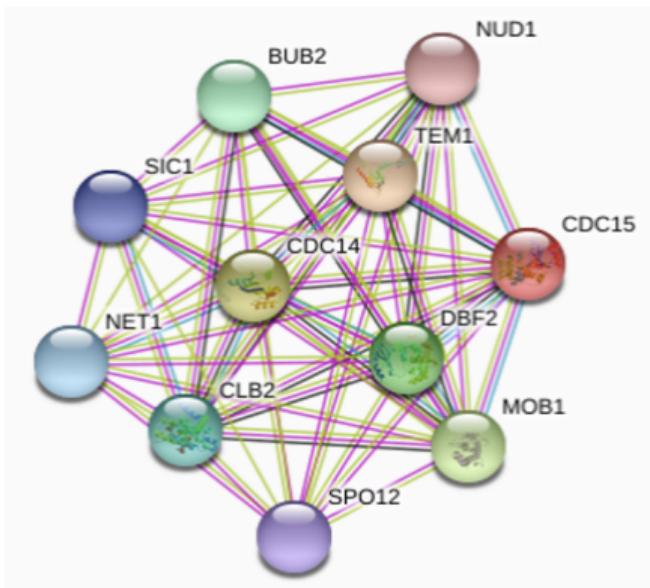


[Show all sequences with this architecture.](#)

Center for Data Science

# Protein-Protein Interaction Database

- ▶ STRING: <https://string-db.org/>
- ▶ Search Tool for the Retrieval of Interacting Genes/Proteins
- ▶ Database of protein/protein interactions
- ▶ Information from numerous sources:
  - ▶ experimental data
  - ▶ computational prediction methods
  - ▶ public text collections
- ▶ Expressed as interaction graphs:
  - ▶ Nodes: Network nodes represent proteins
  - ▶ Edges: Edges represent protein-protein associations



## Curated resources

# Data vs Annotation Database

- ▶ RefSeq: curated nonredundant biological sequences  
<https://www.ncbi.nlm.nih.gov/refseq/>
  - ▶ Source: Genbank (INSDC)
  - ▶ Annotated: Community collaboration, automated computer, NCBI staff curation
- ▶ Advantages of using RefSeq
  - ▶ Non-redundancy
  - ▶ Curated, validated
  - ▶ Format consistency
  - ▶ Distinct accession series
  - ▶ Updates to reflect current sequence data and biology

# RefSeq Annotations

## mRNAs and Proteins

NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA

## Gene Records

NG_123456	Reference Genomic Sequence
-----------	----------------------------

## Chromosome

NC_123455	Microbial replicons, organelle
AC_123455	Alternate assemblies

## Assemblies

NT_123456	Contig
NW_123456	WGS Supercontig

# Avoid using outdated or abandoned tools / databases

Keep in mind the following:

- ▶ Check when it was last updated (ideally recently)
- ▶ Check for an associated peer-reviewed publication
- ▶ Check for a github repo
  - ▶ How recent was the last commit
  - ▶ Are they responsive to issues?

# Session info

```
sessionInfo()

## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.4.1
##
## Matrix products: default
## BLAS:    /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK:  /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics   grDevices  utils       datasets   methods    base
##
## loaded via a namespace (and not attached):
## [1] compiler_4.4.2    fastmap_1.2.0    cli_3.6.5      tools_4.4.2
## [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10    rmarkdown_2.29
## [9] knitr_1.50        xfun_0.52       digest_0.6.37   rlang_1.1.6
## [13] evaluate_1.0.3
```