

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

2024-06-12

Section 1

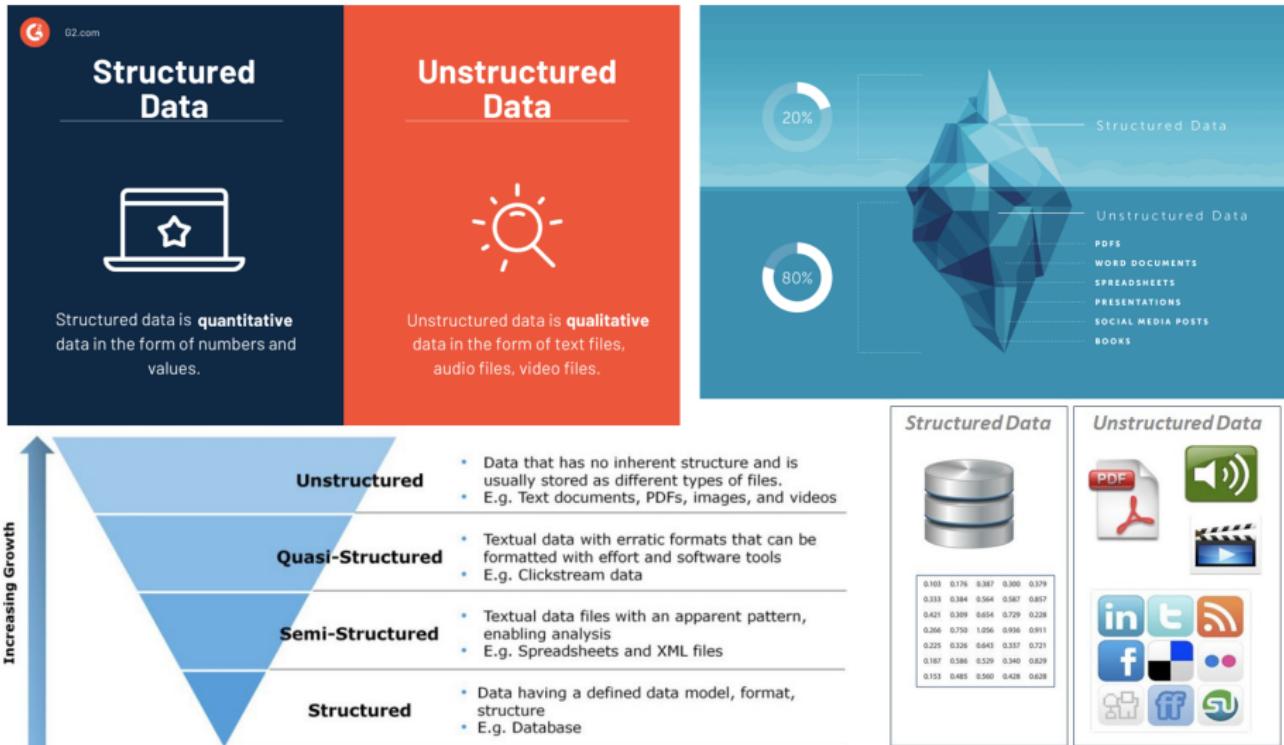
Biological Databases and resources

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	AAGGTAAAAA GGCAATCAG CACAGCC	TCAACCTGG AGTCAGAGG CT	C	—	A
M-05554_1	12.MMHAP3 1FLD3.seq	procollagen, type III, alpha	TGCGCGAGAA GCTGAAGTC TA	TTTTGAGGT 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 CTGAAAAGAA 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:

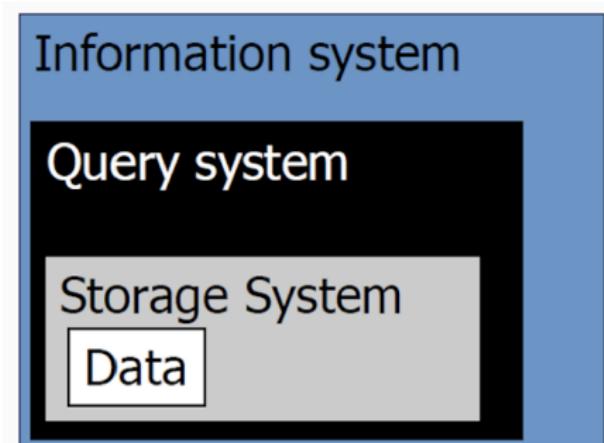
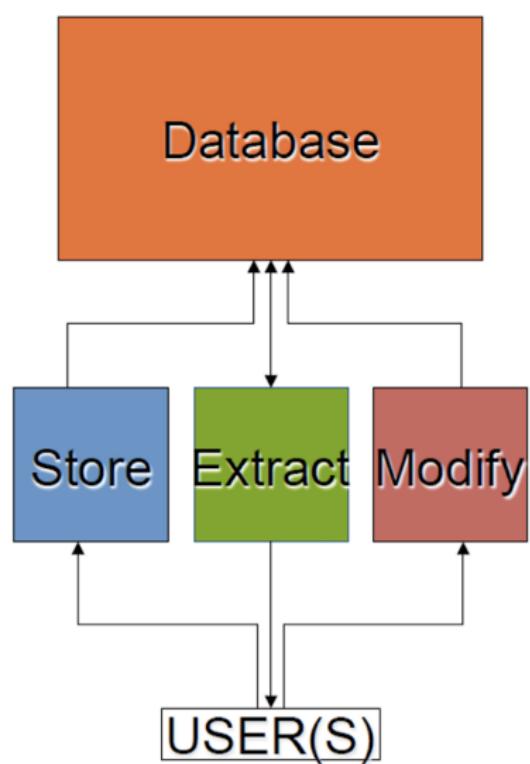
- Some metadata:
File, symlink, etc
- Unstructured

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)

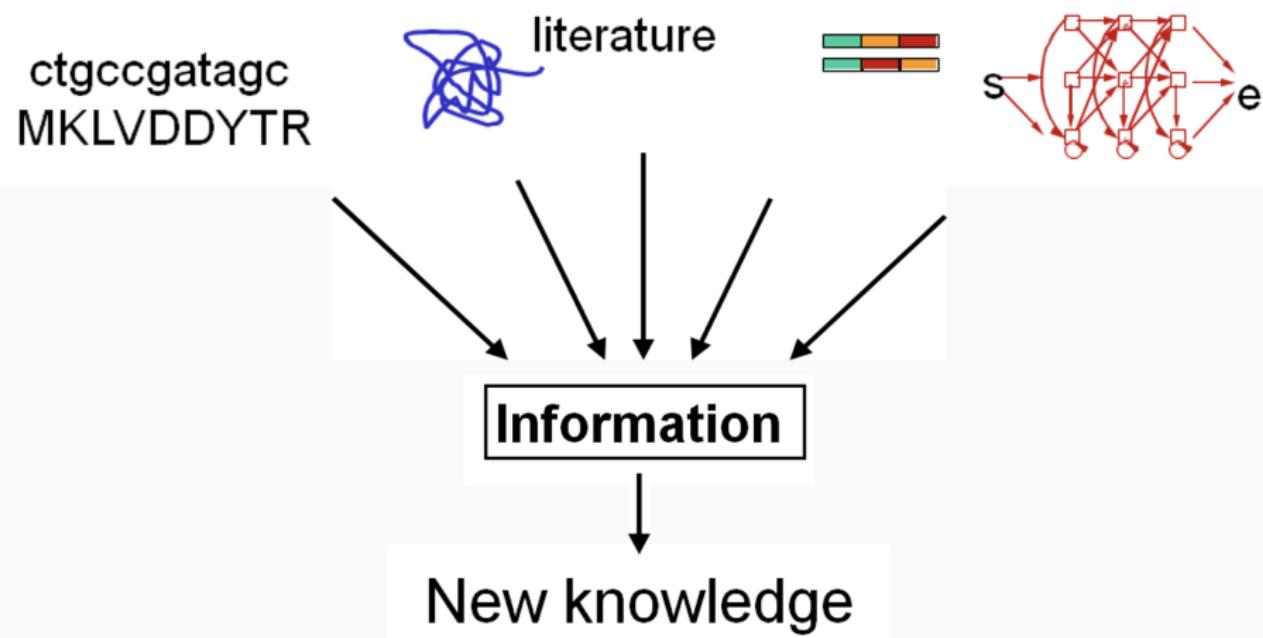
- Complex, versatile, tables

Object-oriented databases (ODBMS)

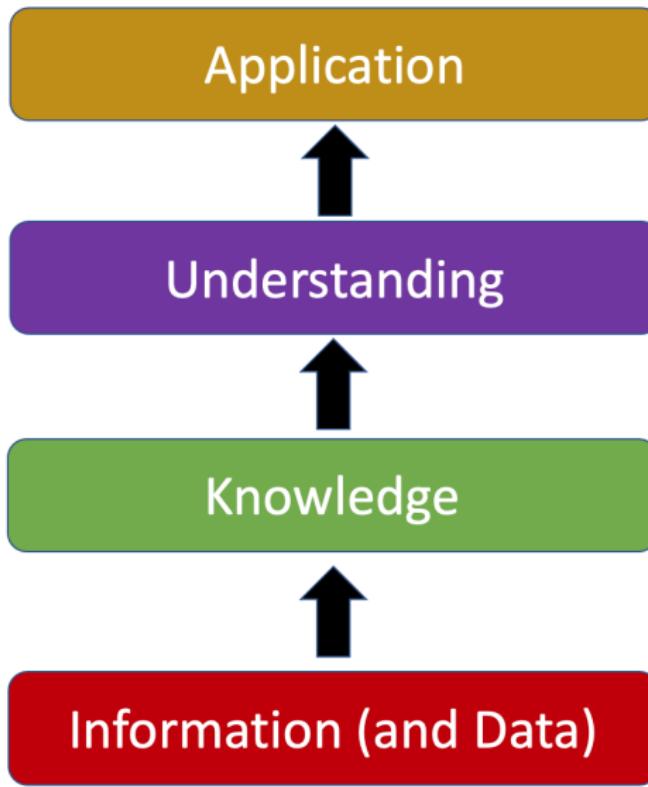
- Data warehouses and distributed databases

Unstructured databases (object store DBs)

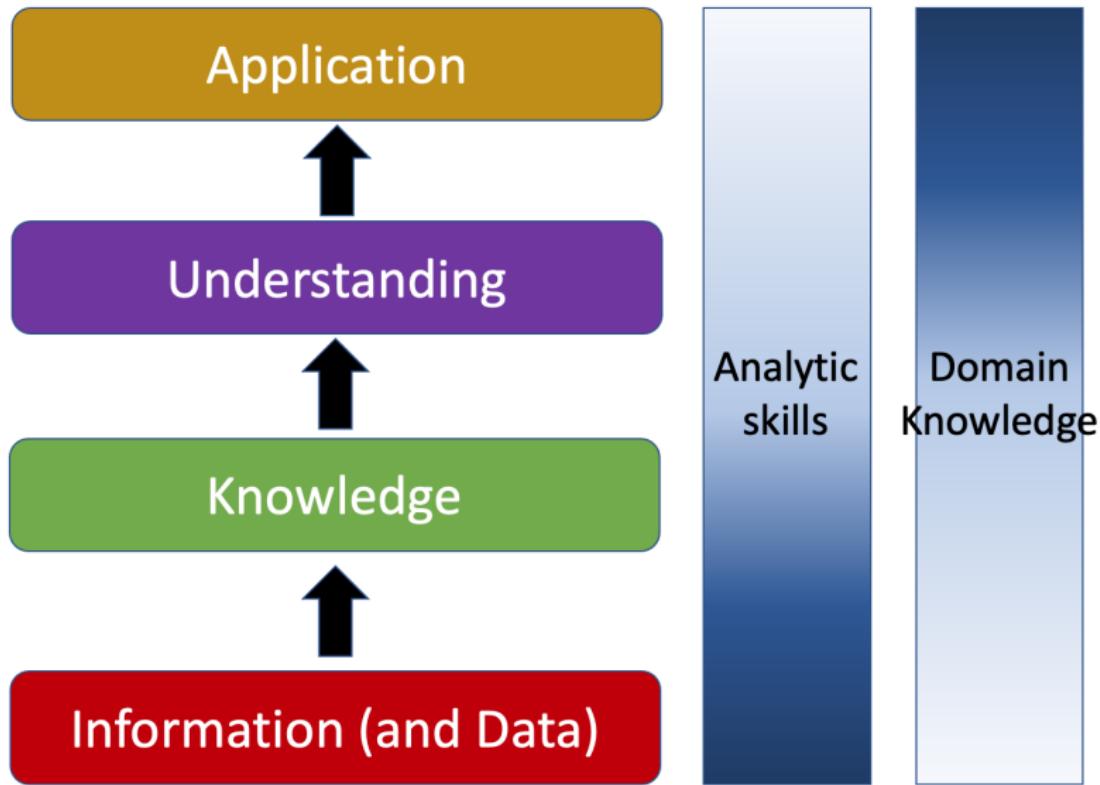
Where do the data come from?



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):

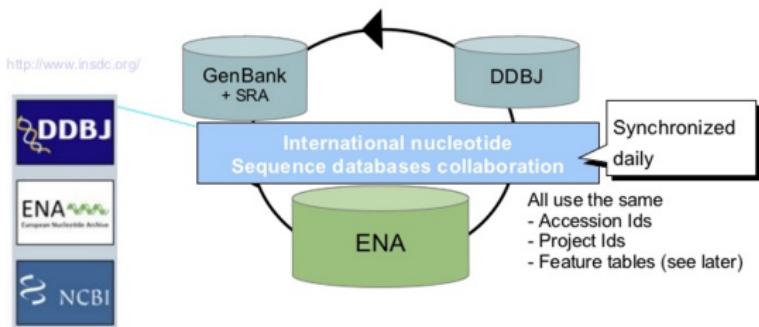
- **PubMed** - search for citations/papers
- **BLAST** - search for sequences
- **Nucleotide** - all nucleotide sequences (DNA, etc)
- **Genome** - published genomes
- **Protein** - amino acid sequences
- **SNP (dbSNP)** - all human SNPs
- **dbGAP** - Controlled access datasets

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 term '(microRNA[Title]) AND bastola[Author]' is entered in the search bar. The results page displays a single article titled 'Contribution of bioinformatics prediction in microRNA-based cancer therapeutics.' by Banwait JK and Bastola DR. The article is from *Adv Drug Deliv Rev*, 2015 Jan;81:94-103. The abstract discusses the use of bioinformatics to predict microRNAs in cancer therapeutics, mentioning their role in early diagnosis, prognosis, and treatment. The page includes links to Elsevier full-text and PMC full-text, as well as options to save items and add to favorites.

NCBI Resources How To Sign in to NCBI

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed (microRNA[Title]) AND bastola[Author] Search Help

Create RSS Create alert Advanced

Format: Abstract Send to Full text links

Elsevier FULL-TEXT ARTICLE PMC Full text

Save items Add to Favorites

Similar articles Review A review of computational approaches detecting miRNAs [Front Biosci (Landmark Ed). 2017]

Adv Drug Deliv Rev. 2015 Jan;81:94-103. doi: 10.1016/j.addr.2014.10.030. Epub 2014 Nov 6.

Contribution of bioinformatics prediction in microRNA-based cancer therapeutics.

Banwait JK¹, Bastola DR².

[Author information](#)

Abstract

Despite enormous efforts, cancer remains one of the most lethal diseases in the world. With the advancement of high throughput technologies massive amounts of cancer data can be accessed and analyzed. Bioinformatics provides a platform to assist biologists in developing minimally invasive biomarkers to detect cancer, and in designing effective personalized therapies to treat cancer patients. Still, the early diagnosis, prognosis, and treatment of cancer are an open challenge for the research community. MicroRNAs (miRNAs) are small non-coding RNAs that serve to regulate gene expression. The discovery of deregulated miRNAs in cancer cells and tissues has led many to

Google Scholar

Google Scholar is another alternative for finding publications:



W. Evan Johnson

FOLLOW

Professor of Medicine
Verified email at rutgers.edu

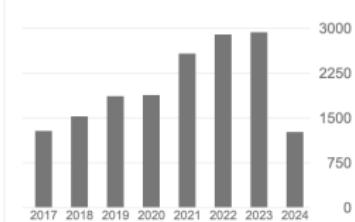
Data Science Computational Biology Bioinformatics Metagenomics Tuberculosis

<input type="checkbox"/> TITLE			CITED BY	YEAR
<input type="checkbox"/> Adjusting batch effects in microarray expression data using empirical Bayes methods WE Johnson, C Li, A Rabinovic <i>Biostatistics</i> 8 (1), 118-127			7202	2007
<input type="checkbox"/> The sva package for removing batch effects and other unwanted variation in high-throughput experiments JT Leek, WE Johnson, HS Parker, AE Jaffe, JD Storey <i>Bioinformatics</i> 28 (6), 882-883			4502	2012
<input type="checkbox"/> Tackling the widespread and critical impact of batch effects in high-throughput data 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"/> ComBat-seq: batch effect adjustment for RNA-seq count data Y Zhang, G Parmigiani, WE Johnson <i>NAR genomics and bioinformatics</i> 2 (3), lqaa078			708	2020
<input type="checkbox"/> Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing J O'Rawe, T Jiang, G Sun, Y Wu, W Wang, J Hu, P Bodily, L Tian, ... <i>Genome medicine</i> 5, 1-18			546	2013

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
------------	-------------

not available

available

Based on funding mandates

Co-authors

Section 2

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

- Title
 - Taxonomy
 - Citation

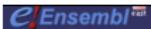
Header

Features (AA seq)

DNA Sequence

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

 Human (GRCh38.p10) ▾ Location: 13:32,315,474-32,400,266 Gene: BRCA2

Gene: BRCA2 ENSG00000139618

Description	BRCA2, DNA repair associated [Source HGNC Symbol, Acc FAD, FAD, FANC01, BRCC2, FANC, FAD1, XRCC11]				
Synonyms					
Location	Chromosome 13: 32,315,474-32,400,266 forward strand. GRCh38 CM000675.2				
About this gene	This gene has 7 transcripts (splice variants). 88 orthologues				
Transcripts	Hide transcript table				
Show/hide columns (1 hidden)					
Name	Transcript ID	bp	Protein	Biotype	CCDS
BRCA2-201	ENST00000380152.7	11986	3418aa	Protein coding	CCDS9344.e
BRCA2-206	ENST00000544455.5	10984	3418aa	Protein coding	CCDS9344.e
BRCA2-202	ENST00000470094.1	842	195aa	Nonsense mediated decay	-
BRCA2-203	ENST00000528762.1	495	64aa	Nonsense mediated decay	-
BRCA2-207	ENST00000614259.1	7950	No protein	Processed transcript	-
BRCA2-204	ENST0000053093.6	2011	No protein	Processed transcript	-
BRCA2-205	ENST00000533776.1	523	No protein	Retained intron	-

Summary

Section 3

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:

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 ${accs[i]};
    gzip ${accs[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)

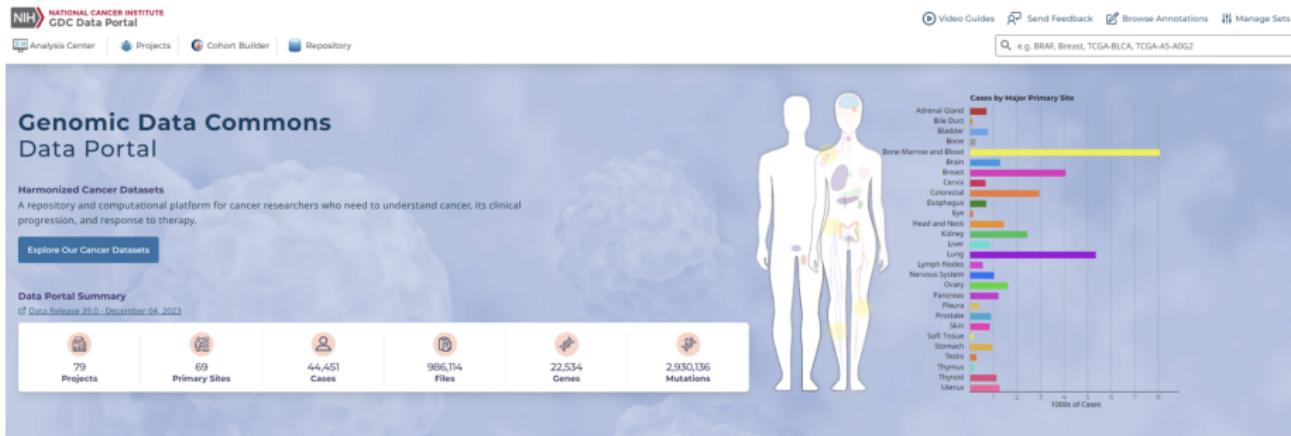
# Extract expression data
expression_data <- exprs(geo_search[[1]])

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

Section 4

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 with a dark blue header and a light blue footer. The main content area has a dark blue background featuring a DNA helix and a brain scan. At the top left, there's a message about the API V2 release. On the right, there's a search bar and navigation links for 'About GTEX', 'Publications', 'Access Biospecimens', 'FAQs', and 'Contact'.

Current Release (V8)

- Tissue & Sample Statistics
- Tissue & Sample Data (Excel spreadsheet)
- Access & Download Data
- Release History
- How to cite GTEx?

The GTEx Portal is the Expression GTEx project, an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 5444 well-defined tissue sites across nearly 18000 individuals, primarily for molecular assays including RNA, WES, and RNA-seq. Remaining samples are available from the GTEx Database. The GTEx Portal provides user access to data including gene expression, GTEx, and Histology Images.

Developmental GTEx

The Developmental Genome-Tissue Expression (dGTE) Project is a new effort to study development-specific genetic effects on gene expression. The goal of the project is to establish a molecular and data analysis resource as well as a tissue bank to study the regulation of gene expression in healthy rodent, healthy reference, neonatal, pediatric, and adolescent tissues, building on the Ontogeny of Human Expression (OEHU) project.

Explore GTEX

Category	Tool	Description
Gene	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
Tissue	Histology Viewer	Browse and search all data by tissue
		Browse and search GTEx histology images
Single Cell	Data Overview	Learn more about available single cell data
	Multi-Gene Single Cell Query	Browse and search single cell expression by gene and tissue
Expression	Multi-Gene Query	Browse and search expression by gene and tissue
	Transcript Browser	Visualize transcript expression and isoform structures

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

DNA regulatory elements

ENCODE contains a host of information about DNA regulatory elements

The screenshot displays the ENCODE project's website interface. At the top, there is a navigation bar with links for ENCODE, Data, Encyclopedia, Materials & Methods, Help, and a shopping cart icon. Below the navigation bar is a horizontal menu bar with three main categories: "Functional genomics" (blue), "Functional characterization" (orange), and "Encyclopedia of elements" (green). The "Functional genomics" section contains the following items:

- Ruch Alzheimer's
- Protein knockdown (Draplin)
- ENCORE
- Immune cells
- Functional genomic series
- Region search

The "Functional characterization" section contains the following items:

- EN-TEX
- Computational and integrative products
- Stem cell differentiation
- Mouse development
- Single-cell experiments
- Encyclopedia browser

The "Encyclopedia of elements" section contains the following items:

- Deeply profiled cell lines
- Human donors
- Imputed experiments
- Reference epigenome
- RNA-seq
- ChIP-seq experiments

<https://www.encodeproject.org>

Section 5

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

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 >>

BRAC_{partial} [Poeciliopsis prolifica]
106 aa protein
Accession: JAO55668.1 GI: 958322777
[BioProject](#) [Nucleotide](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

BRAC [Pseudomonas putida BIRD-1]
371 aa protein
Accession: ADR58867.1 GI: 313497501
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#) [Related Sequences](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

branched-chain amino acid transport protein BraC [Pseudomonas aeruginosa]

Genpept

Chain A, Structure Of The Btb (tramtrack And Bric A Gigaxonin

PDB: 2PPI_A

[Identical Proteins](#) [FASTA](#) [Graphics](#)
[Go to:](#)

LOCUS 2PPI_A 144 aa linear PRI 26-OCT-26
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: Mndb_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; Cetarrhini; Hominidae; Homo.

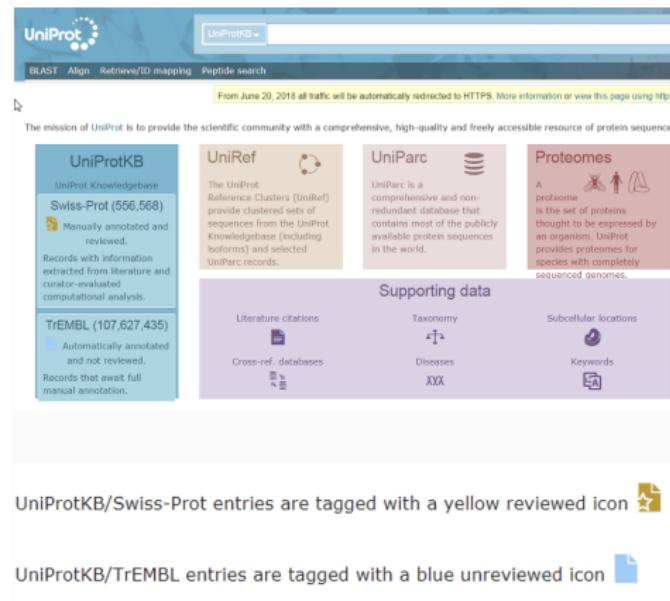
REFERENCE 1 (residues 1 to 144)
AUTHORS Amos,A., Turnbull,A.P., Tickle,J., Keates,T., Bullock,A., S.P., Burgess-Brown,N., Debreczeni,J.E., Ugochukwu,E., Umeano,C., Pike,A.C.W., Papagrigorou,E., Sundstrom,M., Arrowsmith,C.H., Weigelt,J., Edwards,A., Von Delft,F. and Knapp,S.

TITLE Structure Of The Btb (Tramtrack And Bric A Brac) Domain Human Gigaxonin
JOURNAL Unpublished
REFERENCE 2 (residues 1 to 144)
AUTHORS Amos,A., Turnbull,A.P., Tickle,J., Keates,T., Bullock,A., Savitsky,P., N. Brown, Debreczeni,J.E., Ugochukwu,E., Umeano,C., Pike,A.C.W., Papagrigorou,E., Sundstrom,M., Arrowsmith,C.H., Weigelt,J., Edwards,A., Delft, Knapp,S. and Structural Genomics Consortium (Sgc).

//
/region_name="BTB/POZ domain; pfam00651"
/db_xref="CDD:279045"
46..96
/region_name="Domain 2"
/note="NCBI Domains"
49..144
/region_name="BTB"
/note="Broad-Complex, Tramtrack and Bric a brac; smart00225"
/db_xref="CDD:197585"
49..55
/sec_str_type="sheet"
/note="strand 1"
56..62
/sec_str_type="sheet"
/note="strand 2"
63..70
/sec_str_type="helix"
/note="helix 2"
71..79
/sec_str_type="helix"
/note="helix 3"
89..94
/sec_str_type="sheet"
/note="strand 3"
98..109
/sec_str_type="helix"
/note="helix 4"
120..130
/sec_str_type="helix"
/note="helix 5"
133..141
/sec_str_type="helix"
/note="helix 6"
ORIGIN
1 mhhhhhssg vdgtlenly qsmavsdpqh aarlrlaiss freesrfcd ahlvldgeeip
61 vqknilaas pyirtkllyn pkddgdstyk ielegisvmv mreildyifs qgirlnedti
121 qdvvqaadll lltdlktlcc eflc

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 sequences.

UniProtKB
UniProt Knowledgebase
Swiss-Prot (566,568)
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 UniRef Reference Clusters (UniRef100) provide clustered sets of sequences from the UniProt Knowledgebase (including TrEMBL) 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.

Proteomes
A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

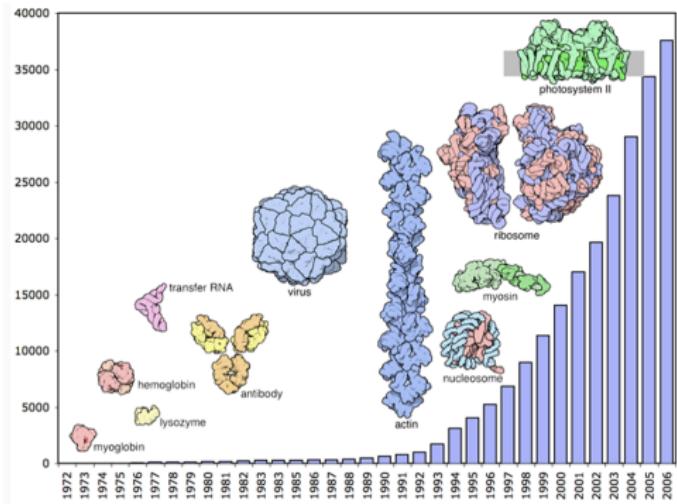
- Literature citations
- Cross-ref. databases
- Taxonomy
- Diseases
- Subcellular locations
- Keywords

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

PDB PROTEIN DATA BANK 149174 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

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

Advanced Search | Browse by Annotations

PDB-101 PDB EMDataResource Protein Data Bank Foundation

Structure Summary 3D View Annotations Sequence Sequence Similarity Structure Similarity Experiment

Biological Assembly 1 5X3E Display Files Download Files

kinase 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 pTRH11*

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

3D View: Structure | Electron Density | Ligand Interaction

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

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A

Biological assembly 1 assigned by authors and generated by PISA (software)

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	7	0.240
Clashscore	0	7
Ramachandran outliers	1.1%	0
Sidechain outliers	13.0%	0
RSRZ outliers	Worse	Better

This is version 1.0 of the entry. See complete history.

Literature Download Primary Citation

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)

139 architectures 3730 sequences 4 Interactions 568 species 103 structures

Summary

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

XIWG39_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:EIT75516.1, ECO:0000313|EnsemblFungi:GGTG_054497D} (1022 residues)

Show all sequences with this architecture.

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

V4B7N4_LOTG1 [Lottia gigantea (Giant owl limpet)] Uncharacterized protein {ECO:0000313|EMBL:ESO84639.1} (791 residues)

Show all sequences with this architecture.

There are 106 sequences with the following architecture: Argol1, PAZ, Piwi

LSKTH8_PTEAL [Pteropus alecto (Black flying fox)] Piwi-like protein 1 {ECO:0000313|EMBL:ELK14246.1} (821 residues)

Show all sequences with this architecture.

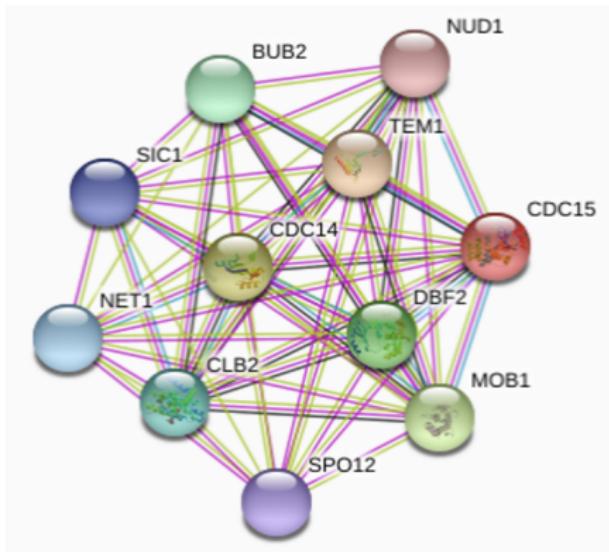
There are 92 sequences with the following architecture: Gly-rich_Ago1, ArgoN, Argol1, PAZ, Argol2, Argomid, Piwi

V4TFX0_9ROSI [Citrus clementina] Uncharacterized protein {ECO:0000313|EMBL:ESR54581.1} (1036 residues)

Show all sequences with this architecture.

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



Section 6

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 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.2.1
##
## 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
##
## 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_4.4.0  fastmap_1.2.0   cli_3.6.2       tools_4.4.0
## [5] htmltools_0.5.8.1 rstudioapi_0.16.0 yaml_2.3.8     rmarkdown_2.27
## [9] knitr_1.47     xfun_0.44      digest_0.6.35   rlang_1.1.4
## [13] evaluate_0.23
```