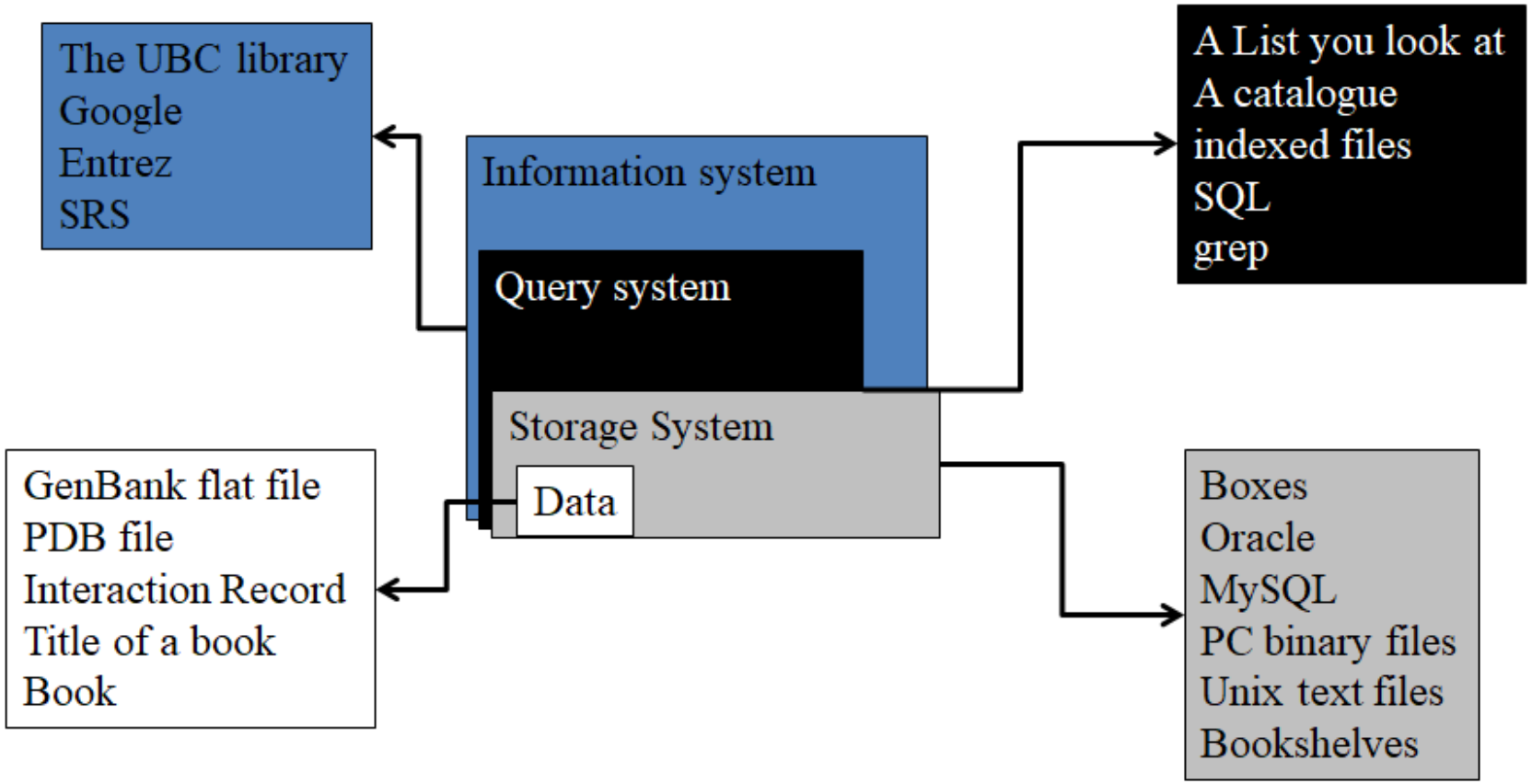
Biological Databases

Outline of discussion

- **Databases and its features**
- **Significance of databases in dealing with large amounts of data**
- **Types of databases**
- **Data integration**
- **Extraction of data from online biological databases**
- **Major online biological databases and their features**

What are databases?

- **A database is an organized collection of structured information, or data, typically stored electronically in a computer system**.
- It consists of basic units called records or entries.
- For example, a protein database would have protein entries as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence).



Databank vs Database vs Data warehouse?

Types of Database Management System

A database is usually controlled by a database management system (DBMS). Together, the data and the DBMS, along with the applications that are associated with them, are referred to as a database system or just database.

Creating a database means you can remember the rules of the database rather than the locations of individual files and so find your way around more easily.

There are majorly three types of database management systems viz. (1) flat file indexing systems, (2) relational DBMS and (3) object-oriented DBMS.

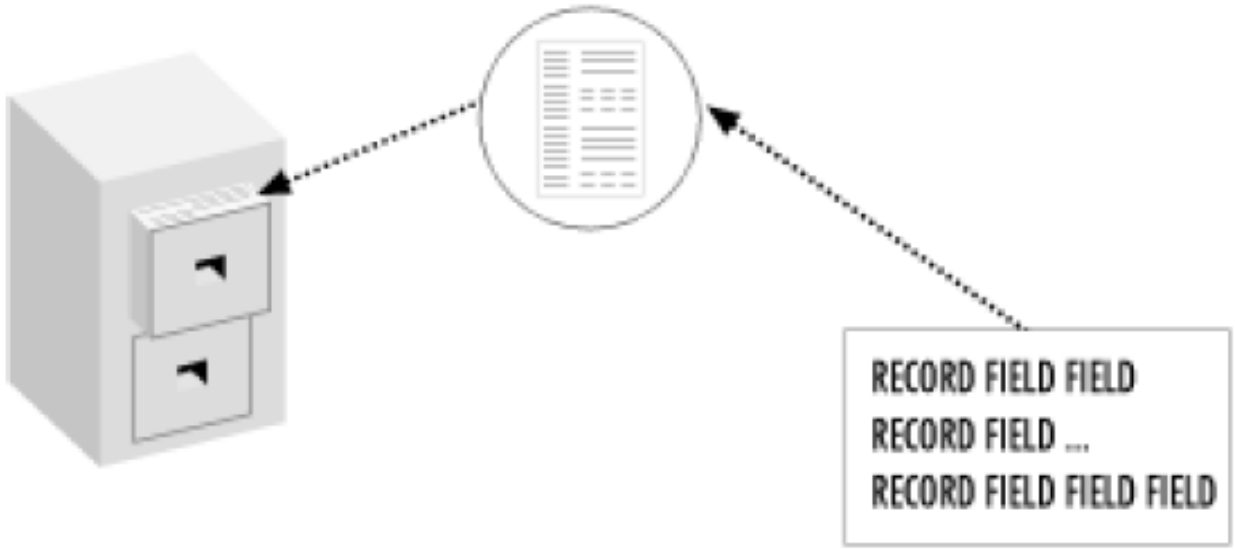
1. Flat File Databases

A flat file database is not truly a database, it is simply an ordered collection of similar files.

Aarij M Hussain	210106001	B.Tech. (BSBE)	BSBE	Bioinformatics
Abhinav P Singh	210106002	B.Tech. (BSBE)	BSBE	Biophysics
----------	----------	----------	------	----------
----------	----------	----------	------	----------
----------	----------	----------	------	----------
Yashraj Verma	210106085	B.Tech. (BSBE)	BSBE	Immunology

Flat File Databases

The emphasis in **formatting data** for a flat file database is **at the character level** i.e. at the level of how the data would appear if it were printed on a page.



Flat file databases are made useful by **ordering and indexing**. A collection of flat files on a computer file system can be ordered and stored in labeled folders exactly the same way as a collection of printed papers are ordered in a file cabinet Drawer.

The relationship of a flat file to a flat-file database

Flat file databases are often made **searchable by indexing**. An index pulls out a particular attribute from a file and pairs the attribute value in **the index with a filename and location**. It is analogous to a book index, which for example tells you where in a book you will find the word "genome.“

Flat file databases in biology

The PDB began by using flat files in the well-known PDB format. The PDB now uses an object-oriented database backend to support database queries and file access.

Beyond the PDB, flat-file databases are still widely used by biologists. Many users of biological sequence data store and access sequences locally using the **Sequence Retrieval System (SRS)**, a flat file indexing system designed with biological data in mind.

Drawbacks of flat file database management systems

However, as flat file collections grow larger and larger, working with them becomes inefficient. An index is one-dimensional, so it is difficult (though not impossible) to make connections between attributes within an indexed flat file database.

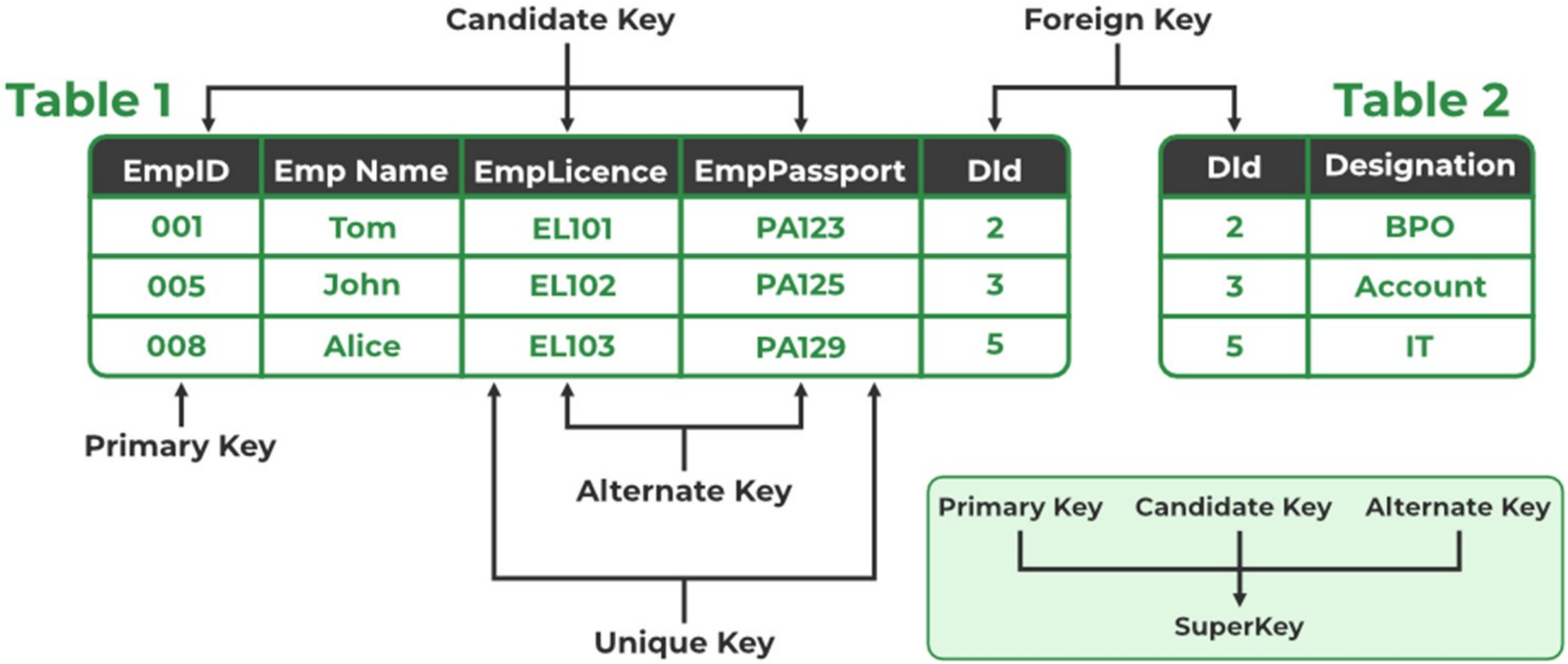
Types of Database Management System

2. Relational Databases: In a relational database, the information is stored in a collection of tables.

Roll No.	Name	Degree	Department	Course taken
210106001	Aarij Modhubhai Hussain	B.Tech. (BSBE)	BSBE	Bioinformatics
210106002	Abhinav Pratap Singh	B.Tech. (BSBE)	BSBE	Biophysics
----------	----------	----------	----------	----------
----------	----------	----------	----------	----------
210106085	Yashraj Verma	B.Tech. (BSBE)	BSBE	Immunology

The form of the tables follows rules that are uniform across the database, so you can access all the tables about student details.

Relational Databases

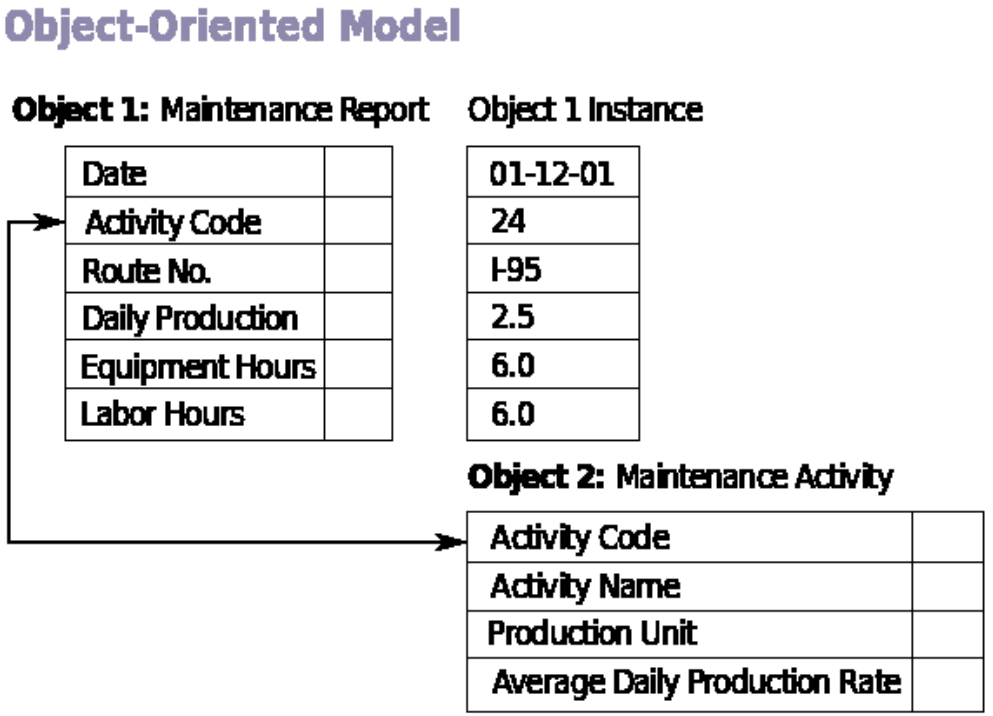


- **Advantages:** Manageability, Flexibility, Avoid errors. **Challenges:** Scalability (single server), Performance, Relationships.

Types of Database Management System

3. Object-Oriented DBMS

An OODBMS is a database management system in which information is represented in the form of objects as used in object-oriented programming.

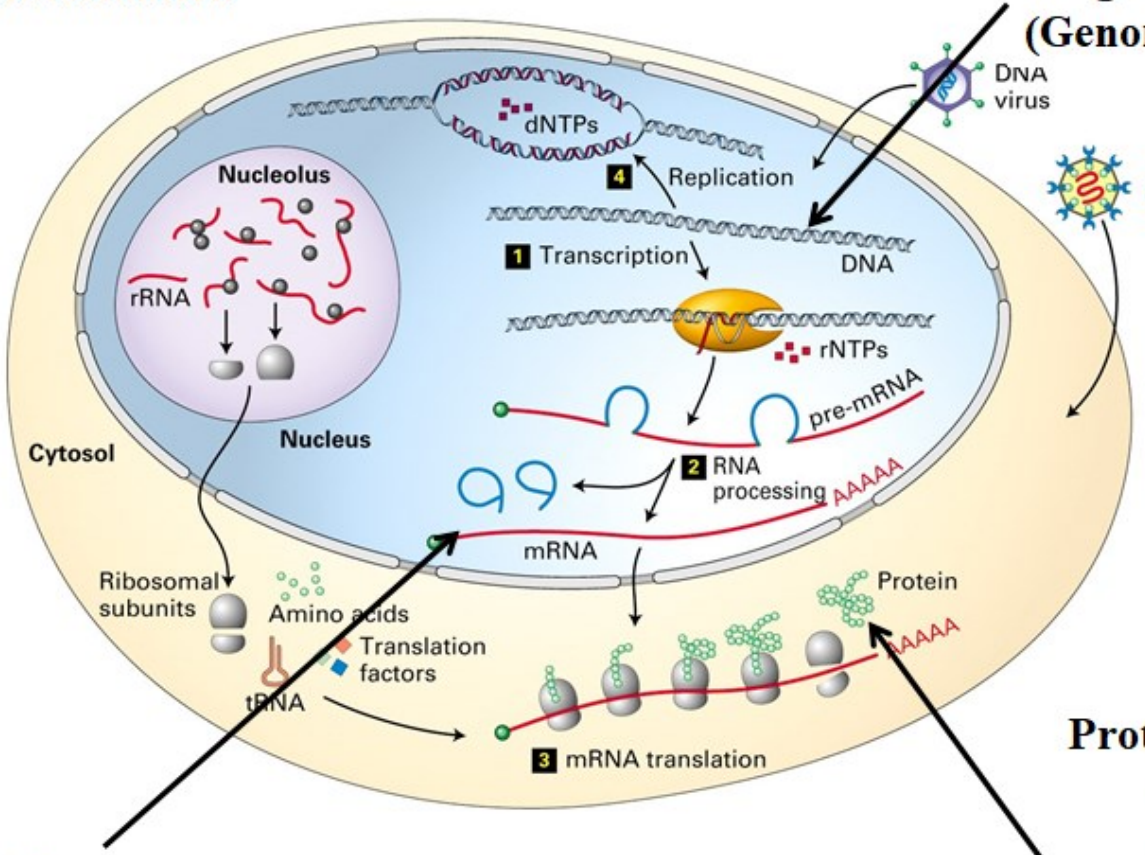


- **Advantages:** Faster access and performance. **Challenges:** Versatility, No standard query language, difficult to learn.

Biological databases: data

Literature information

Original DNA Sequences (Genomes)



This is only for a cell type, there are millions of such types.

Expressed DNA sequences
(= mRNA, = cDNA sequences)
Expressed Sequence Tags (ESTs)

Protein Sequences
- -Inferred
- -Direct sequencing

Protein structures
- -Experiments
- -Models

How many types of cells in humans?

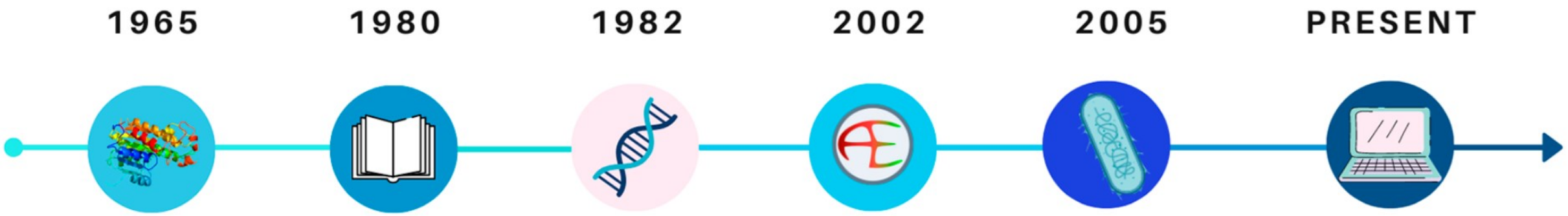
How many cells in humans?

How many genome sequences are available?

How many species are there on the Earth?

In fact, the scientific community has now generated data beyond the exabyte (10^{18}) level.

A brief history of biological databases



1965 — Margaret Dayhoff created the Atlas of Protein Sequence and Structure

1980 — The EMBL Data Library created a catalog of biological data

1982 — GenBank became a public database.

2002 — ArrayExpress released

2005 — The first high-throughput sequencing systems was used to sequence a full *E. coli* genome

PRESENT — Many directories for multi-omic data were created. Let's explore some databases!

Types of biological databases

- **Primary (archival)**
  - Nucleic acid
    - GenBank
    - EMBL
    - DDBJ
  - Protein
    - Swiss-Prot
    - TrEMBL
    - PIR
  - Structure
    - PDB
    - MMDB
    - ModBase
    - CSD

Curators

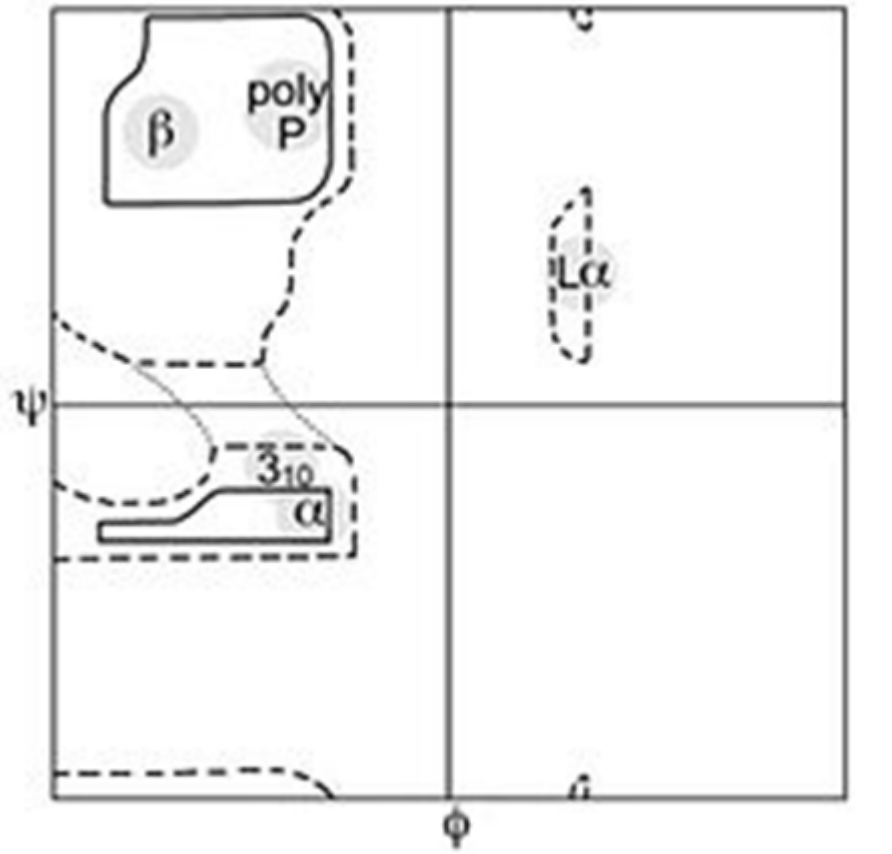
- **Secondary (curated)**
  - Nucleic acid
    - RefSeq
    - FlyBase
    - OMIM
  - Protein
    - Prosite
    - PRINTS
    - Pfam
  - Structure
    - LySDB
    - NDB

Composite databases: the initial data are taken from the primary database, and then they are merged together based on certain conditions. Example: UniProtKB (SwissProt + TrEMBL).

How do you define the ‘Perfect’ database?

- Comprehensive, but easy to search
- Annotated, but not “too annotated”
- A simple, easy to understand its structure
- Cross-referenced
- Minimum redundancy
- Easy retrieval of data
- Accuracy
- Up-to-date
- Good Interface
- Batch search/download options
- Simplify the information space by specialization

Bonus: Allows you to make discoveries



Reasons to search databases

- When obtaining a new DNA sequence, one needs to know whether it has already been deposited in the databanks, or whether they contain any homologous sequences (sequences which are derived from a common ancestry) exist there.
- Given a putative coding ORF, one can search for homologous proteins (similar in their folding or structure of function).
- To find similar non-coding DNA stretches in the database: repeat elements or regulatory sequences, for instance.
- There are other uses for specific purpose, like locating false priming sites for a set of PCR oligonucleotides.

DNA vs Protein searches

Protein sequence should be used for DB similarity search whenever possible because:

- There are very different DNA sequences that code for similar protein sequences.
- When comparing DNA sequences, one gets significantly more random matches than with proteins as (a) DNA is composed of 4 letters hence two unrelated DNA sequences are expected to have 25% similarity, (b) In contrast, proteins are made up of 20 letters, thus sensitivity of the comparison is improved, (c) DNA DBs are much larger and grow faster than protein DBs, implies bigger DB mean more random hits.
- For DNA, usually identity matrices is used while for proteins more sensitive matrices such as PAM or BLOSUM are used, which allows a better search.
- Proteins are rarely mutated during evolution, thus searching them reveals remote evolutionary relationships.

How to perform Database Searching

The amount of biological data is increasing so rapidly that it requires methods to search these information. There are three data retrieval systems in molecular biology (1) Sequence retrieval system (SRS), (2) Entrez and (3) DBGET.

These systems allow text searching of multiple databases and provide links to relevant information for entries that match the search criteria. The three systems differ in the databases they search and the links they have to the other information.

SRS
SRS is a homogeneous interface over 80 DBs developed at EBI, UK. It includes DBs of sequences, metabolic pathways, transcription factors, genomes, mappings, mutations and locus specific mutations.

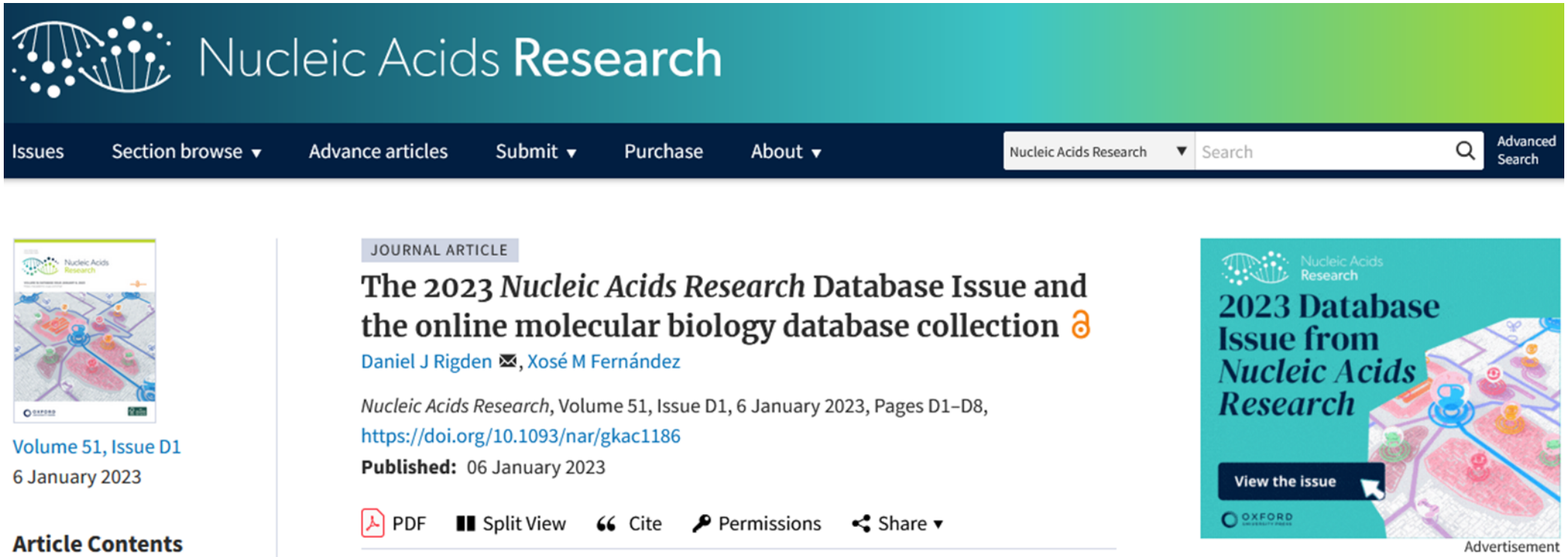
Entrez
Entrez is a DB and retrieval system at NCBI. It is point to explore distinct but integrated DBs. It is easy to use but offers more limited information to search.

DBGET
DBGET is retrieval system developed at Univ. of Tokyo and provide information to 20 DBs, one at a time.

What can be discovered about a gene by a database search?

- A little or a lot, depending on the gene
  - **Evolutionary information**: homologous genes, taxonomic distributions, allele frequencies, synteny, etc.
  - **Genomic information**: chromosomal location, introns, UTRs, regulatory regions, shared domains, etc.
  - **Structural information**: associated protein structures, fold types, structural domains.
  - **Expression information**: expression specific to particular tissues, developmental stages, phenotypes, diseases, etc.
  - **Functional information**: enzymatic/molecular function, pathway/cellular role, localization, role in diseases.

Where to find biological databases?

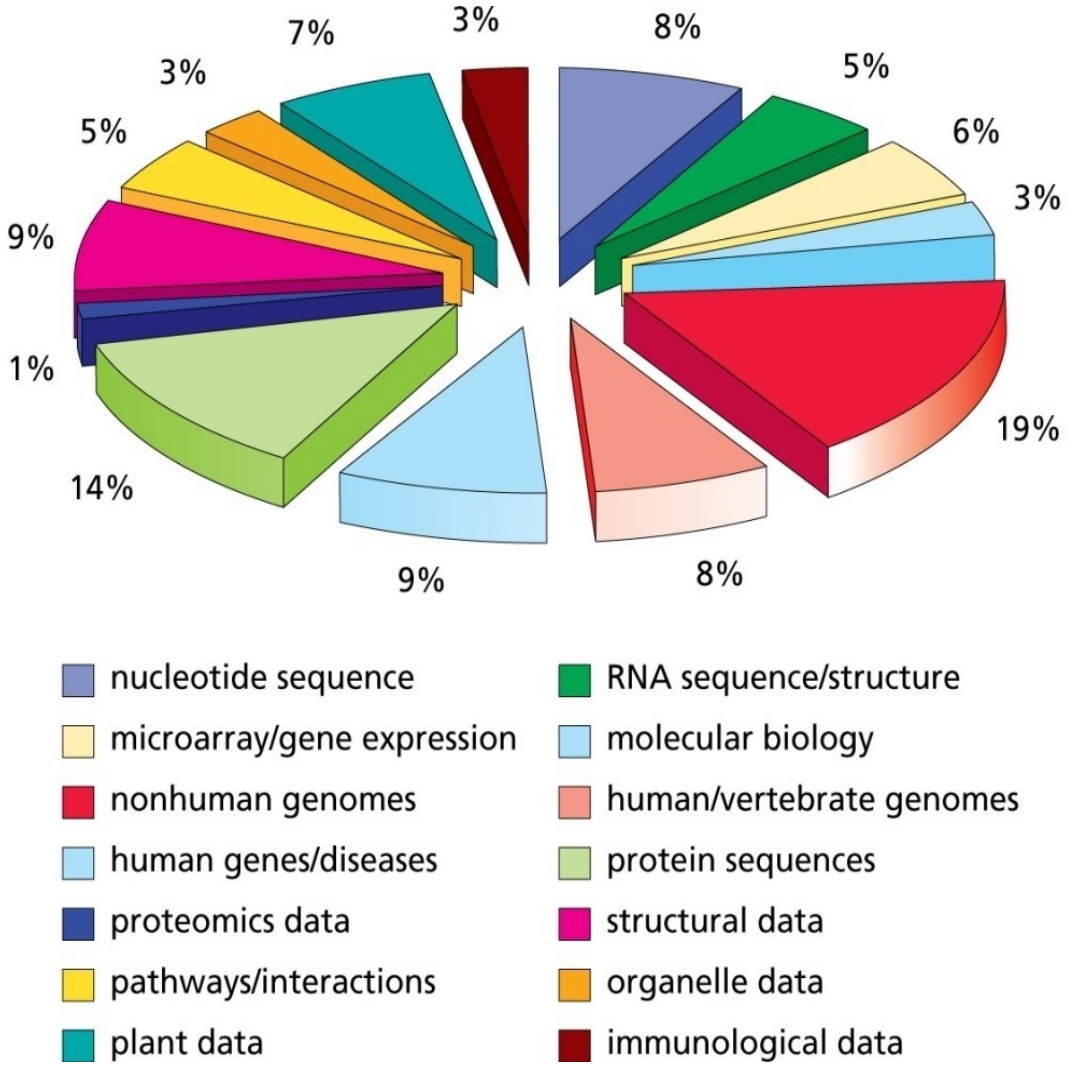


The total number of NAR online databases has been expanded to **~1764** in 2023 (**1645,** 2022).

Where to find biological databases?

These articles are divided into different categories:

1. Nucleotide Sequence Databases
2. RNA sequence databases
3. Protein sequence databases
4. Structure Databases
5. Genomics Databases (non-vertebrate)
6. Metabolic and Signaling Pathways
7. Human and other Vertebrate Genomes
8. Human Genes and Diseases
9. Microarray Data and other Gene Expression DBs
10. Proteomics Resources
11. Other Molecular Biology Databases
12. Organelle databases
13. Plant databases
14. Immunological databases
15. Cell biology

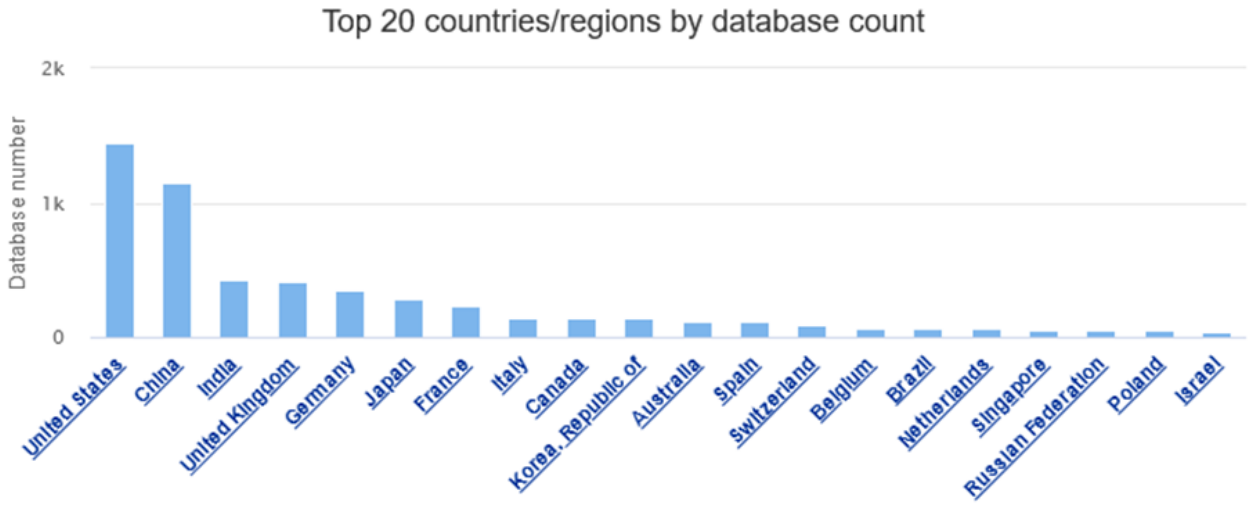


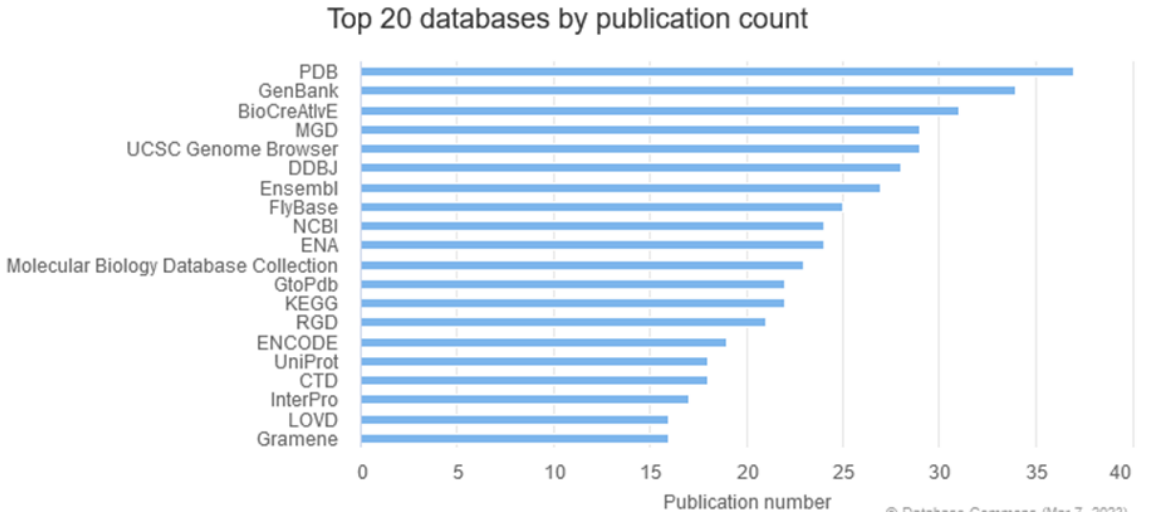
Some interesting online databases

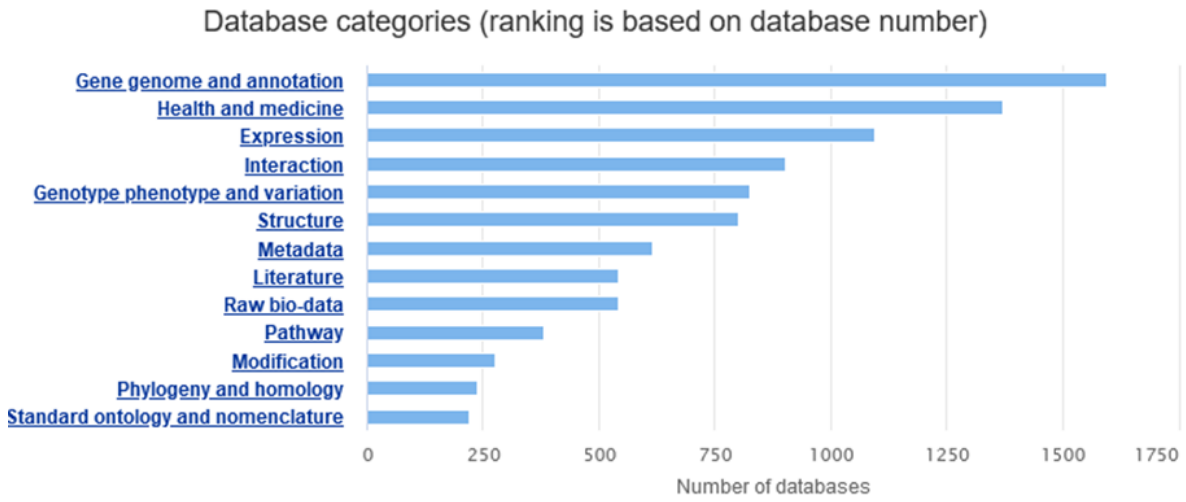
Database name	URL	Brief description
Addgene Vector Database	http://www.addgene.org/vector-database/	Plasmid vectors from publications and commercial sources
AnimalTFDB	http://bioinfo.life.hust.edu.cn/AnimalTFDB/	Contains 125,135 TF genes and 80,060 cofactor genes from 97 animal genomes.
BacFITBase	http://www.tartaglialab.com/bacfitbase/	Bacterial genes relevant to host infection
BBCancer	http://bbcancer.renlab.org	Blood-based biomarkers for cancer
CancerPPD	http://crdd.osdd.net/raghava/cancerppd/	Experimentally validated anticancer peptides
EHFPI	http://biotech.bmi.ac.cn/ehfpi/	Essential Host Factors for Pathogenic Infection
EuRBPDB	http://EuRBPDB.syshospital.org	Eukaryotic RNA binding proteins
MoonProt	http://www.moonlightingproteins.org/	Moonlighting proteins
SuperDRUG2	http://cheminfo.charite.de/superdrug2/	A one stop resource for approved/ marketed drugs

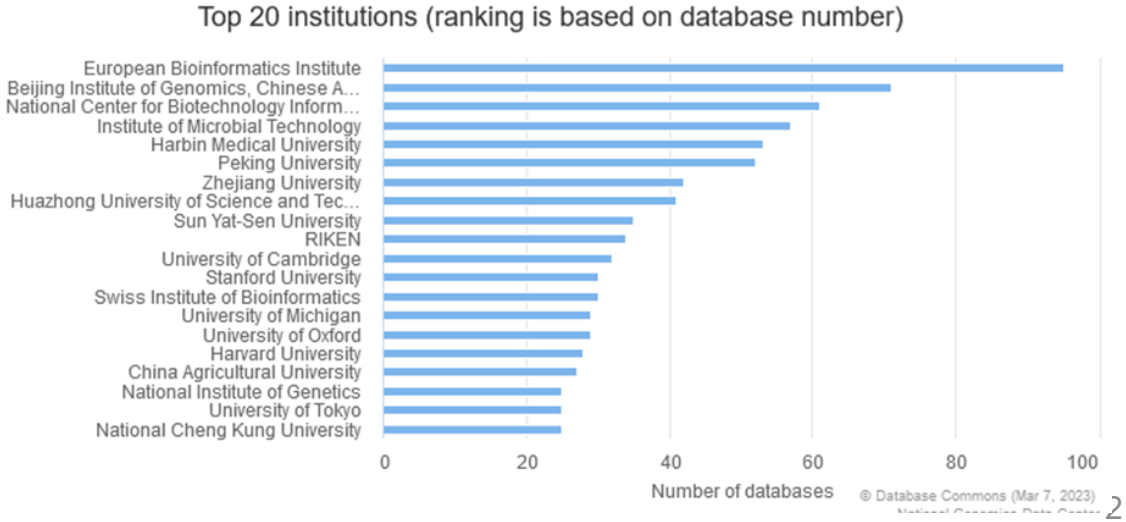
Where to find biological databases?

Database Commons: a catalog of worldwide biological databases: No. of databases: 5910 (2023), No. of publications from: 9023 (2023), Categorized into: 13, No. of species: 1528, No. of countries: 72, No. of institutions: 1997. Highly cited databases: David, KEGG STRING, Pfam, ENCODE.









Where to find biological databases?

Pathguide: It contains information about **702** biological pathway related resources and molecular interaction related resources.

Bio.Tools: It strives to provide a comprehensive registry of software and databases, facilitating researchers from across the spectrum of biological and biomedical science to find, understand, utilize and cite the resources they need in their day-to-day work.

OMICtools: Leveraging life science data to reveal exciting new insights.

Metabases

BioGraph – A knowledge discovery service based on the integration of 21 heterogeneous databases.
Bioinformatic Harvester – Integrating 26 major protein/gene resources.
ConsensusPathDB – A molecular functional interaction database (contains 215541 unique interactions, 4601 pathways from overall 30 dbs).
Enzyme Portal – Integrates enzyme information such as small-molecule chemistry, biochemical pathways and drug compounds.
euGenes – (Genomic information of eukaryotic organisms).
GeneCards – (searchable integrated information of human genes).
mGen – Integrates GenBank, Refseq, EMBL and DDBJ.
PathogenPortal – A repository linking to the Bioinformatics Resource Centers sponsored by the National Institute of Allergy and Infectious Diseases (NIAID).
SOURCE – a unification tool which dynamically collects and compiles data from many scientific databases, and thereby attempts to encapsulate the genetics and molecular biology of genes from the genomes of *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* into easy to navigate GeneReports.
Big Data Centre - Resources with significant updates in the past year include BioProject (a biological project library), BioSample (a biological sample library), Genome Sequence Archive (GSA, a data repository for archiving raw sequence reads), Genome Warehouse (GWH, a centralized resource housing genome-scale data), Genome Variation Map (GVM, a public repository of genome variations), Science Wikis (a catalog of biological knowledge wikis for community annotations) and IC4R (Information Commons for Rice).

Some specific biological databases

GenBank

The first DNA sequence database, established in 1979, was the Gene Sequence Database (GSDB), now known as GenBank.

GenBank (https://www.ncbi.nlm.nih.gov/genbank/) is a comprehensive, public database that contains 19.6 trillion base pairs from over 2.9 billion nucleotide sequences for 5,04,000 formally described species.

NCBI, in cooperation with EMBL and other international organizations, provides the most complete collection of DNA sequence data in the world, as well as PubMed, a taxonomy database, and an alternate access point for protein sequence and structure data.

Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage.

Growth of GenBank

Genetic Sequence Data Bank (June 15, 2023)

NCBI-GenBank Flat File Release 254.0

Sequences: 2,43,560,863

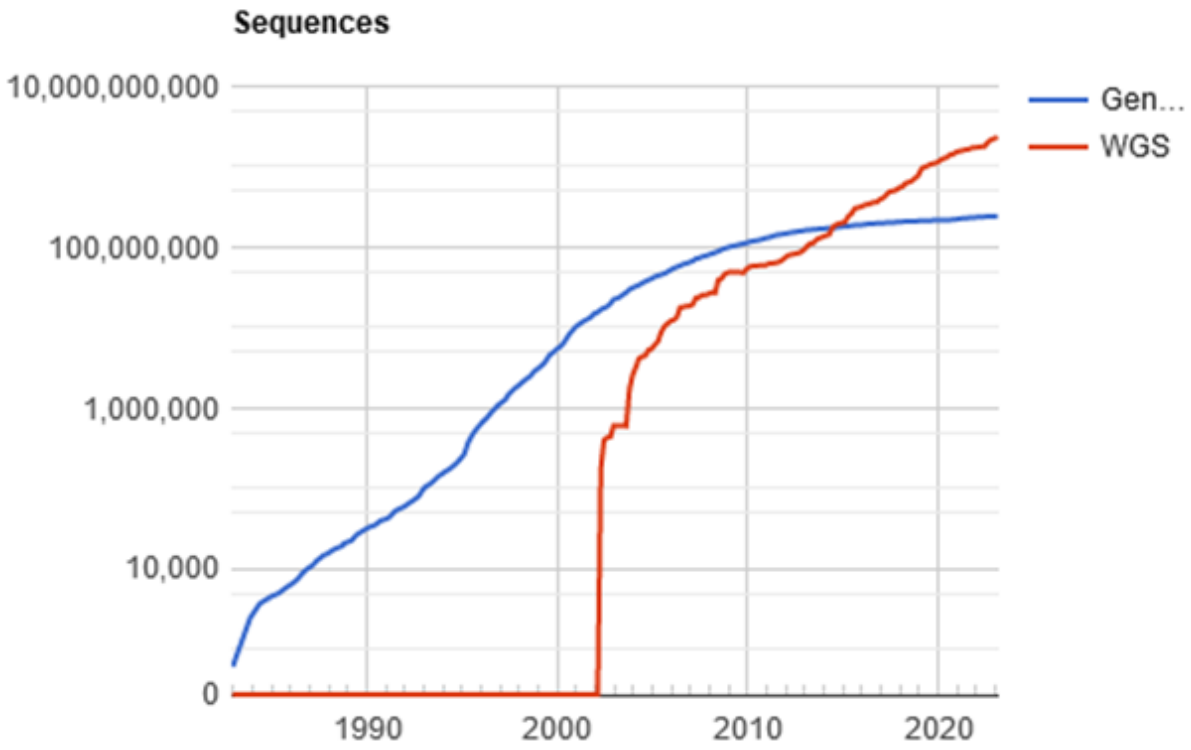


Table: Growth of GenBank Division

Division	Description	2022	2023	Increase	% Increase
BCT	Bacteria	130518385589	166217792419	35699406830.00	27.35
ENV	Environmentalsamples	7394414660	8516518905	1122104245.00	15.18
EST	Expressedsequencetags	43324455796	43330114068	5658272.00	0.01
GSS	Genomesurveysequences	26380049011	26380049011	0.00	0.00
HTC	High-throughputcDNA	737423641	740853492	3429851.00	0.47
HTG	High-throughputgenomic	27800219072	27801878633	1659561.00	0.01
INV	Invertebrates	108680334593	269338221858	160657887265.00	147.83
MAM	Othermammals	28568850588	41720029494	13151178906.00	46.03
PAT	Patentsequences	29588418021	30938105095	1349687074.00	4.56
PHG	Phages	935884237	1158493277	222609040.00	23.79
PLN	Plants	350590744188	484803006831	134212262643.00	38.28
PRI	Primates	15165437356	15619743253	454305897.00	3.00
ROD	Rodents	23336550435	66092410483	42755860048.00	183.21
STS	Sequencetaggedsites	640923137	640923137	0.00	0.00
SYN	Synthetic	7994601379	8030787249	36185870.00	0.45
TLS	Targetedlocistudies	39930167315	43852280645	3922113330.00	9.82
TSA	Transcriptomeshotgundata	454757992932	511680950707	56922957775.00	12.52
UNA	Unannotated	4421782	4436341	14559.00	0.33
VRL	Viruses	39351597469	187366647663	148015050194.00	376.13
VRT	Othervertebrates	85320979451	99921122967	14600143516.00	17.11
WGS	Wholegenomeshotgundata	13888187863722	17511809676629	3623621812907.00	26.09

GenBank sequences are organized into 21 divisions, each of which is represented by a three-letter abbreviation.

Sayers et al., 2022, Nucleic Acids Res.

Submission of data to GenBank

Direct submissions from scientists. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries.

BankIt: a WWW-based submission tool with wizards to guide the submission process.

Sequin: NCBI's stand-alone submission tool with wizards to guide the submission process.

Tbl2asn: a command-line program, automates the creation of sequence records for submission to GenBank using many of the same functions as Sequin. It is used primarily for submission of complete genomes and large batches of sequences.

Receiving an Accession Number for your Manuscript
GenBank will provide accession numbers for submitted sequences, usually within two working days.

GenBank submission types

Standard: GenBank accepts mRNA or genomic sequence data directly determined by the submitter.

EST, STS (Sequence-tagged site), GSS (Genome Survey Sequence), HTG (High-Throughput Genomic) Sequences, Complete Microbial Genomes, WGS (Whole Genome Shotgun) Sequences, TSA (Transcriptome Shotgun Assembly) Sequences, TPA (Third Party Annotation)

The following data is not accepted by GenBank

Non-contiguous sequences, Primer sequences, Protein sequences with no underlying nucleotide submission, Sequence containing a mix of genomic and mRNA sequence, Sequences without a physical counterpart (consensus sequences), Sequences with length less than 200 nucleotides.

A sample of GenBank record – abstract syntax notation

```
LOCUS       eIF4E                   2881 bp    DNA     linear   INV 27-OCT-2005
DEFINITION  Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E)
            gene, alternative splice products, complete cds.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Drosophila melanogaster (fruit fly)
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila.
REFERENCE   1  (bases 1 to 2881)
  AUTHORS   Burnett,F.M., van der Waals,J.D. and Szent-Gyorgi,A.
  TITLE     Environmental influences on the expansion of germline tandem
            repeats in several species of Galapagos finches
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 2881)
  AUTHORS   Burnett,F.M., van der Waals,J.D. and Szent-Gyorgi,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-OCT-2005) Evolutionary Biology Department, Oxbridge
            University, 1859 Tennis Court Lane, Camford OX1 2BH, United Kingdom
FEATURES             Location/Qualifiers
     source          1..2881
                     /organism="Drosophila melanogaster"
                     /mol_type="genomic DNA"
                     /strain="Oregon R"
     gene            join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
                     /gene="eIF4E"
     CDS             join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
                     /gene="eIF4E"
                     /codon_start=1
                     /product="eukaryotic initiation factor 4E-II"
                     /translation="MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETG
                     EPAGNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTV
                     EDFWSLYNHIKPPSEIKLGSDYSLFKKNIRPMWEDAANKQGGRWVITLNKSSKTDLDN
                     LWLDVLLCLIGEAFDHSDQICGAVINIRGKSNKISIWTADGNNEEAALEIGHKLRDAL
                     RLGRNNSLQYQLHKDTMVKQGSNVKSIYTL"
```

Access to GenBank

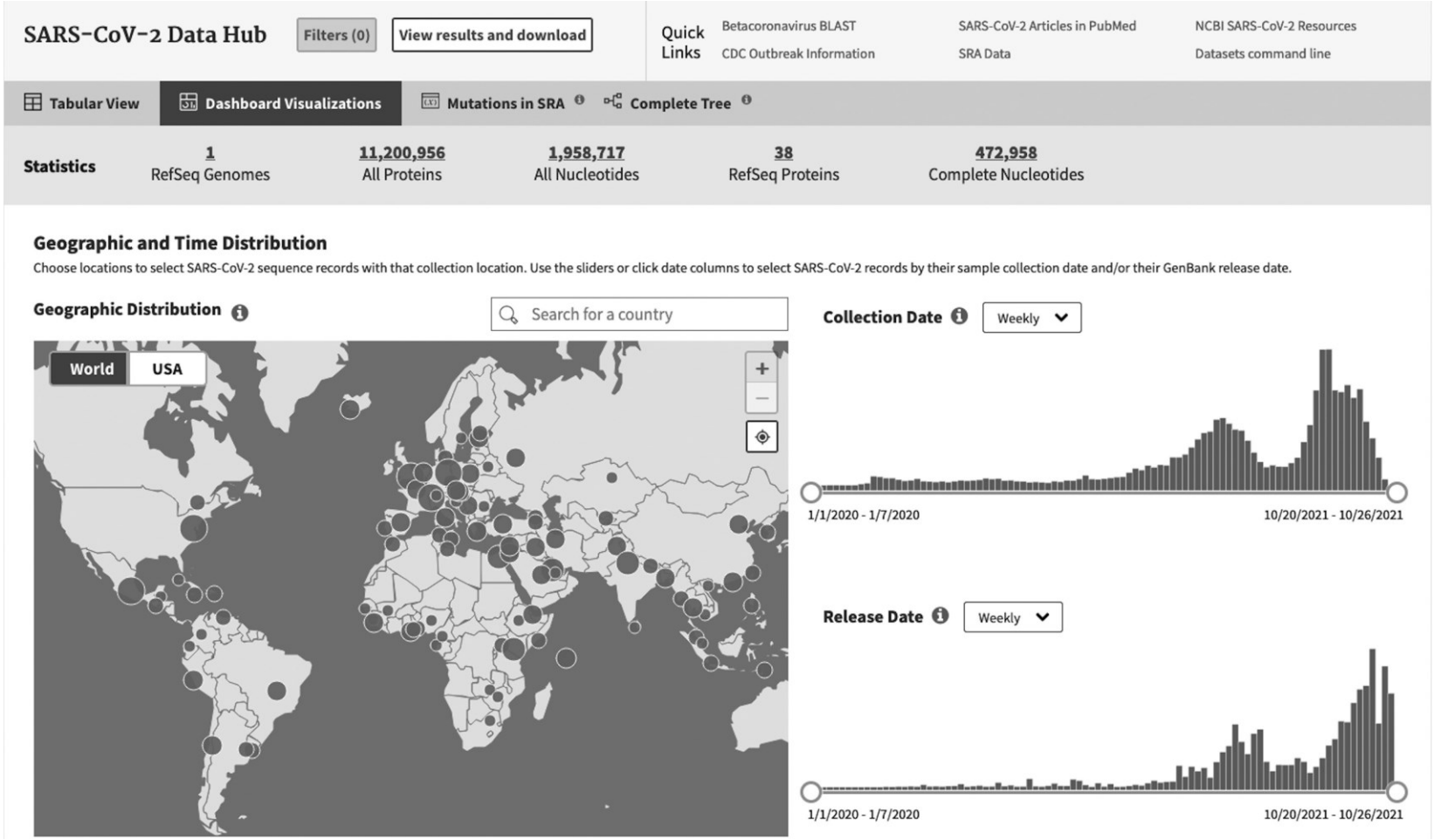
- Search GenBank for sequence identifiers and annotations with **Entrez** Nucleotide, which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags) and dbGSS (Genome Survey Sequences).

- Search and align GenBank sequences to a query sequence using BLAST.

- Search, link and download sequences programmatically using NCBI e-utilities.

SARS-CoV-2 resources

NCBI Datasets provides downloads for over 1.5 million complete SARS-CoV-2 genomes (https://www.ncbi.nlm.nih.gov/datasets/coronavirus/genomes) as well as a new taxonomy page for this virus (https://www.ncbi.nlm.nih.gov/data-hub/taxonomy/2697049/).

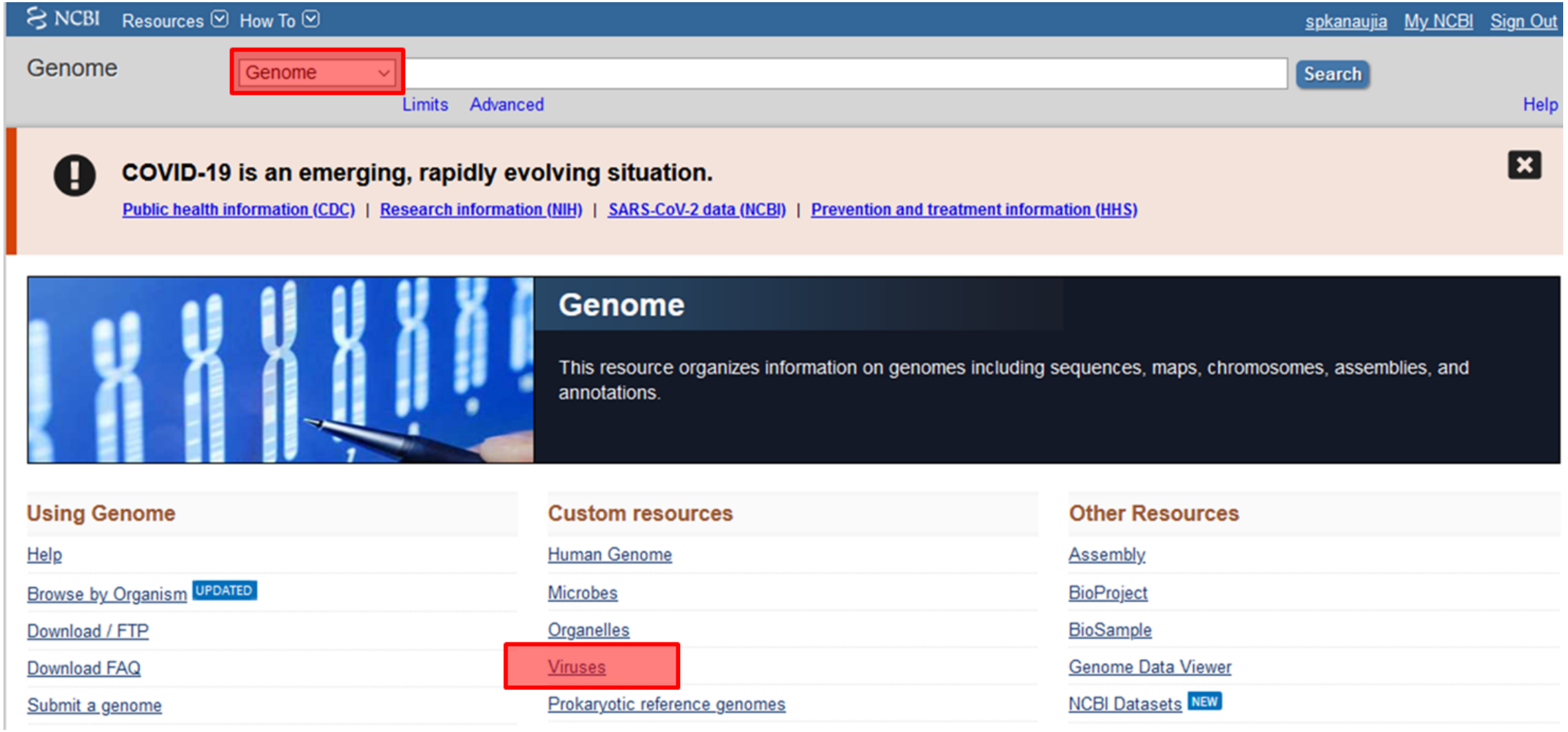


SARS-CoV-2 Data Hub in the NCBI Virus resource.

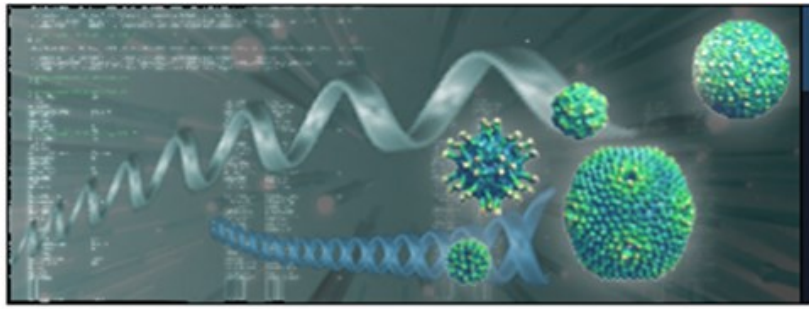
Sayers et al., 2023, Nucleic Acids Res.

An interesting observation

An interesting observation



An interesting observation



Viral Genomes

This resource provides viral genome sequence data and related information.

Explore Viral Genome Sequences

- Viral genome browser
- Browse viral genomes by family
- View all RefSeq and Neighbor nucleotide records

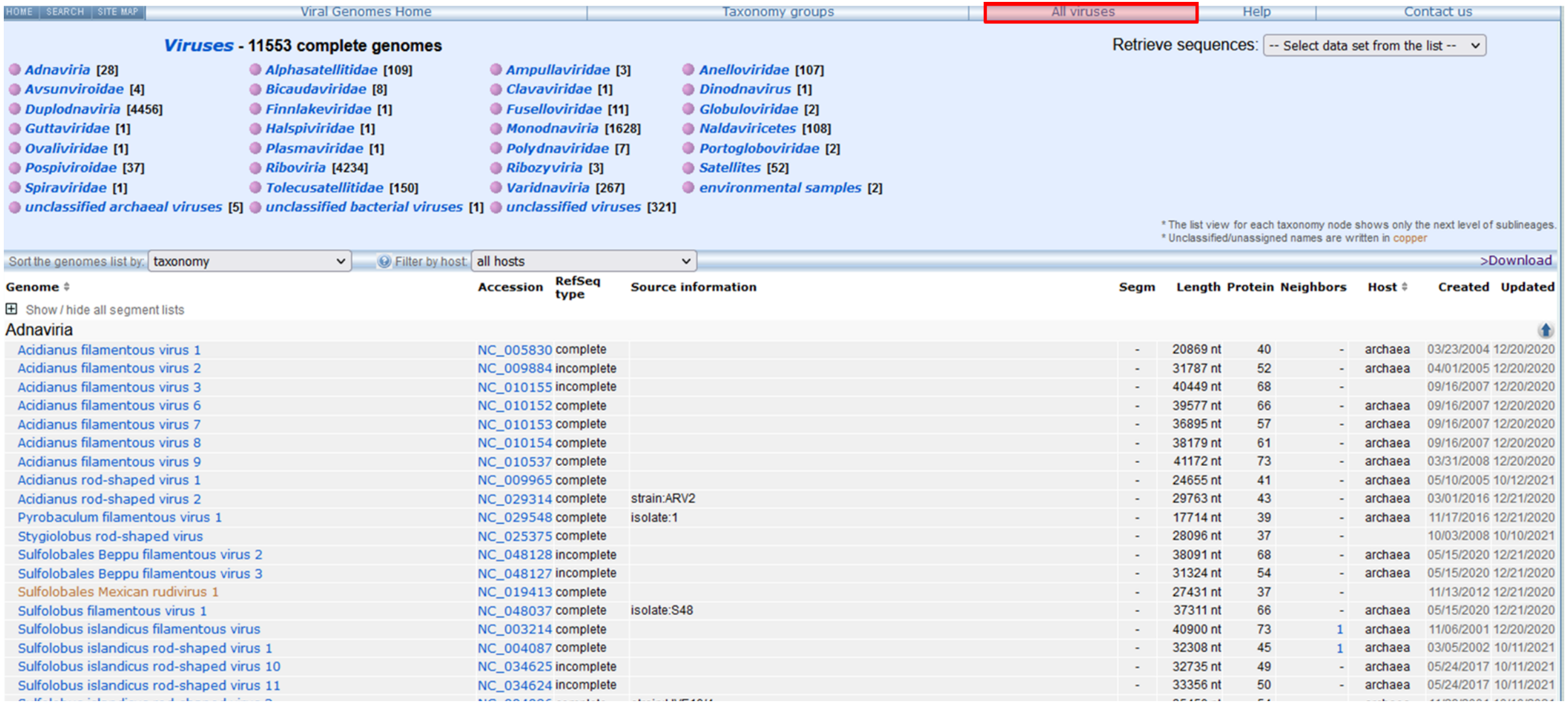
Resource Tools

- Retrovirus Resource
- Virus Variation Resource
- Pairwise Sequence Comparison Tool (PASC)
- Protein Clusters
- Viral Genotyping Tool

Virus Variation Resource

- Influenza virus
- Dengue virus
- West Nile virus
- MERS coronavirus
- Ebolavirus
- Rotavirus
- Zika virus

An interesting observation



HOME | SEARCH | SITE MAP | Viral Genomes Home | Taxonomy groups | All viruses | Help | Contact us

Viruses - 11553 complete genomes

Retrieve sequences: -- Select data set from the list --

- Adnaviria [28]
- Alphasatellitidae [109]
- Ampullaviridae [3]
- Anelloviridae [107]
- Avsunviroidae [4]
- Bicaudaviridae [8]
- Clavaviridae [1]
- Dinodnavirus [1]
- Duplodnaviria [4456]
- Finnlakeviridae [1]
- Fuselloviridae [11]
- Globuloviridae [2]
- Guttaviridae [1]
- Halspiviridae [1]
- Monodnaviria [1628]
- Naldaviricetes [108]
- Ovaliviridae [1]
- Plasmaviridae [1]
- Polydnaviridae [7]
- Portogloboviridae [2]
- Pospiviroidae [37]
- Riboviria [4234]
- Ribozyviria [3]
- Satellites [52]
- Spiraviridae [1]
- Tolecusatellitidae [150]
- Varidnaviria [267]
- environmental samples [2]
- unclassified archaeal viruses [5]
- unclassified bacterial viruses [1]
- unclassified viruses [321]

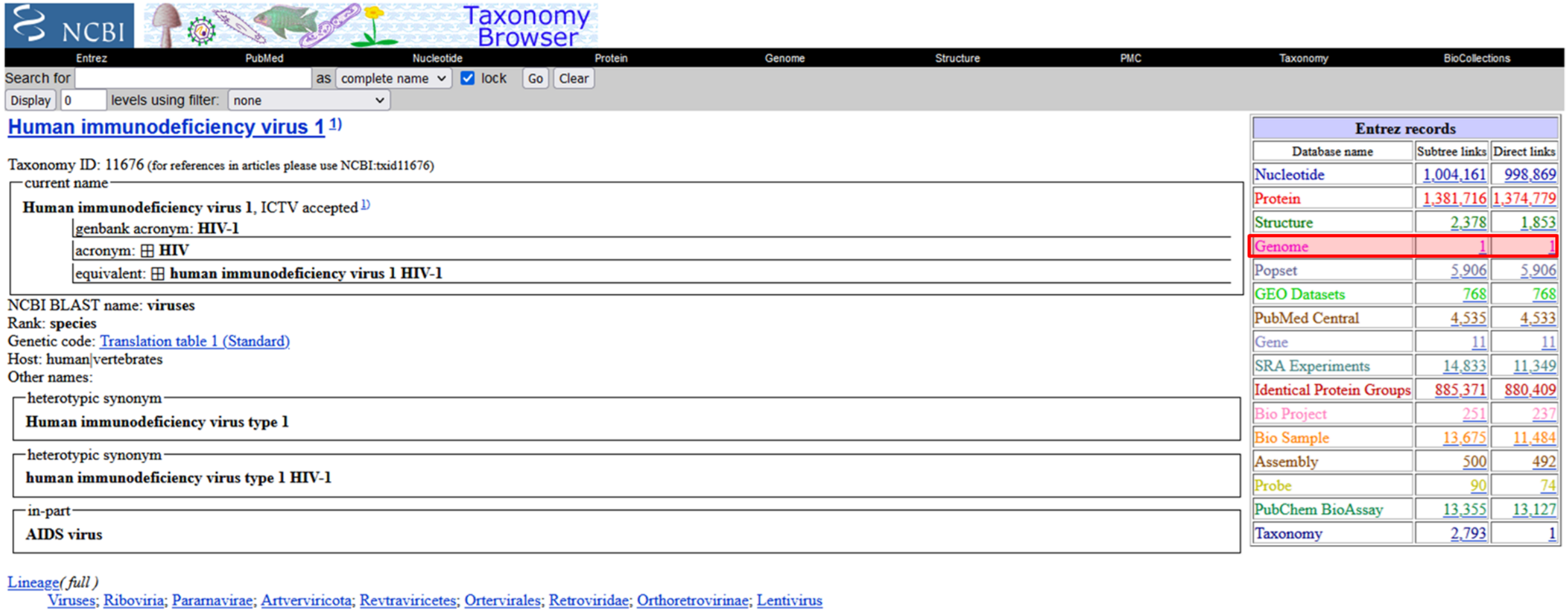
* The list view for each taxonomy node shows only the next level of sublineages.
* Unclassified/unassigned names are written in copper

Sort the genomes list by: taxonomy | Filter by host: all hosts | >Download

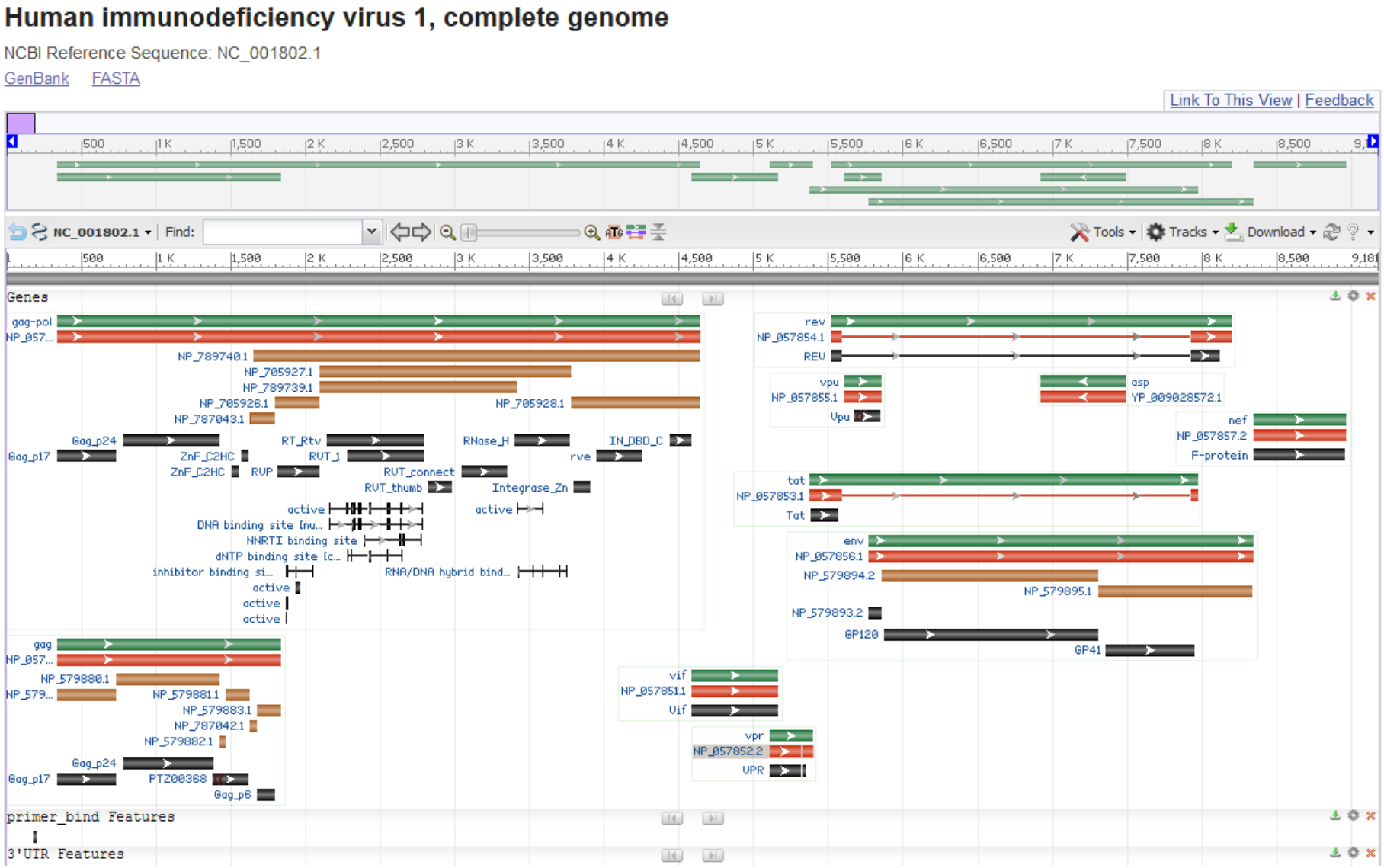
Show / hide all segment lists

Genome	Accession	RefSeq type	Source information	Segm	Length	Protein	Neighbors	Host	Created	Updated
Adnaviria										
Acidianus filamentous virus 1	NC_005830	complete		-	20869 nt	40	-	archaea	03/23/2004	12/20/2020
Acidianus filamentous virus 2	NC_009884	incomplete		-	31787 nt	52	-	archaea	04/01/2005	12/20/2020
Acidianus filamentous virus 3	NC_010155	incomplete		-	40449 nt	68	-		09/16/2007	12/20/2020
Acidianus filamentous virus 6	NC_010152	complete		-	39577 nt	66	-	archaea	09/16/2007	12/20/2020
Acidianus filamentous virus 7	NC_010153	complete		-	36895 nt	57	-	archaea	09/16/2007	12/20/2020
Acidianus filamentous virus 8	NC_010154	complete		-	38179 nt	61	-	archaea	09/16/2007	12/20/2020
Acidianus filamentous virus 9	NC_010537	complete		-	41172 nt	73	-	archaea	03/31/2008	12/20/2020
Acidianus rod-shaped virus 1	NC_009965	complete		-	24655 nt	41	-	archaea	05/10/2005	10/12/2021
Acidianus rod-shaped virus 2	NC_029314	complete	strain:ARV2	-	29763 nt	43	-	archaea	03/01/2016	12/21/2020
Pyrobaculum filamentous virus 1	NC_029548	complete	isolate:1	-	17714 nt	39	-	archaea	11/17/2016	12/21/2020
Stygiolobus rod-shaped virus	NC_025375	complete		-	28096 nt	37	-		10/03/2008	10/10/2021
Sulfolobales Beppu filamentous virus 2	NC_048128	incomplete		-	38091 nt	68	-	archaea	05/15/2020	12/21/2020
Sulfolobales Beppu filamentous virus 3	NC_048127	incomplete		-	31324 nt	54	-	archaea	05/15/2020	12/21/2020
Sulfolobales Mexican rudivirus 1	NC_019413	complete		-	27431 nt	37	-		11/13/2012	12/21/2020
Sulfolobus filamentous virus 1	NC_048037	complete	isolate:S48	-	37311 nt	66	-	archaea	05/15/2020	12/21/2020
Sulfolobus islandicus filamentous virus	NC_003214	complete		-	40900 nt	73	1	archaea	11/06/2001	12/20/2020
Sulfolobus islandicus rod-shaped virus 1	NC_004087	complete		-	32308 nt	45	1	archaea	03/05/2002	10/11/2021
Sulfolobus islandicus rod-shaped virus 10	NC_034625	incomplete		-	32735 nt	49	-	archaea	05/24/2017	10/11/2021
Sulfolobus islandicus rod-shaped virus 11	NC_034624	incomplete		-	33356 nt	50	-	archaea	05/24/2017	10/11/2021

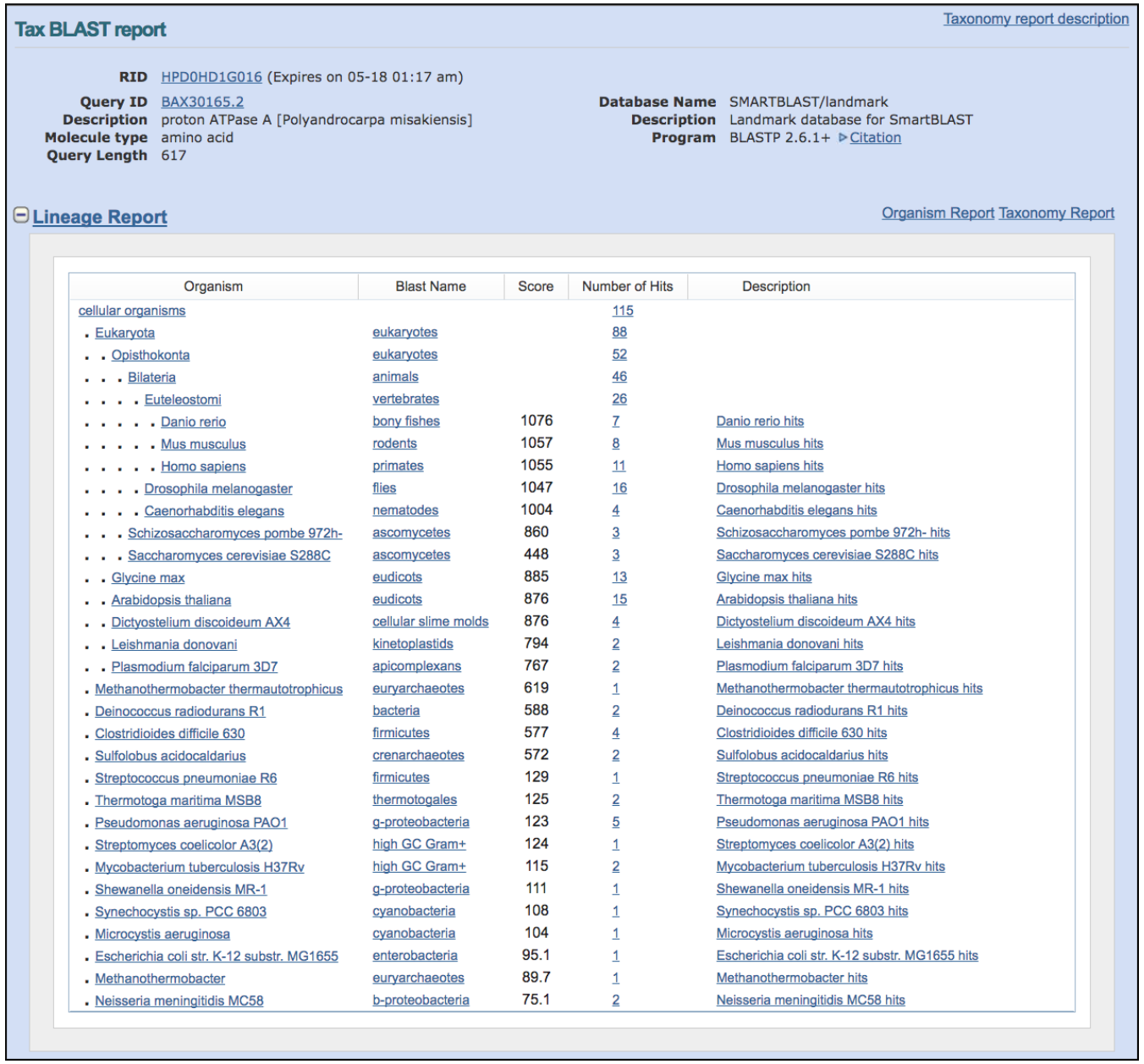
An interesting observation



An interesting observation



An interesting observation



Tax BLAST report — Taxonomy report description

RID HPD0HD1G016 (Expires on 05-18 01:17 am)
Query ID BAX30165.2
Description proton ATPase A [Polyandrocarpa misakiensis]
Molecule type amino acid
Query Length 617
Database Name SMARTBLAST/landmark
Description Landmark database for SmartBLAST
Program BLASTP 2.6.1+ ▷Citation

Lineage Report — Organism Report Taxonomy Report

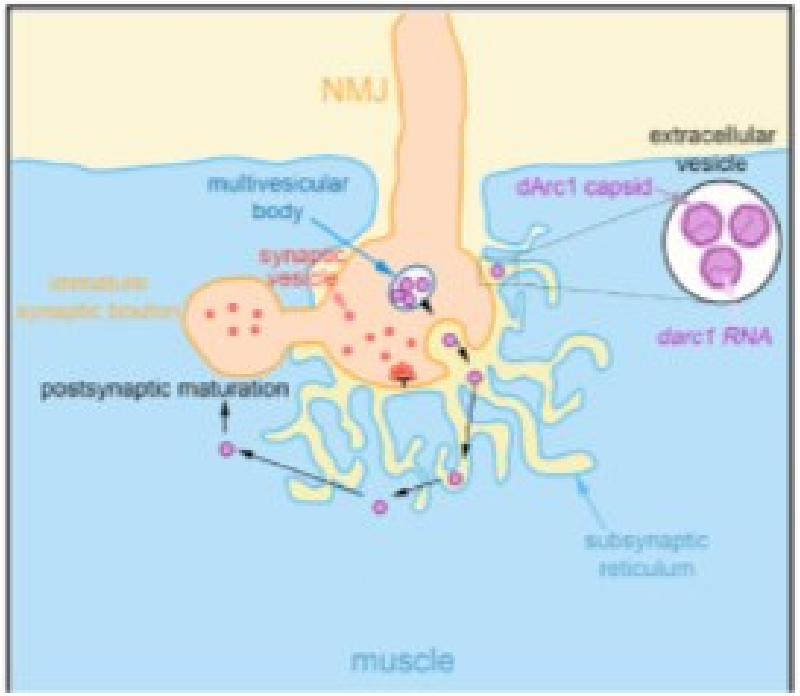
Organism	Blast Name	Score	Number of Hits	Description
cellular organisms			115	
. Eukaryota	eukaryotes		88	
. . Opisthokonta	eukaryotes		52	
. . . Bilateria	animals		46	
. . . . Euteleostomi	vertebrates		26	
. . . . . Danio rerio	bony fishes	1076	7	Danio rerio hits
. . . . . Mus musculus	rodents	1057	8	Mus musculus hits
. . . . . Homo sapiens	primates	1055	11	Homo sapiens hits
. . . . Drosophila melanogaster	flies	1047	16	Drosophila melanogaster hits
. . . . Caenorhabditis elegans	nematodes	1004	4	Caenorhabditis elegans hits
. . . Schizosaccharomyces pombe 972h-	ascomycetes	860	3	Schizosaccharomyces pombe 972h- hits
. . . Saccharomyces cerevisiae S288C	ascomycetes	448	3	Saccharomyces cerevisiae S288C hits
. . Glycine max	eudicots	885	13	Glycine max hits
. . Arabidopsis thaliana	eudicots	876	15	Arabidopsis thaliana hits
. . Dictyostelium discoideum AX4	cellular slime molds	876	4	Dictyostelium discoideum AX4 hits
. . Leishmania donovani	kinetoplastids	794	2	Leishmania donovani hits
. . Plasmodium falciparum 3D7	apicomplexans	767	2	Plasmodium falciparum 3D7 hits
. Methanothermobacter thermautotrophicus	euryarchaeotes	619	1	Methanothermobacter thermautotrophicus hits
. Deinococcus radiodurans R1	bacteria	588	2	Deinococcus radiodurans R1 hits
. Clostridioides difficile 630	firmicutes	577	4	Clostridioides difficile 630 hits
. Sulfolobus acidocaldarius	crenarchaeotes	572	2	Sulfolobus acidocaldarius hits
. Streptococcus pneumoniae R6	firmicutes	129	1	Streptococcus pneumoniae R6 hits
. Thermotoga maritima MSB8	thermotogales	125	2	Thermotoga maritima MSB8 hits
. Pseudomonas aeruginosa PAO1	g-proteobacteria	123	5	Pseudomonas aeruginosa PAO1 hits
. Streptomyces coelicolor A3(2)	high GC Gram+	124	1	Streptomyces coelicolor A3(2) hits
. Mycobacterium tuberculosis H37Rv	high GC Gram+	115	2	Mycobacterium tuberculosis H37Rv hits
. Shewanella oneidensis MR-1	g-proteobacteria	111	1	Shewanella oneidensis MR-1 hits
. Synechocystis sp. PCC 6803	cyanobacteria	108	1	Synechocystis sp. PCC 6803 hits
. Microcystis aeruginosa	cyanobacteria	104	1	Microcystis aeruginosa hits
. Escherichia coli str. K-12 substr. MG1655	enterobacteria	95.1	1	Escherichia coli str. K-12 substr. MG1655 hits
. Methanothermobacter	euryarchaeotes	89.7	1	Methanothermobacter hits
. Neisseria meningitidis MC58	b-proteobacteria	75.1	2	Neisseria meningitidis MC58 hits

Cells hack virus-like protein to communicate

Cell

Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons

Graphical Abstract



Authors

James Ashley, Benjamin Cordy, Diandra Lucia, Lee G. Fradkin, Vivian Budnik, Travis Thomson

Correspondence

vivian.budnik@umassmed.edu (V.I
travis.thomson@umassmed.edu (T

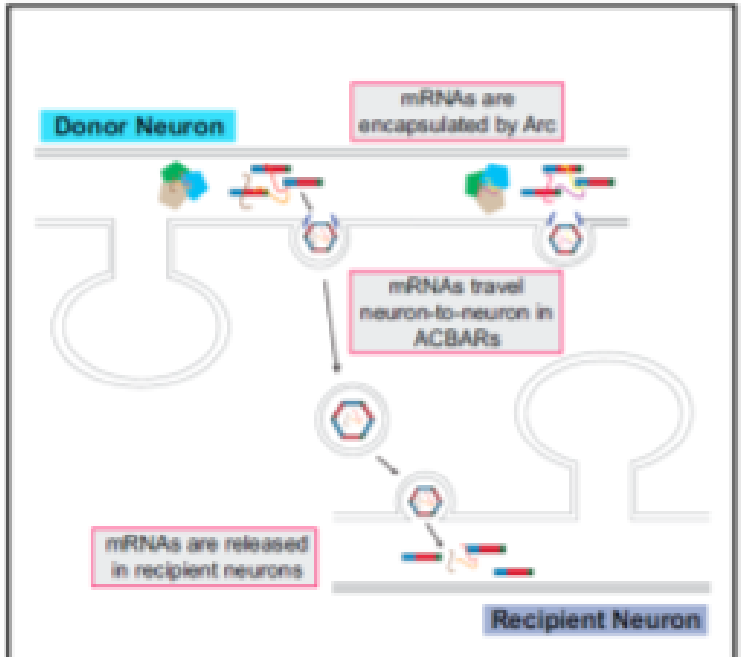
In Brief

The neuronal protein Arc is evoluti
related to retrotransposon Gag pr
and forms virus-like capsid structi
transmit mRNA between cells in th
nervous system.

Cell

The Neuronal Gene *Arc* Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer

Graphical Abstract



Authors

Elissa D. Pastuzyn, Cameron E. D
Rachel B. Kearns, ..., John A.G. E
Cédric Feschotte, Jason D. Shep

Correspondence

jason.shepherd@neuro.utah.edu

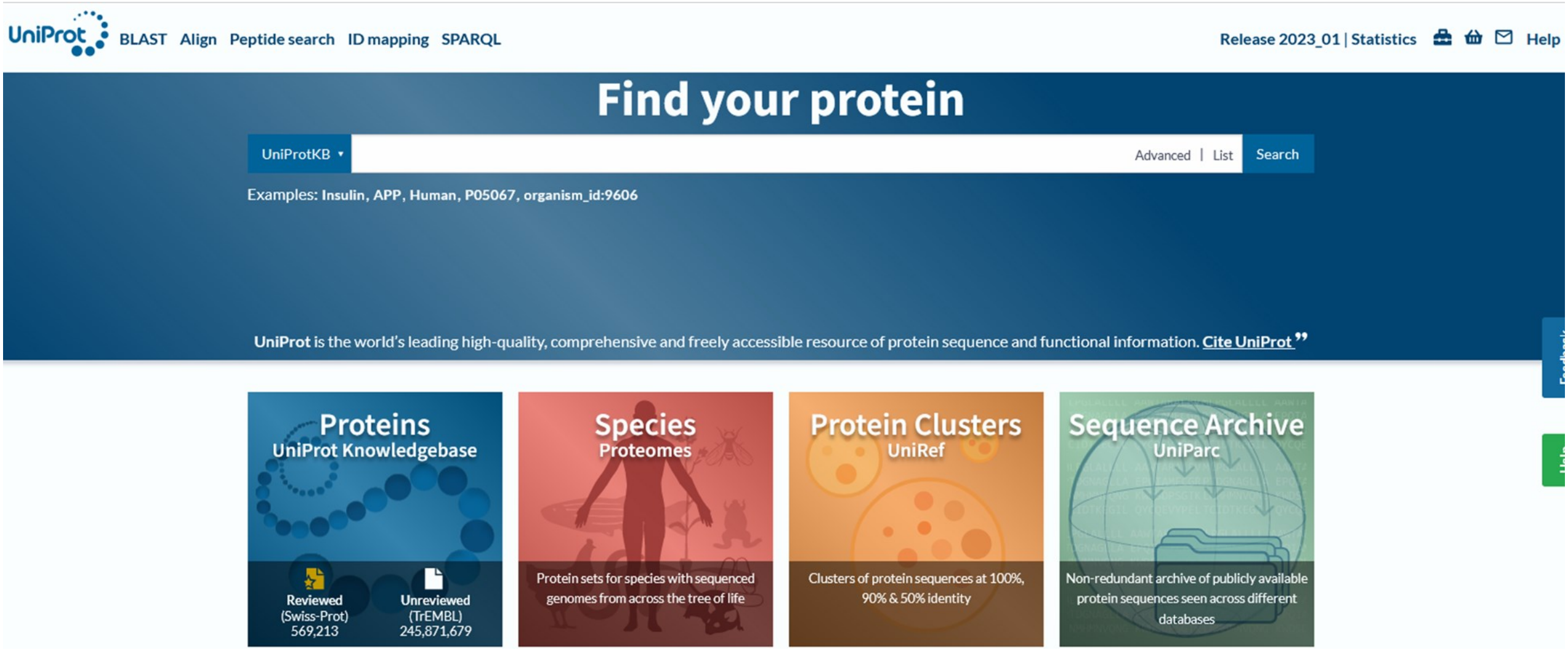
In Brief

The neuronal protein Arc is evolu
related to retrotransposon Gag p
and forms virus-like capsid struc
that can transfer mRNA between
the nervous system.

Endogenous retroviruses (ERVs) are remnants of ancient retroviral infections, and comprise nearly 8% of the human genome.

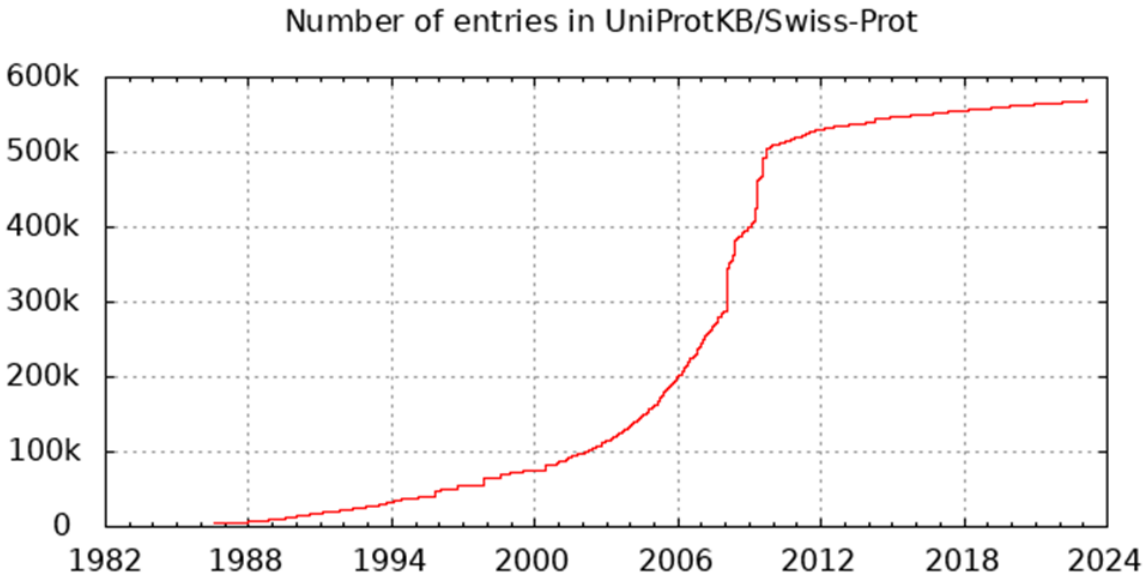
UniProtKB

UniProtKB

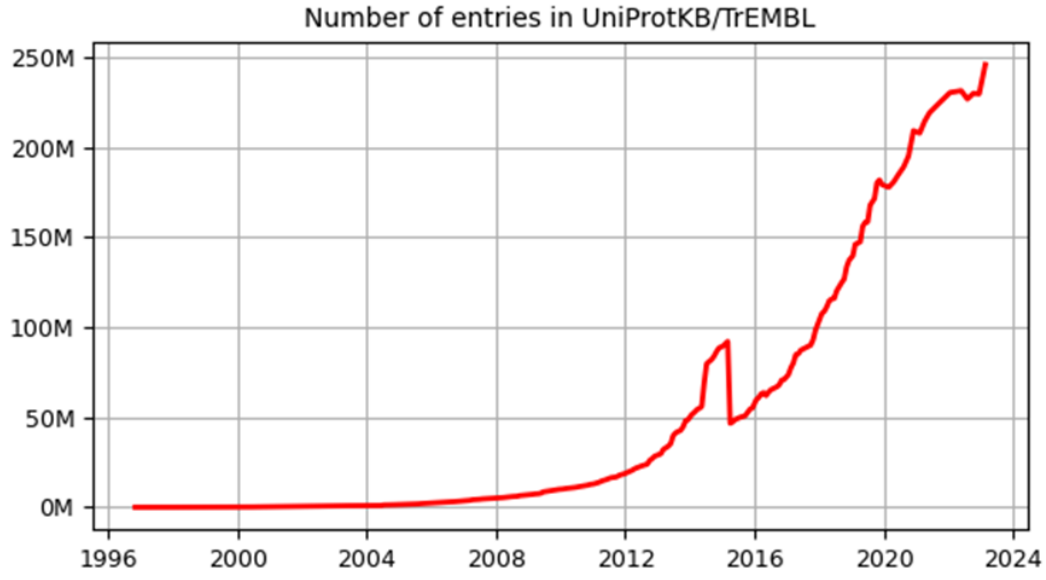

UniProt
BLAST
Align
Peptide search
ID mapping
SPARQL
Release 2023_01 | Statistics
Help
Find your protein
UniProtKB
Advanced | List
Search
Examples: Insulin, APP, Human, P05067, organism_id:9606
UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt
Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 569,213
Unreviewed (TrEMBL) 245,871,679
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

Growth of UniProtKB

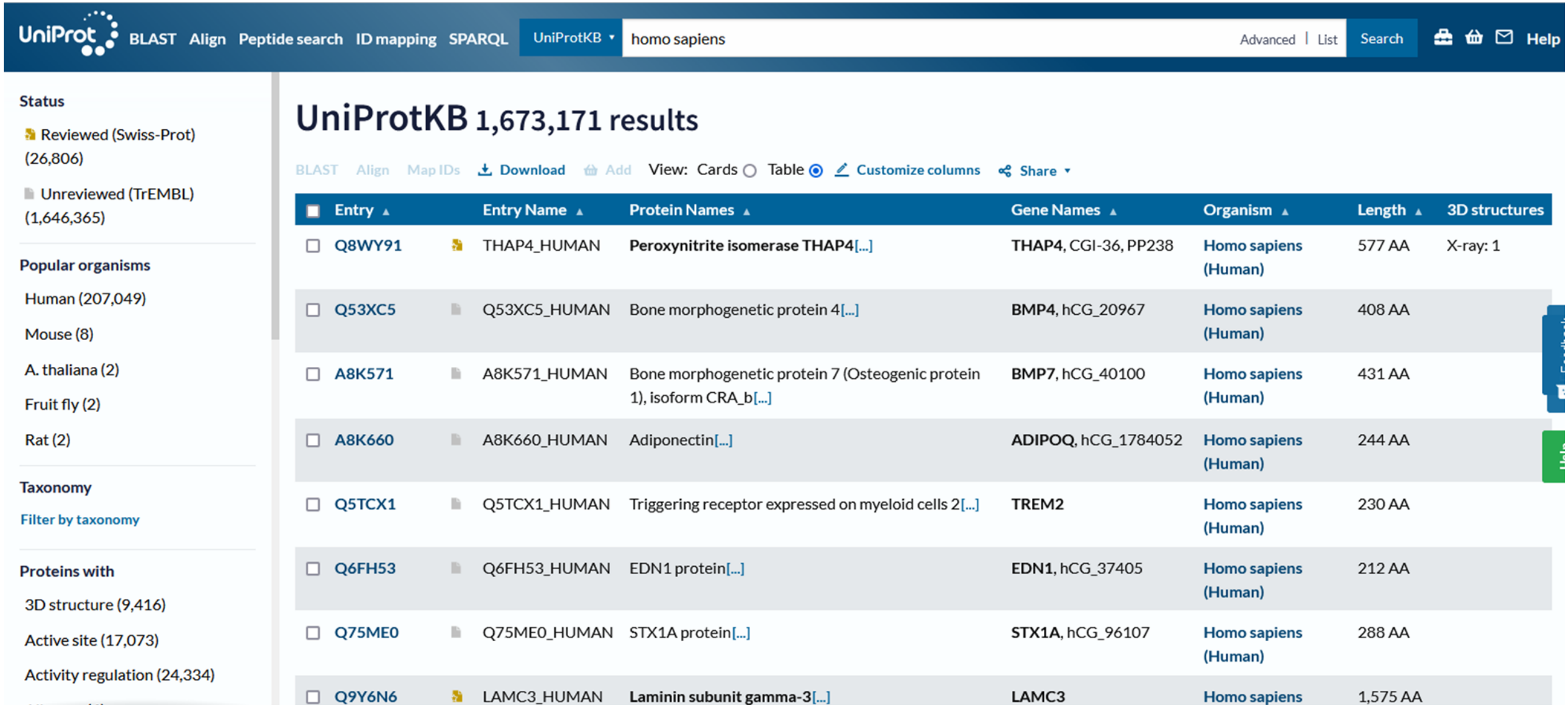
Release **2023_03** of **28-Jun-2023** of **UniProtKB/Swiss-Prot** contains **5,69,793** sequence entries, curated from **2,93,323** unique references and comprising **206,004,162** amino acids.



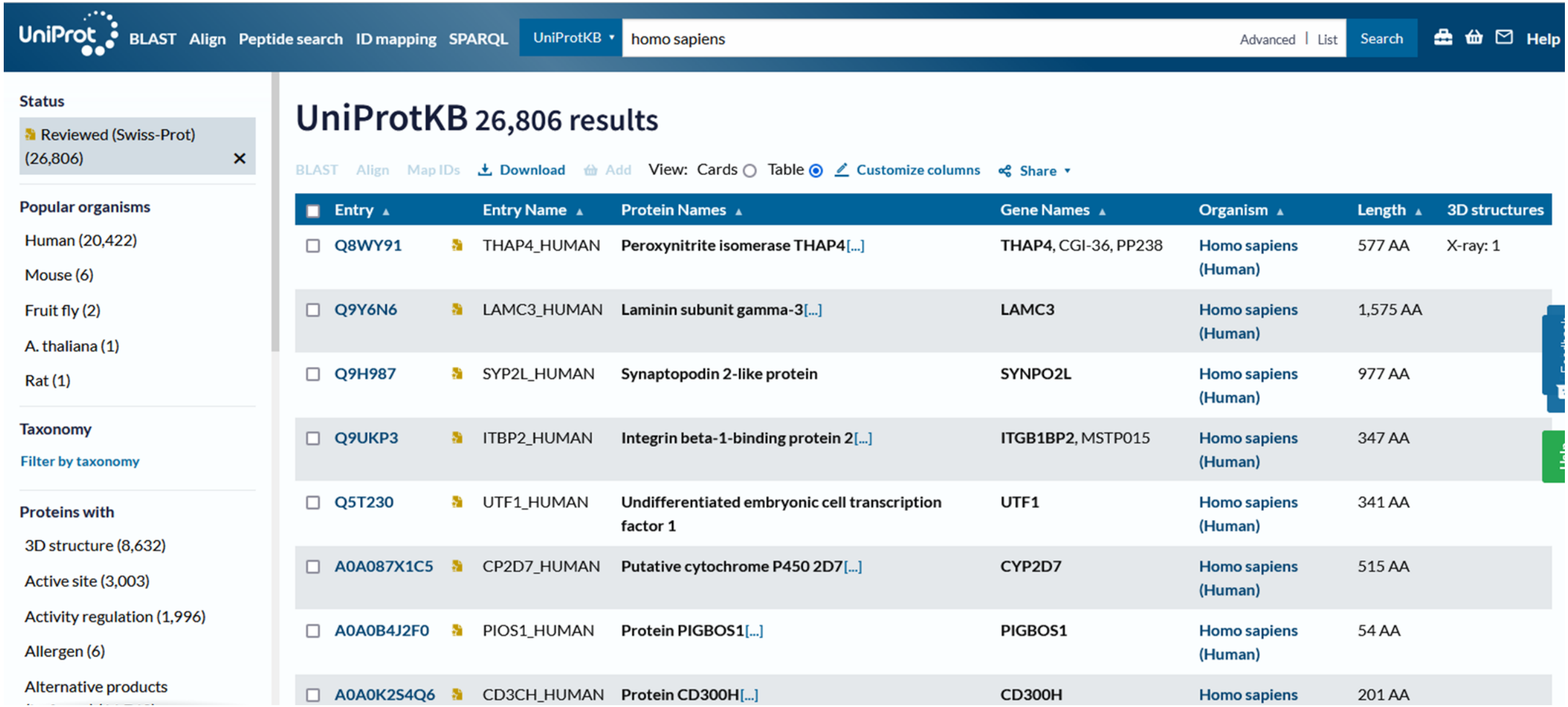
Release **2023_03** of **28-Jun-2023** of **UniProtKB/TrEMBL** contains **248,272,897** sequence entries, comprising **86,599,718,967** amino acids.



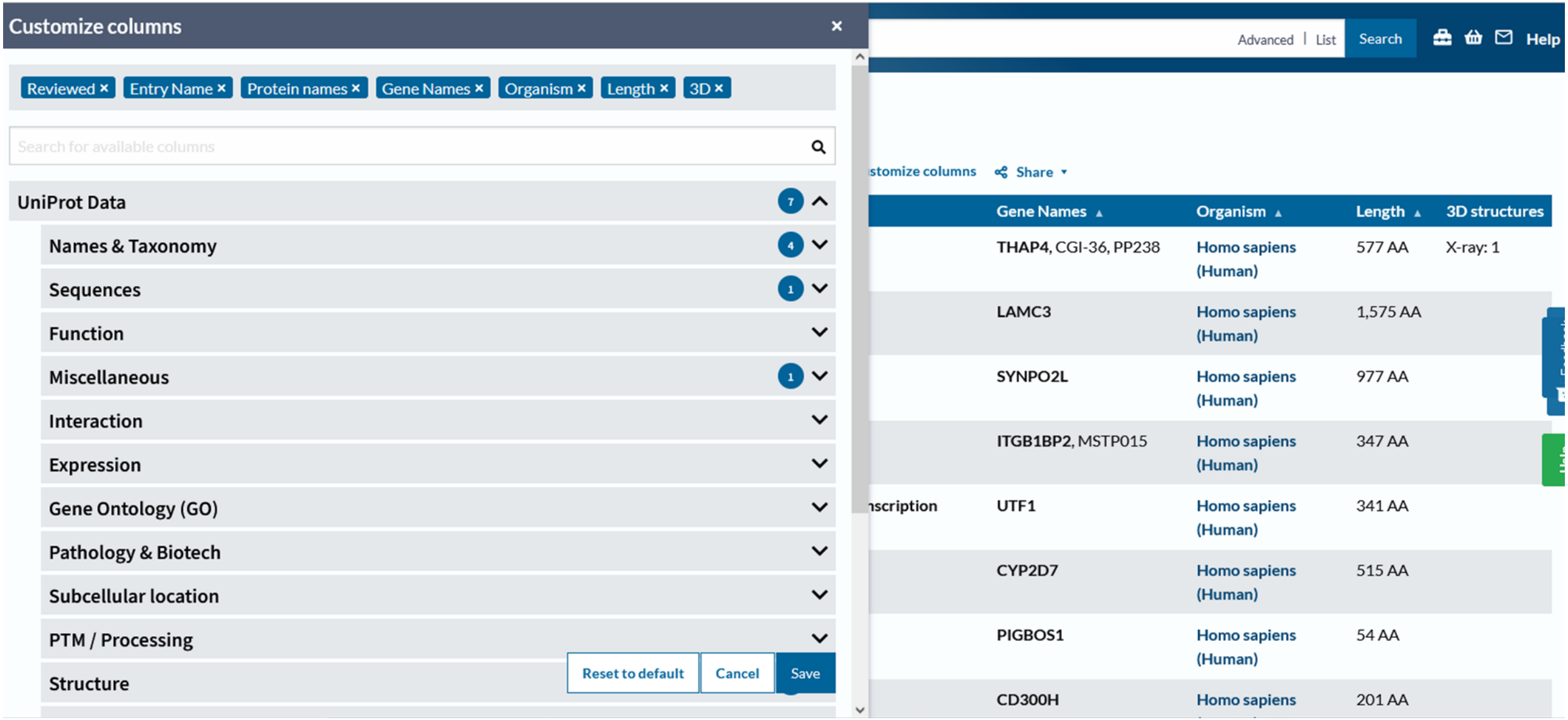
UniProtKB: Search



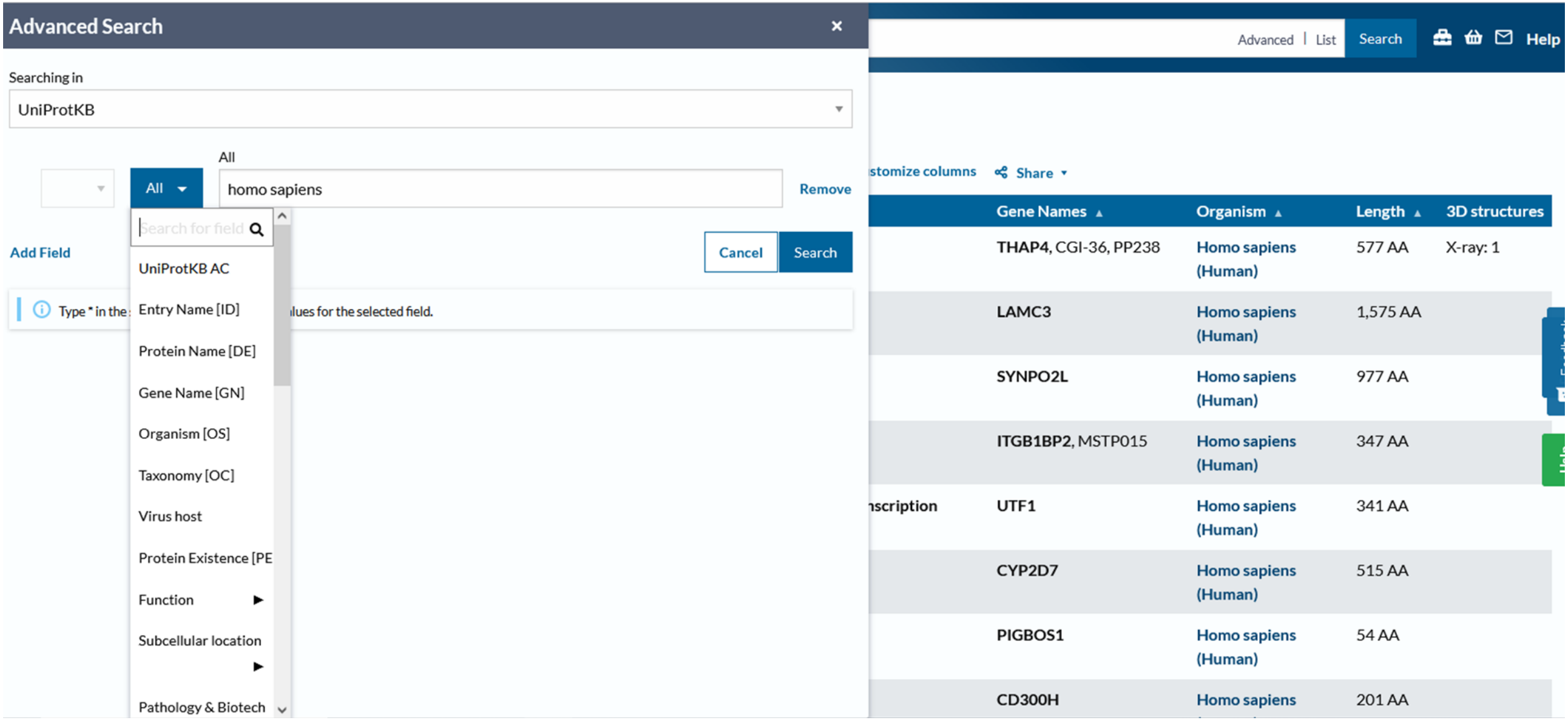
UniProtKB: customize results



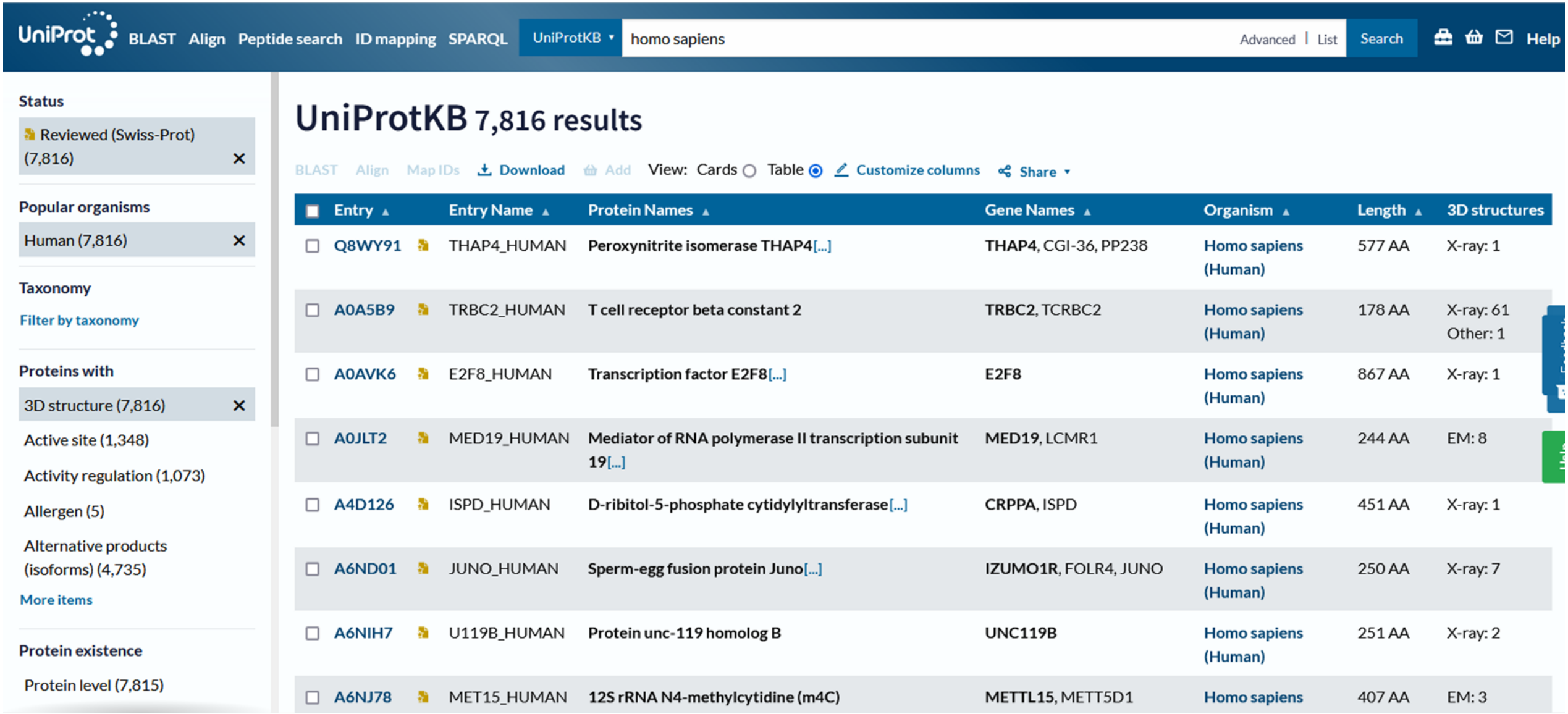
UniProtKB: customize results



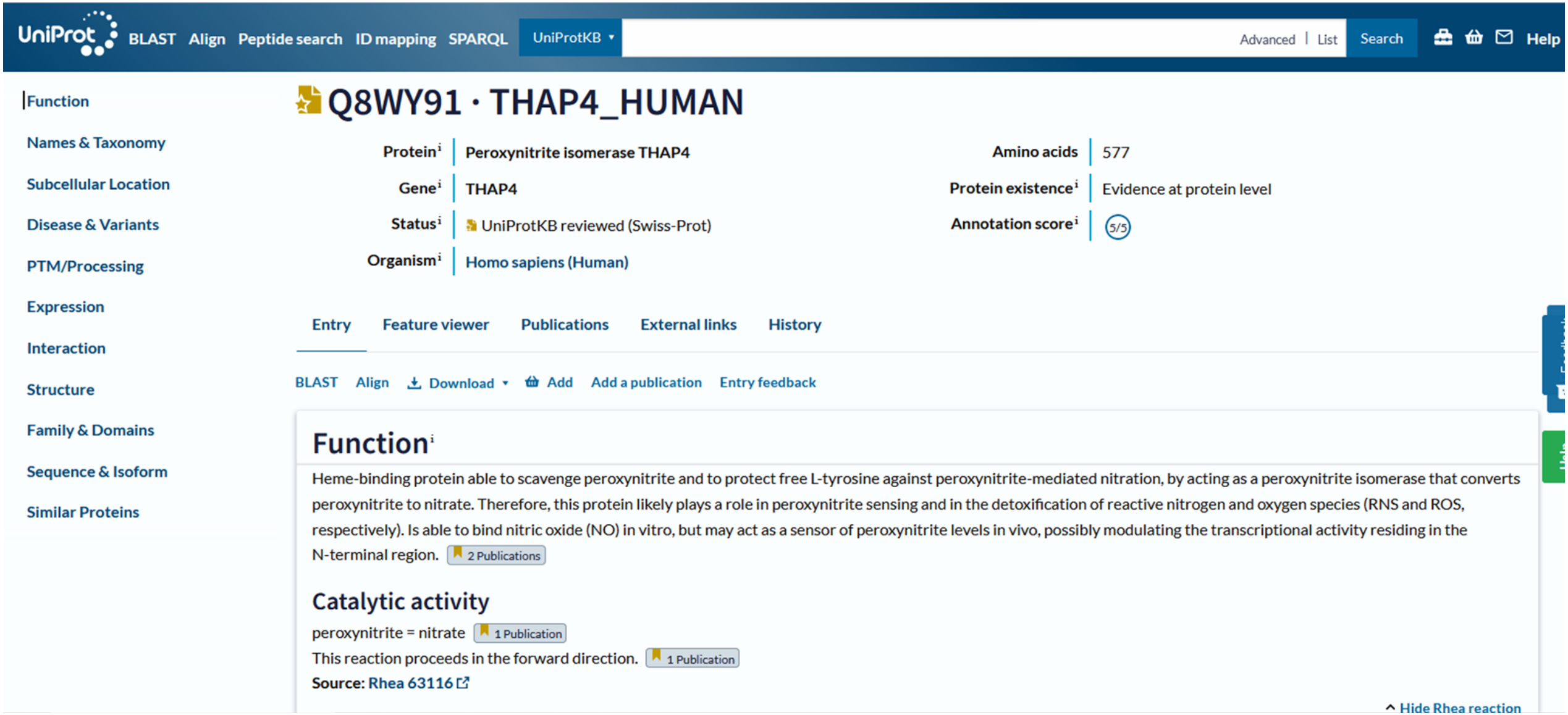
UniProtKB: advanced search



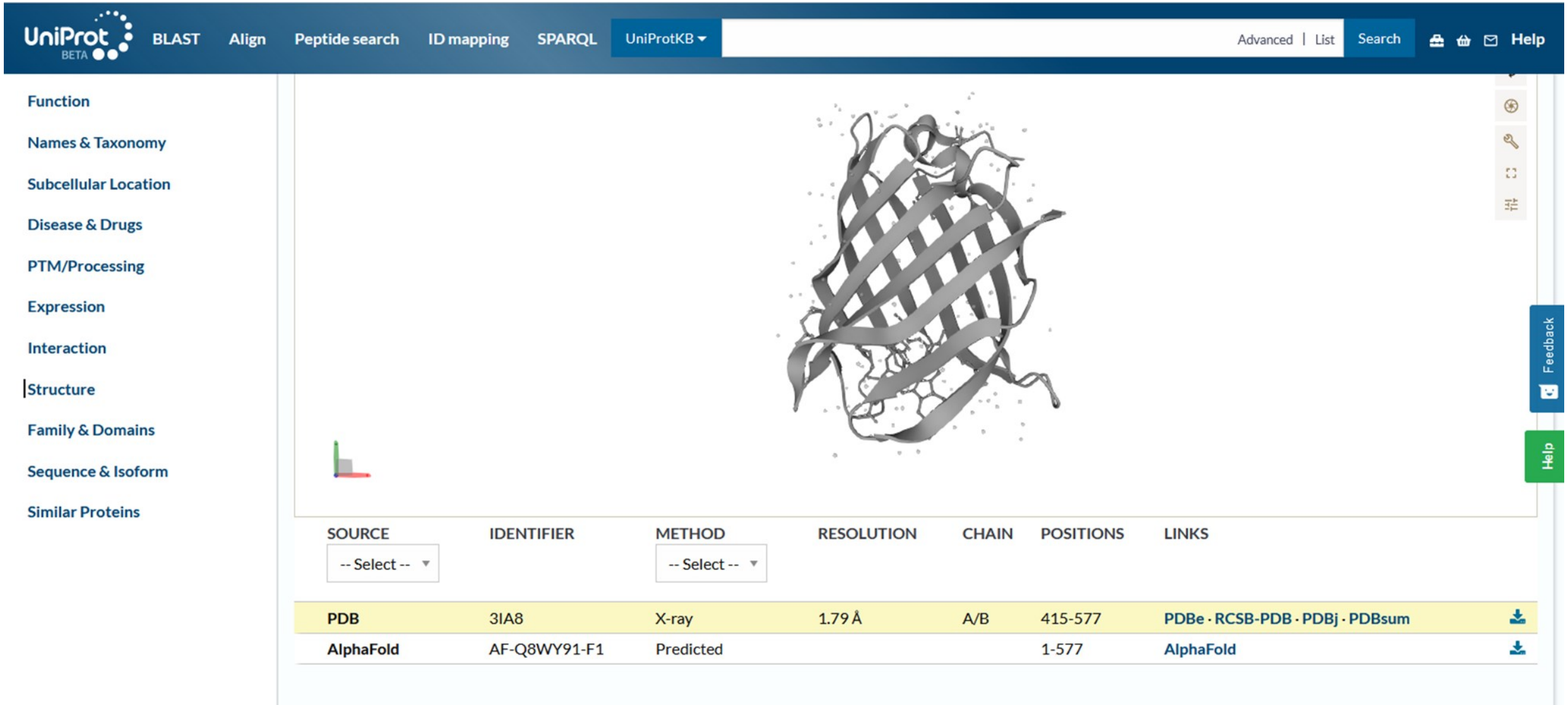
UniProtKB: refine search



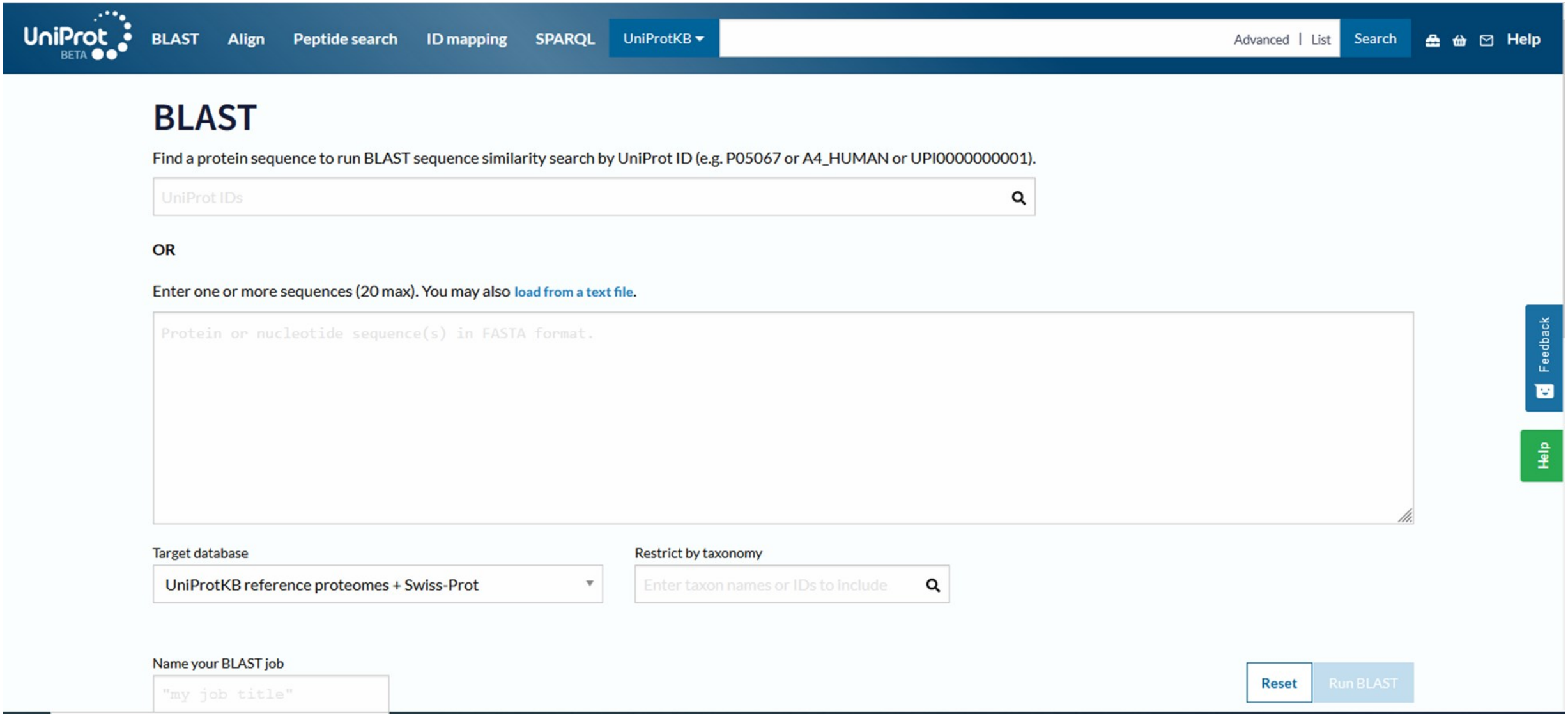
UniProtKB: exploring a protein



UniProtKB: exploring a protein



Tools on the UniProtKB web site



Protein Data Bank

Protein Data Bank (PDB)

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK

207,791 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

3D Structures | Enter search term(s), Entry ID(s), or sequence | Include CSM

Advanced Search | Browse Annotations Help

PDB-101 wwPDB EMDataResource NAKB wwPDB Foundation PDB-Dev

New: More Computed Structure Models (CSM) available Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

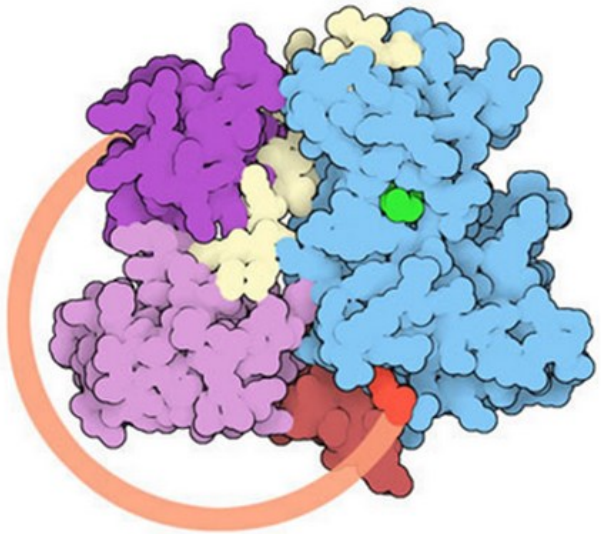
RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive

Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

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

July **Molecule of the Month**



c-Abl Protein Kinase and Imatinib

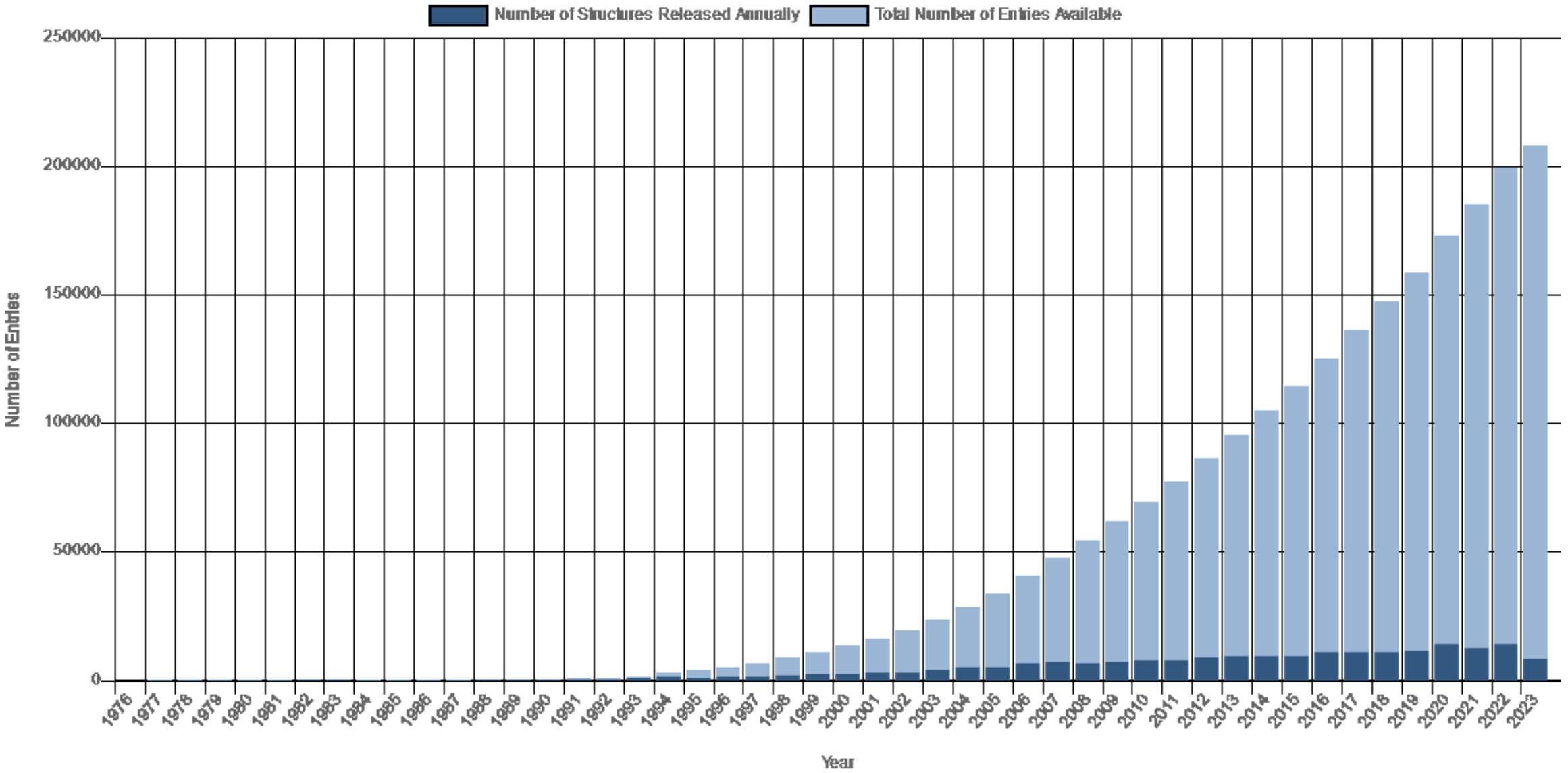
PDB Current Holdings Breakdown

Molecular type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	1,56,853	12,236	10,865	192	72	32	**1,80,250**
Protein + Oligosaccharide	9,180	34	1,906	7	1	0	**11,128**
Protein + Nucleic acid	8,200	283	3,420	7	0	0	**11,910**
Nucleic acid (only)	2,698	1,454	108	13	2	1	**4,276**
Other	164	32	9	0	0	0	**205**
Oligosaccharide (only)	11	6	0	1	0	4	**22**
Total	**1,77,106**	**14,045**	**16,308**	**220**	**75**	**37**	**2,07,791**

63,708 Structures of Human Sequences
16,385 Nucleic Acid Containing Structures
999,251 AlphaFoldDB

As of July 25, 2023

Growth of PDB



Search suggestions

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | About | Documentation | Careers | MyPDB | Contact us

RCSB PDB PROTEIN DATA BANK

201,979 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

3D Structures | sar-cov-2 | Include CSM

Advanced Search | Browse Annotations | Help

PDB-101 | wwPDB | EMDataResource | NUCLEIC ACID DATABASE | wwPDB Foundation | PDB-Dev

Search | Query History | Browse Annotations | MyPDB

QUERY: Full Text = ""sars-cov-2""

JSON | MyPDB Login

Advanced Search Query Builder

Help

Search Summary | This query matches 3,322 **Structures**.

Refinements

Structure Determination Methodology

- experimental (3,322)

Scientific Name of Source Organism

- Severe acute respiratory syndrome coronavirus 2 (2,964)
- Homo sapiens (1,028)
- Mus musculus (104)
- synthetic construct (86)
- Lama glama (66)
- Severe acute respiratory syndrome-related coronavirus (46)
- Vicugna pacos (37)

-- Tabular Report --

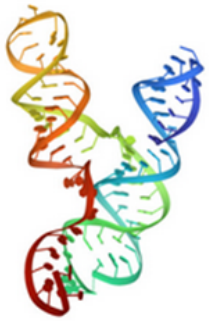
All | Selected

1 to 25 of 3,322 **Structures**

Page 1 of 133

25

Sort by ↓ Score



3D View

6XRZ

Download File | View File

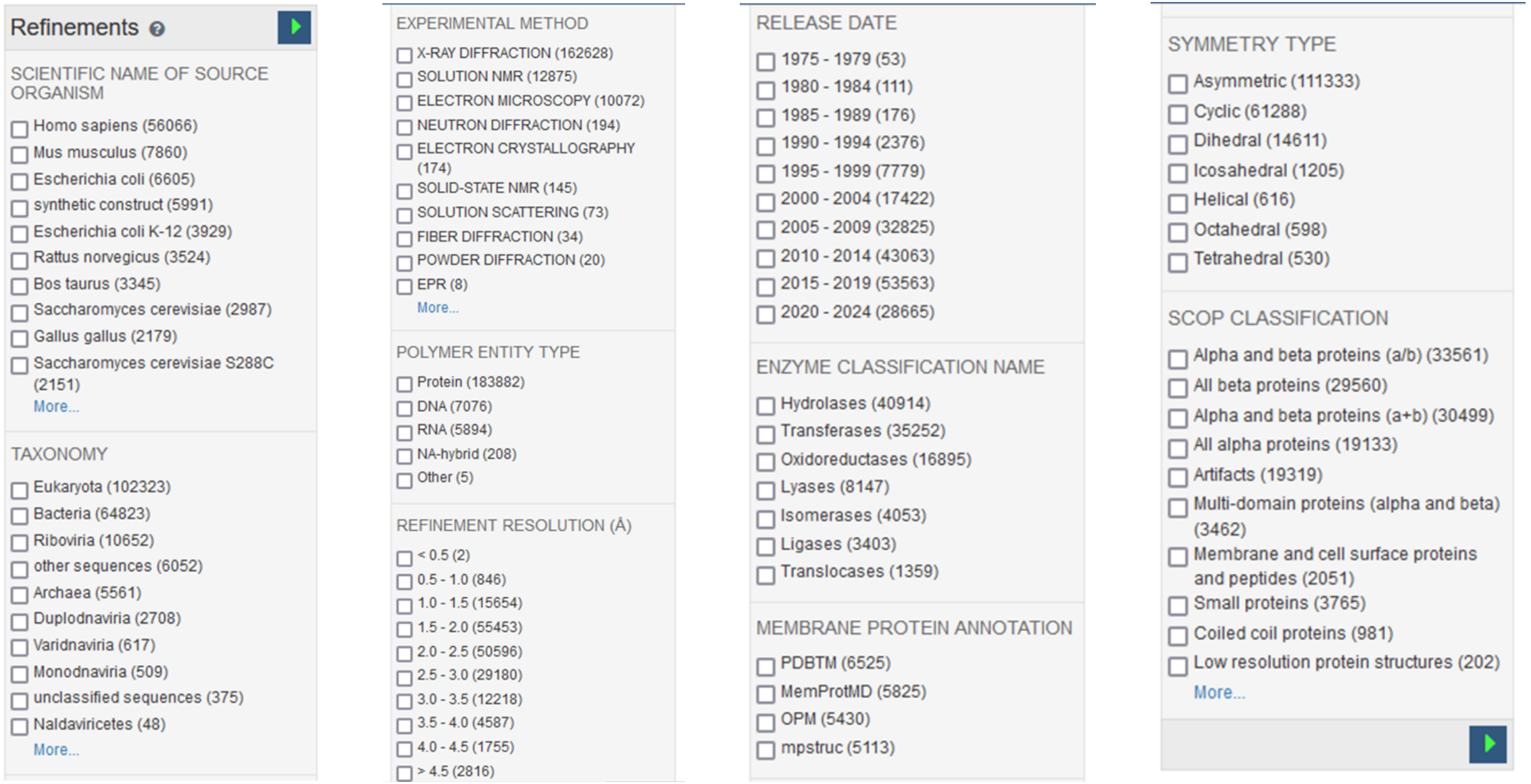
The 28-kDa Frameshift Stimulation Element from the SARS-CoV-2 RNA Genome

Zhang, K., Zheludev, I., Hagey, R., Wu, M., Haslecker, R., Hou, Y., Kretsch, R., Pintilie, G., Rangan, R., Kladwang, W., Li, S., Pham, E., Souibgui, C., Baric, R., Sheahan, T., Souza, V., Glenn, J., Chiu, W., Das, R.

(2020) Biorxiv

Released	2020-08-19
Method	ELECTRON MICROSCOPY 6.9 Å
Organisms	Severe acute respiratory syndrome coronavirus 2
Macromolecule	Frameshift Stimulation Element from the SARS-CoV-2 RNA Genome (nucleic acid)

Query refinements



Molecular descriptions

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | About | Documentation | Careers | MyPDB | Contact us

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome | Ligands | Versions

Biological Assembly 1

3D View: Structure | 1D-3D View | Electron Density | Validation Report | Ligand Interaction

Global Symmetry: Cyclic - C2 (3D View)
Global Stoichiometry: Homo 2-mer - A2

Find Similar Assemblies

Biological assembly 1 assigned by authors and

Usage | Display Files | Download Files

7P35

Structure of the SARS-CoV-2 3CL protease in complex with rupintrivir

PDB DOI: 10.2210/pdb7P35/pdb

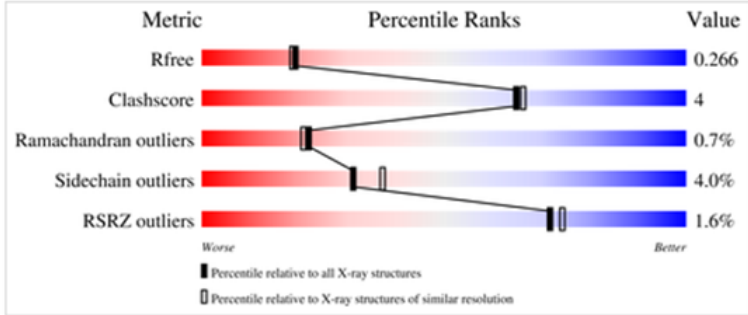
Classification: VIRAL PROTEIN
Organism(s): Severe acute respiratory syndrome coronavirus 2
Expression System: Escherichia coli
Mutation(s): No

Deposited: 2021-07-07 **Released:** 2021-07-21
Deposition Author(s): Fabrega-Ferrer, M., Perez-Saavedra, J., Herrera-Morande, A., Coll, M.
Funding Organization(s): Spanish Ministry of Science, Innovation, and Universities, Spanish National Research Council

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.26 Å
R-Value Free: 0.258
R-Value Work: 0.200

wwPDB Validation | 3D Report | Full Report



Ligand Structure Quality Assessment

Molecular descriptions

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers MyPDB Contact us

Macromolecules

Find similar proteins by: Sequence (by identity cutoff) | 3D Structure

Entity ID: 1

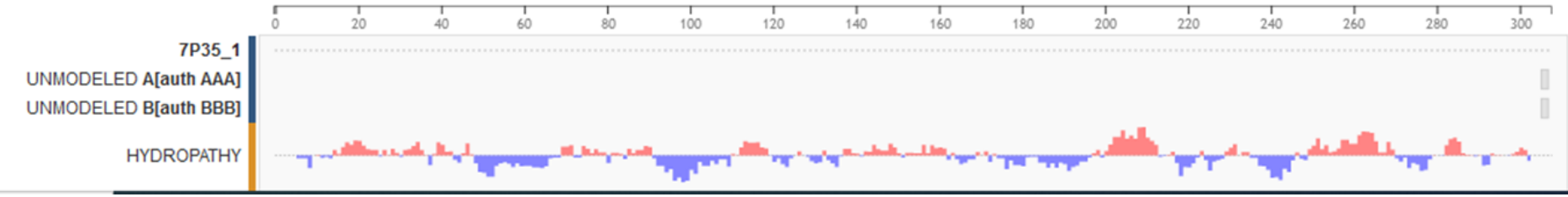
Molecule	Chains	Sequence Length	Organism	Details	Image
3C-like proteinase	A [auth AAA], B [auth BBB]	306	Severe acute respiratory syndrome coronavirus 2	**Mutation(s)**: 0 **Gene Names:** ORF1ab **EC:** 2.7.7.48 (PDB Primary Data), 3.4.19.12 (PDB Primary Data), 3.4.22.69 (PDB Primary Data), 3.6.4.12 (PDB Primary Data), 3.6.4.13 (PDB Primary Data)	

Entity Groups

Sequence Clusters	30% Identity	50% Identity	70% Identity	90% Identity	95% Identity	100% Identity

Protein Feature View Expand

Reference Sequence 7P35_1



Sequence information

7P35

Usage | Display Files | Download Files

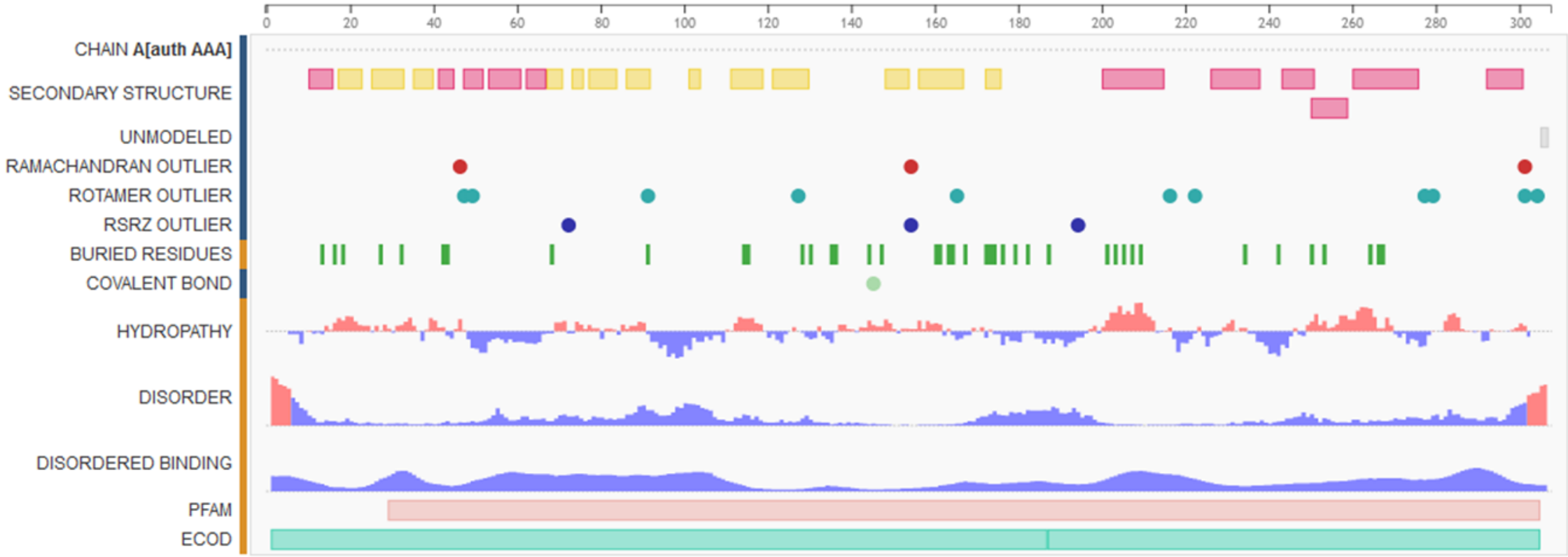
Structure of the SARS-CoV-2 3CL protease in complex with rupintrivir

Help

CHAIN

A [auth...

3c-like proteinase - Severe acute respiratory syndrome coronavirus 2 View Features in 3D



A sample of PDB file format

```
HEADER    VIRAL PROTEIN/PROTEIN BINDING           01-JUN-20   7C8J
TITLE     STRUCTURAL BASIS FOR CROSS-SPECIES RECOGNITION OF COVID-19 VIRUS SPIKE
TITLE    2 RECEPTOR BINDING DOMAIN TO BAT ACE2
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: ANGIOTENSIN-CONVERTING ENZYME;
COMPND   3 CHAIN: A;
COMPND   4 EC: 3.4.-.-;
COMPND   5 ENGINEERED: YES;
COMPND   6 MOL_ID: 2;
COMPND   7 MOLECULE: SARS-COV-2 RECEPTOR BINDING DOMAIN;
COMPND   8 CHAIN: B;
COMPND   9 FRAGMENT: UNP RESIDUES 333-527;
COMPND  10 SYNONYM: S GLYCOPROTEIN,E2,PEPLOMER PROTEIN;
COMPND  11 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: RHINOLOPHUS MACROTIS;
SOURCE   3 ORGANISM_COMMON: BIG-EARED HORSESHOE BAT;
SOURCE   4 ORGANISM_TAXID: 196889;
SOURCE   5 GENE: ACE2;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE   8 MOL_ID: 2;
SOURCE   9 ORGANISM_SCIENTIFIC: SEVERE ACUTE RESPIRATORY SYNDROME CORONAVIRUS
SOURCE  10 2;
SOURCE  11 ORGANISM_COMMON: 2019-NCOV;
SOURCE  12 ORGANISM_TAXID: 2697049;
SOURCE  13 GENE: S, 2;
SOURCE  14 EXPRESSION_SYSTEM: INSECT CELL EXPRESSION VECTOR PTIE1;
SOURCE  15 EXPRESSION_SYSTEM_TAXID: 266783
KEYWDS    COVID-19, RECEPTOR BINDING DOMAIN (RBD), RHINOLOPHUS MACROTIS, BATS,
KEYWDS   2 ACE2, VIRAL PROTEIN-PROTEIN BINDING COMPLEX
EXPDTA    X-RAY DIFFRACTION
AUTHOR    K.F.LIU,J.WANG,S.G.TAN,S.NIU,L.L.WU,Y.F.ZHANG,X.Q.PAN,Y.M.MENG,
AUTHOR   2 Q.CHEN,Q.H.WANG,H.W.WANG,J.X.QI,G.F.GAO
REVDAT   2   03-FEB-21 7C8J    1       REMARK
REVDAT   1   27-JAN-21 7C8J    0
JRNL        AUTH    K.LIU,S.TAN,S.NIU,J.WANG,L.WU,H.SUN,Y.ZHANG,X.PAN,X.QU,P.DU,
JRNL        AUTH 2 Y.MENG,Y.JIA,Q.CHEN,C.DENG,J.YAN,H.W.WANG,Q.WANG,J.QI,
JRNL        AUTH 3 G.F.GAO
JRNL        TITL    CROSS-SPECIES RECOGNITION OF SARS-COV-2 TO BAT ACE2.
JRNL        REF     PROC.NATL.ACAD.SCI.USA        V. 118       2021
JRNL        REFN                    ESSN 1091-6490
JRNL        PMID    33335073
JRNL        DOI     10.1073/PNAS.2020216118
```

```
SEQRES   1 A  707  THR THR GLU ASP GLU ALA LYS LYS PHE LEU ASP LYS PHE
SEQRES   2 A  707  ASN SER LYS ALA GLU ASP LEU SER TYR GLU SER SER LEU
SEQRES   3 A  707  ALA SER TRP ASP TYR ASN THR ASN ILE SER ASP GLU ASN
SEQRES   4 A  707  VAL GLN LYS MET ASP GLU ALA GLY ALA LYS TRP SER ALA
SEQRES   5 A  707  PHE TYR GLU GLU GLN SER LYS LEU ALA LYS ASN TYR PRO
SEQRES   6 A  707  LEU GLU GLU ILE GLN ASN ASP THR VAL LYS ARG GLN LEU
SEQRES   7 A  707  GLN ILE LEU GLN GLN GLN SER GLY SER PRO VAL LEU SER GLU
SEQRES   8 A  707  ASP LYS SER LYS ARG LEU ASN SER ILE LEU ASN ALA MET
SEQRES   9 A  707  SER THR ILE TYR SER THR GLY LYS VAL CYS LYS PRO ASN
SEQRES  10 A  707  ASN PRO GLN GLU CYS LEU LEU GLU PRO GLY LEU ASP
SEQRES  11 A  707  ASN ILE MET GLY THR SER LYS ASP TYR ASN GLU ARG LEU
SEQRES  12 A  707  TRP ALA TRP GLU GLY TRP ARG ALA GLU VAL GLY LYS GLN
SEQRES  13 A  707  LEU ARG PRO LEU TYR GLU GLU TYR VAL VAL LEU LYS ASN
SEQRES  14 A  707  GLU MET ALA ARG GLY TYR HIS TYR GLU ASP TYR GLY ASP
SEQRES  15 A  707  TYR TRP ARG ARG ASP TYR GLU THR GLU GLU SER SER GLY
SEQRES  16 A  707  PRO GLY TYR SER ARG ASP GLN LEU MET LYS ASP VAL ASP
SEQRES  17 A  707  ARG ILE PHE THR GLU ILE LYS PRO LEU TYR GLU HIS LEU
SEQRES  18 A  707  HIS ALA TYR VAL ARG ALA LYS LEU MET ASP THR TYR PRO
SEQRES  19 A  707  LEU HIS ILE SER PRO THR GLY CYS LEU PRO ALA HIS LEU
SEQRES  20 A  707  LEU GLY ASP MET TRP GLY ARG PHE TRP THR ASN LEU TYR
SEQRES  21 A  707  PRO LEU THR VAL PRO PHE GLY GLN LYS PRO ASN ILE ASP
SEQRES  22 A  707  VAL THR ASP ALA MET LEU ASN GLN GLY TRP ASP ALA ASN
SEQRES  23 A  707  ARG ILE PHE LYS GLU ALA GLU LYS PHE PHE VAL SER VAL
SEQRES  24 A  707  SER LEU PRO LYS MET THR GLU GLY PHE TRP ASN LYS SER
SEQRES  25 A  707  MET LEU THR GLU PRO GLY ASP GLY ARG LYS VAL VAL CYS
SEQRES  26 A  707  HIS PRO THR ALA TRP ASP LEU GLY LYS GLY ASP PHE ARG
SEQRES  27 A  707  ILE LYS MET CYS THR LYS VAL THR MET GLU ASP PHE LEU
SEQRES  28 A  707  THR ALA HIS HIS GLU MET GLY HIS ILE GLN TYR ASP MET
SEQRES  29 A  707  ALA TYR ALA SER GLN PRO TYR LEU LEU ARG ASN GLY ALA
SEQRES  30 A  707  ASN GLU GLY PHE HIS GLU ALA VAL GLY GLU VAL MET SER
SEQRES  31 A  707  LEU SER VAL ALA THR PRO LYS HIS LEU LYS THR MET GLY
SEQRES  32 A  707  LEU LEU SER PRO ASP PHE ARG GLU ASP ASP GLU THR GLU
SEQRES  33 A  707  ILE ASN PHE LEU LEU LYS GLN ALA LEU ASN ILE VAL GLY
SEQRES  34 A  707  THR LEU PRO PHE THR TYR MET LEU GLU LYS TRP ARG TRP
SEQRES  35 A  707  MET VAL PHE LYS GLY GLU ILE PRO LYS GLU GLU TRP MET
SEQRES  36 A  707  LYS LYS TRP TRP GLU MET LYS ARG GLU ILE VAL GLY VAL
SEQRES  37 A  707  VAL GLU PRO VAL PRO HIS ASP GLU THR TYR CYS ASP PRO
SEQRES  38 A  707  ALA SER LEU PHE HIS VAL ALA ASN ASP TYR SER PHE ILE
SEQRES  39 A  707  ARG TYR TYR THR ARG THR ILE PHE GLU PHE GLN PHE HIS
SEQRES  40 A  707  GLU ALA LEU CYS ARG ILE ALA GLN HIS ASN GLY PRO LEU
SEQRES  41 A  707  HIS LYS CYS ASP ILE SER ASN SER THR ASP ALA GLY LYS
SEQRES  42 A  707  LYS LEU HIS GLN MET LEU SER VAL GLY LYS SER GLN ALA
SEQRES  43 A  707  TRP THR LYS THR LEU GLU ASP ILE VAL ASP SER ARG ASN
SEQRES  44 A  707  MET ASP VAL GLY PRO LEU LEU ARG TYR PHE LYS PRO LEU
SEQRES  45 A  707  TYR THR TRP LEU GLN GLU GLN ASN ARG LYS SER TYR VAL
SEQRES  46 A  707  GLY TRP ASN THR ASP TRP SER PRO TYR ALA ASP GLN SER
SEQRES  47 A  707  ILE LYS VAL TRP ILE SER LEU LYS SER ALA LEU GLY GLU
```

```
HET     ZN  A 801       1
HETNAM      ZN ZINC ION
FORMUL   3   ZN    ZN 2+
HELIX    1 AA1 THR A   20  ASN A   53  1                                  34
HELIX    2 AA2 ASP A   56  ASN A   82  1                                  27
HELIX    3 AA3 ASN A   90  SER A  103  1                                  14
HELIX    4 AA4 SER A  109  GLY A  130  1                                  22
HELIX    5 AA5 PRO A  146  SER A  155  1                                  10
HELIX    6 AA6 ASP A  157  GLU A  171  1                                  15
HELIX    7 AA7 VAL A  172  TYR A  194  1                                  23
HELIX    8 AA8 ASP A  198  ARG A  204  1                                   7
HELIX    9 AA9 ARG A  205  GLU A  208  5                                   4
HELIX   10 AB1 SER A  218  TYR A  252  1                                  35
HELIX   11 AB2 TRP A  275  ASN A  277  5                                   3
HELIX   12 AB3 LEU A  278  VAL A  283  1                                   6
HELIX   13 AB4 VAL A  293  GLN A  300  1                                   8
HELIX   14 AB5 ASP A  303  VAL A  318  1                                  16
HELIX   15 AB6 THR A  324  SER A  331  1                                   8
HELIX   16 AB7 MET A  366  TYR A  385  1                                  20
HELIX   17 AB8 ALA A  386  GLN A  388  5                                   3
HELIX   18 AB9 PRO A  389  ARG A  393  5                                   5
HELIX   19 AC1 PHE A  400  ALA A  413  1                                  14
HELIX   20 AC2 THR A  414  THR A  420  1                                   7
HELIX   21 AC3 ASP A  431  VAL A  447  1                                  17
HELIX   22 AC4 GLY A  448  GLY A  466  1                                  19
HELIX   23 AC5 PRO A  469  GLU A  471  5                                   3
HELIX   24 AC6 GLU A  472  ILE A  484  1                                  13
HELIX   25 AC7 CYS A  498  SER A  502  5                                   5
HELIX   26 AC8 LEU A  503  ASN A  508  1                                   6
HELIX   27 AC9 PHE A  512  ALA A  533  1                                  22
HELIX   28 AD1 PRO A  538  CYS A  542  5                                   5
HELIX   29 AD2 SER A  547  GLY A  561  1                                  15
HELIX   30 AD3 ALA A  565  ASP A  575  1                                  11
HELIX   31 AD4 VAL A  581  ASN A  599  1                                  19
HELIX   32 AD5 ASN A  636  LYS A  659  1                                  24
HELIX   33 AD6 GLY A  666  GLU A  668  5                                   3
HELIX   34 AD7 PRO A  696  ARG A  716  1                                  21
HELIX   35 AD8 PRO B  337  ASN B  343  1                                   7
HELIX   36 AD9 ASP B  364  ASN B  370  1                                   7
HELIX   37 AE1 ASP B  405  ILE B  410  5                                   6
HELIX   38 AE2 GLY B  416  ASN B  422  1                                   7
HELIX   39 AE3 SER B  438  SER B  443  1                                   6
HELIX   40 AE4 GLY B  502  TYR B  505  5                                   4
SHEET    1 AA1 2 LYS A 131  VAL A 132  0
SHEET    2 AA1 2 LEU A 142  LEU A 143 -1  O  LEU A 142   N  VAL A 132
```

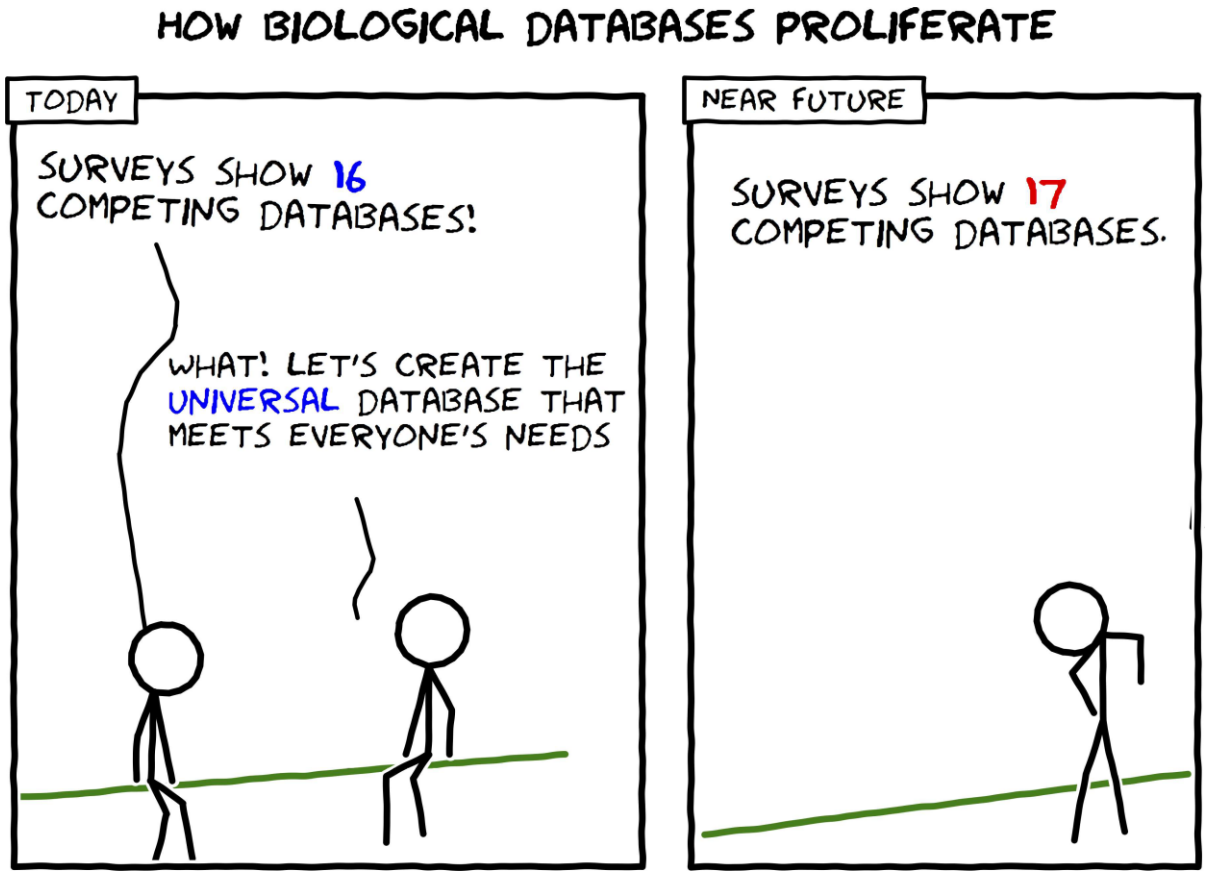
A sample of PDB file format

```
ATOM      1  N   PRO A   4       6.719 -12.134  26.603  1.00 18.91           N
ATOM      2  CA  PRO A   4       6.735 -10.746  27.122  1.00 18.45           C
ATOM      3  C   PRO A   4       6.209  -9.735  26.108  1.00 16.72           C
ATOM      4  O   PRO A   4       6.701  -9.658  24.983  1.00 16.64           O
ATOM      5  CB  PRO A   4       8.174 -10.427  27.495  1.00 20.82           C
ATOM      6  CG  PRO A   4       8.942 -11.387  26.584  1.00 20.17           C
ATOM      7  CD  PRO A   4       8.093 -12.664  26.557  1.00 22.00           C
ATOM      8  N   LEU A   5       5.207  -8.963  26.521  1.00 16.15           N
ATOM      9  CA  LEU A   5       4.605  -7.937  25.674  1.00 14.51           C
ATOM     10  C   LEU A   5       5.700  -6.960  25.244  1.00 14.38           C
ATOM     11  O   LEU A   5       6.564  -6.600  26.042  1.00 15.34           O
ATOM     12  CB  LEU A   5       3.513  -7.204  26.458  1.00 13.81           C
ATOM     13  CG  LEU A   5       2.639  -6.180  25.737  1.00 14.69           C
ATOM     14  CD1 LEU A   5       1.815  -6.864  24.656  1.00 15.29           C
ATOM     15  CD2 LEU A   5       1.725  -5.506  26.754  1.00 15.24           C
```

Ten Simple Rules for Developing Public Biological Databases

So, if you are considering developing a new database, and especially if you are a student or postdoc, please, for the love of science, follow these ten simple rules for creating and maintaining biological databases (and also a similar set of great rules for scientific web resources).

Rule 1: Don't reinvent the wheel



Ten Simple Rules for Developing Public Biological Databases

Rule 2: The three most important things in database development are data quality, data quality, and data quality

Rule 3: Know your users

Rule 4: Use modern technology

Rule 5: Put yourself in your user's shoes

Rule 6: Keep search simple and organized

Rule 7: Give users data where they need it

Rule 8: Support open science

Rule 9: Tell the world

Rule 10: Maintain, update, or retire

Thank You