

Databases

2024-04-10

WHAT IS A DATABASE?

A database is a structured, organized set of data.

In computing terminology a database refers to a software used to store and organize data.

Think of it as a file cabinet where you store data in different sections called tables.

Guess how many biological databases exist worldwide



We can find the answer in...



 **Data Statistics**

6389
DATABASES

13
CATEGORIES

1647
SPECIES

9872
PUBLICATIONS

2141
INSTITUTIONS

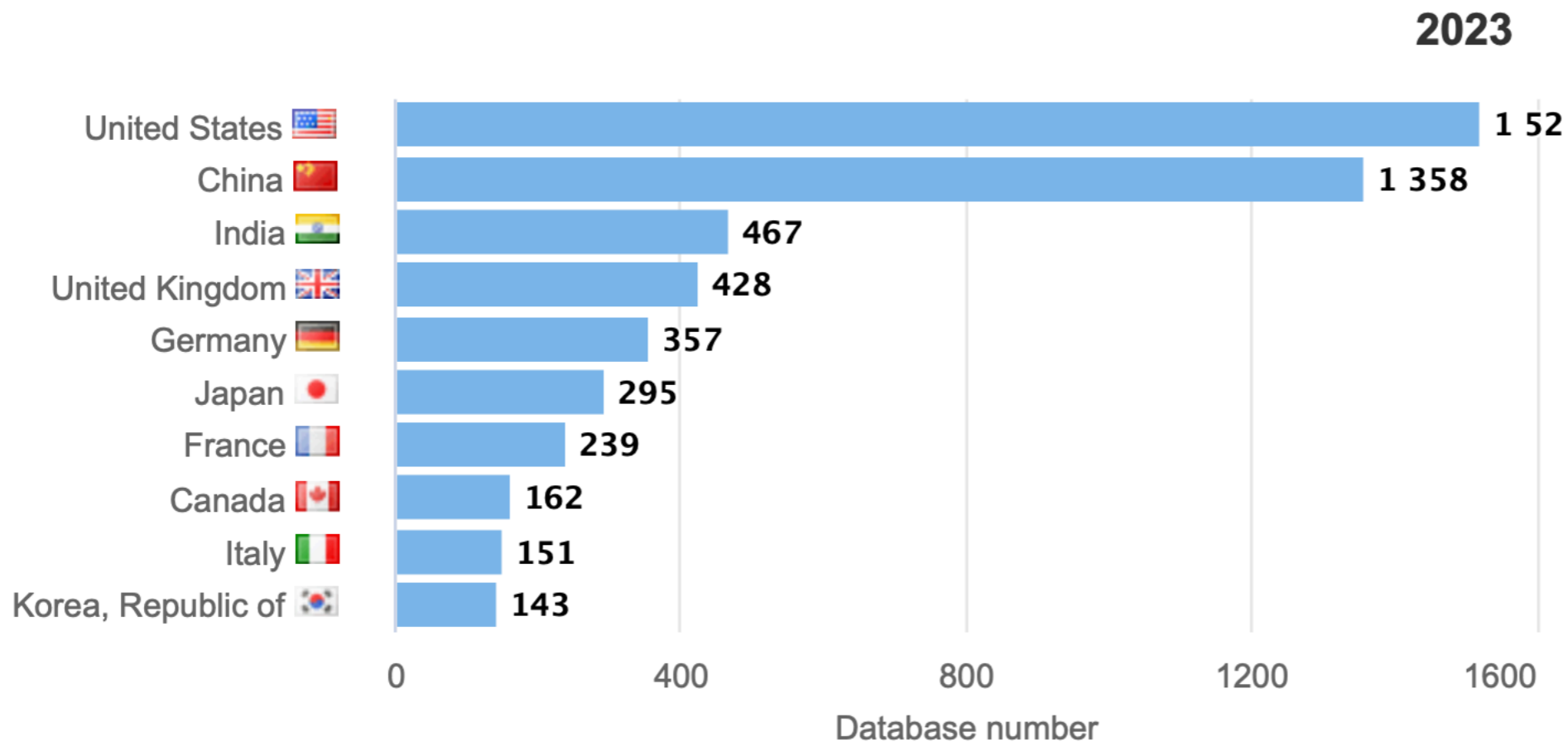
76
COUNTRIES / REGIONS

<https://ngdc.cncb.ac.cn/databasecommons/>

We can find the answer in...



Annual increment of global biological databases



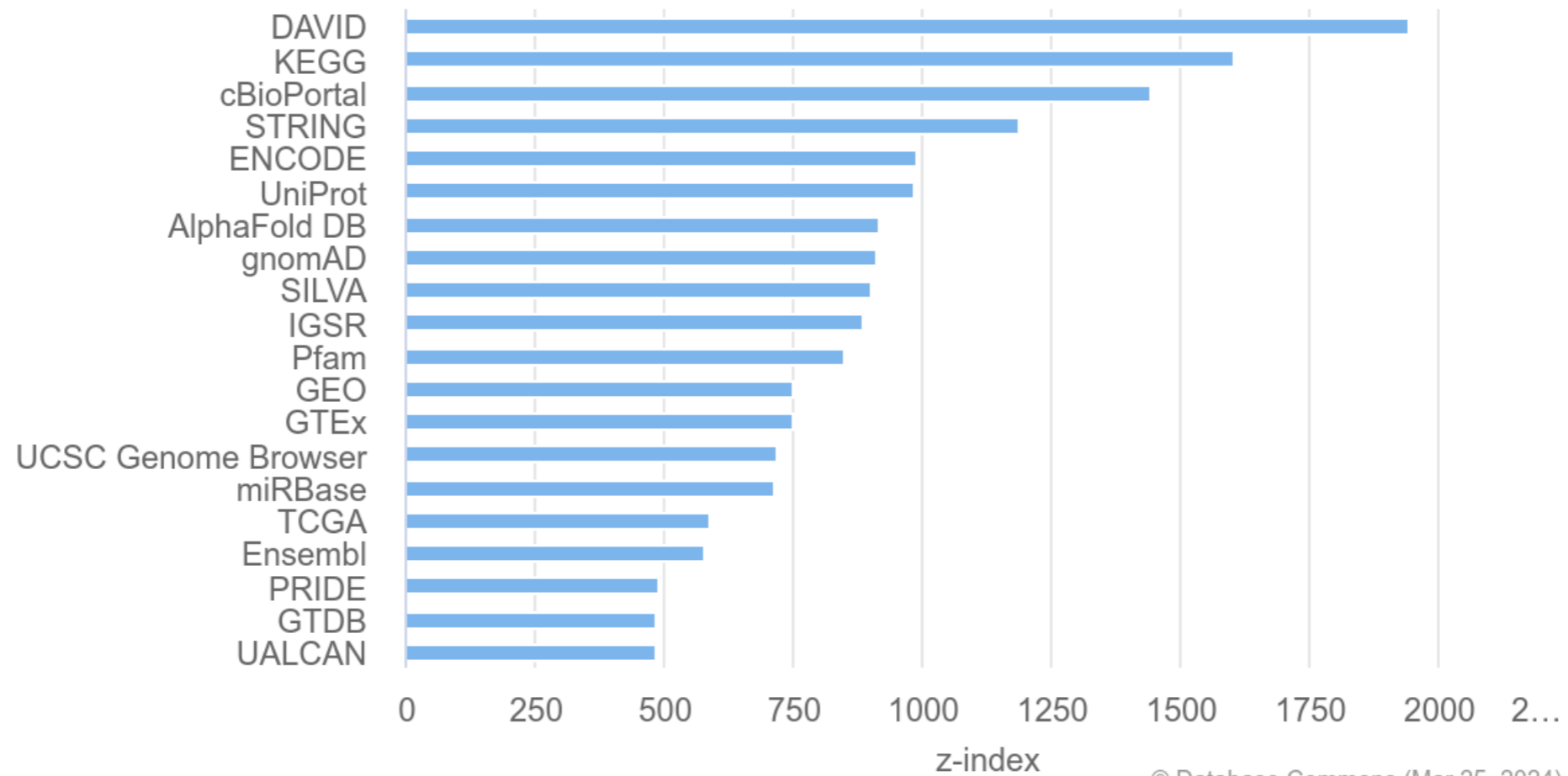
© Database Commons (Mar 13, 2024)
National Genomics Data Center

<https://ngdc.cncb.ac.cn/databasecommons/>

We can find the answer in...



Top 20 databases by *z-index*



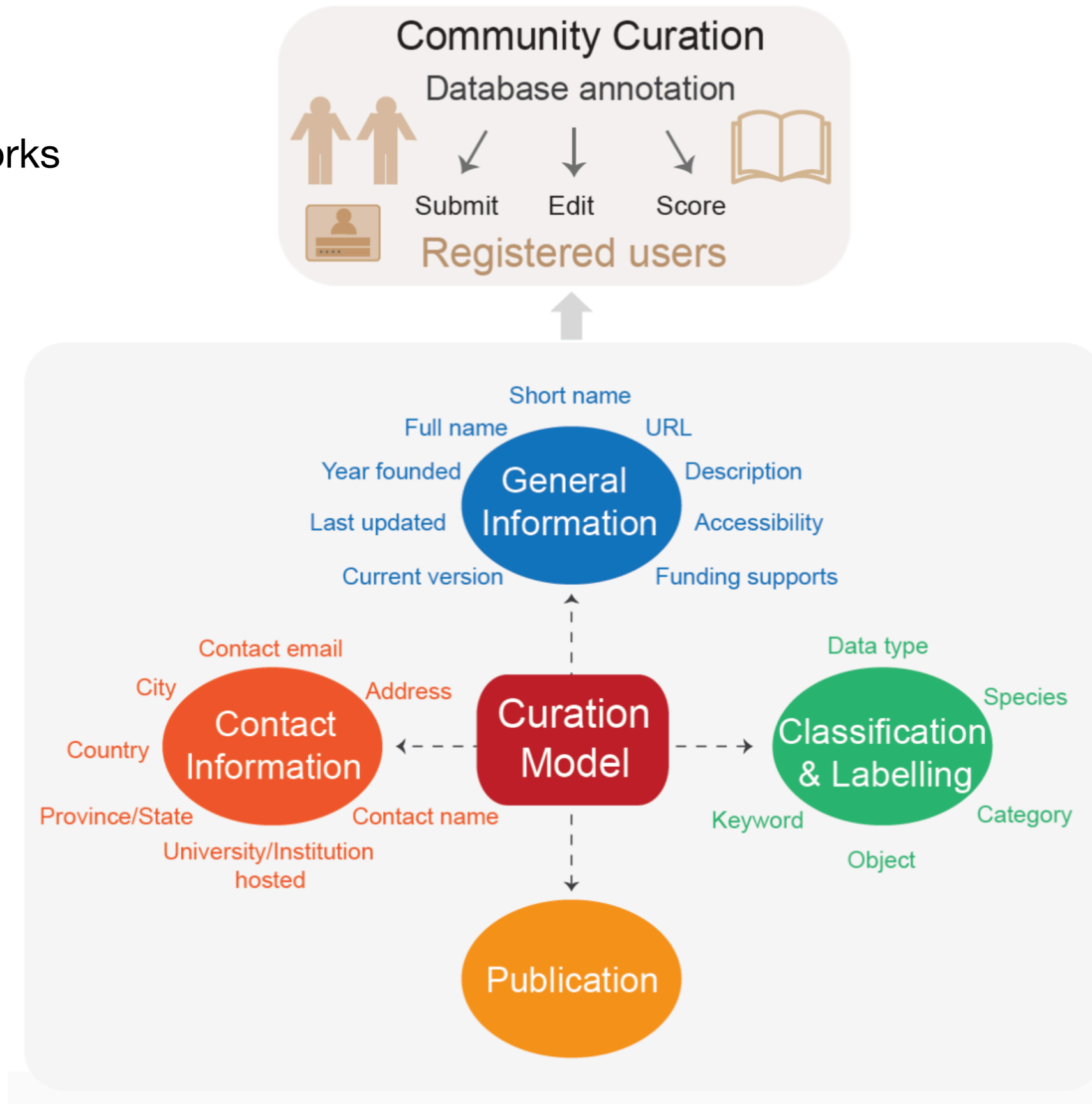
© Database Commons (Mar 25, 2024)
National Genomics Data Center

Z-index: Number of citations/Database age

<https://ngdc.cncb.ac.cn/databasecommons/>

Database Commons

How it works



Database Commons

Classification and labelling

Databases are classified based on data type, data object and database subjects. In addition, major species and keywords are tagged to further indicate the specific fields the database is related with.

- **Data Object**

A database may encompass multiple data objects. In Database Commons, there are a total of 6 data objects as detailed below.

- 1 **Animal**
- 2 **Plant**
- 3 **Fungi**
- 4 **Bacteria**
- 5 **Archaea**
- 6 **Virus**

Database Commons

Classification and labelling

- **Data Type**

A database may encompass multiple data types. In Database Commons, there are a total of 3 data types as detailed below.

- 1 **DNA:** gene/chromosome/genome sequence, DNA mutation/modification, DNA structure, DNA elements including probe, primer, motif, repeat sequence, etc.
- 2 **RNA:** RNA sequence, coding & non-coding transcripts, alternative splicing, RNA editing/modification, RNA probe and primer, RNA motif and structure, RNA expression
- 3 **Protein:** protein sequence, protein motif and domain, protein structure, protein modification, protein-protein interaction, protein expression

Database Commons

Classification and labelling

- **Database Category**

A database may encompass multiple database categories. In Database Commons, there are a total of 13 database categories as detailed below.

- 1 **Raw bio-data:** raw data of nucleic acid/protein sequencing and microarray, and image, digit, video, audio from biological and medical research
- 2 **Gene, genome and annotation:** gene/genetic element annotation, gene structure/family/motif/domain annotation, genome annotation, comparative genome (metagenome, pan-genome) analysis and annotation
- 3 **Genotype, phenotype and variation:** genotypes, phenotypes, multiple-scale variations (including SNP, INDEL, CNV, chromosomal rearrangement and other structural variation), genotype-phenotype associations
- 4 **Phylogeny and homology:** phylogeny reconstruction of genes/species, evolutionary history/process/event among individuals/organisms, homology identification
- 5 **Expression:** RNA/protein expression, expression abundance and pattern, RNA probe or primer used for gene expression detection, differential expression analysis
- 6 **Modification:** DNA modification, post-transcriptional modification of mRNA and non-coding RNA, post-translational modification of protein, modification type/technology/function
- 7 **Structure:** secondary, tertiary and quaternary structure of DNA/RNA/protein, chromatin structure
- 8 **Interaction:** direct (physical) and indirect (functional) associations, including protein-protein interaction, RNA-protein interaction, DNA-protein interaction, gene regulatory interaction, biochemical reaction, antigen and antibody, and genetic interaction
- 9 **Pathway:** biological pathways for metabolic, signaling, gene regulatory analysis
- 10 **Health and medicine:** disease variation/genotype-phenotype association, immune reaction, disease model, clinical biomarker, therapeutic target, drug & chemical compound, pharmacogenomics and pharmacodynamics, electronic health record
- 11 **Standard, ontology and nomenclature:** standard, ontology and nomenclature for biological entities
- 12 **Literature:** literature information, literature/text mining, textual annotation based on literature
- 13 **Metadata:** metadata information for biological entities, e.g., project/sample/experiment/run/database/tool

<https://ngdc.cncb.ac.cn/databasecommons/>

Databases we will see in more detail

We will explore:

- UCSC Genome Browser
- Kyoto Encyclopedia of Genes and Genomes (KEGG)
- The Cancer Genome Browser (TCGA)
- cBioPortal
- The Human Protein Atlas
- Known and Predicted Protein-Protein Interactions (STRING)

UCSC Genome Browser

UCSC Genome Browser

General information

URL: <https://genome.ucsc.edu>

Full name: University of California Santa Cruz Genome Browser Database

Description: Launched in 2001 to showcase the draft human genome assembly, the UCSC Genome Browser database (<http://genome.ucsc.edu>) and associated tools continue to grow, providing a comprehensive resource of genome assemblies and annotations to scientists and students worldwide.

Year founded: 2002

Last update: 2022-04-02

Version:

Accessibility: Manual: Accessible Real time  : Accessible

Country/Region: [United States](#)

Classification & Tag

Data type: [DNA](#) | [Protein](#) | [RNA](#) |

Data object: [Animal](#) |

Database category: [Gene genome and annotation](#) |

Major species: [Homo sapiens](#) | [Mus musculus](#) |

Keywords: [genome assembly](#) | [genome browser](#) | [genome alignment](#) | [SARS-CoV-2](#) |

Ranking

All databases:
14/6000 (99.783%)

14
TOTAL RANK

Gene genome and
annotation:
6/1675 (99.701%)


15,855
CITATIONS 

720.682
Z-INDEX 

Community reviews

 5 Stars (1)

Data quality & quantity: 

Content organization & presentation 

System accessibility & reliability: 

 [Submit a review](#)

Word cloud

hamster graphs tripled 100-species web' stacked
whale providing showcase created students
UCSC Genome Browser Author browsers years
sharkhubssequences past sharing database: increasingly
Box assemblies track annotation tools
hg38 multiple database year genom
mink color align incl s Browser year
CDK20 community 1 align as incl s G de



Tools



hg38



hg19



mm39

- **Genome Browser** - Interactively visualize genomic data
- **BLAT** - Rapidly align sequences to the genome
- **In-Silico PCR** - Rapidly align PCR primer pairs to the genome
- **Table Browser** - Download and filter data from the Genome Browser
- **LiftOver** - Convert genome coordinates between assemblies
- **REST API** - Returns data requested in JSON format
- **Variant Annotation Integrator** - Annotate genomic

News

Mar. 07, 2024 - [New Prediction Scores super track and BayesDel track](#)

Mar. 05, 2024 - [New JASPAR tracks: Human \(hg19/hg38\) - Mouse \(mm39\)](#)

Mar. 01, 2024 - [AbSplice Prediction Scores for hg38](#)

Feb. 21, 2024 - [New DECIPHER Dosage Sensitivity tracks for Human](#)

Feb. 14 2024 - [New GENCODE gene tracks: V45 \(hg19/hg38\) - VM34 \(mm39\)](#)

Feb. 12, 2024 - [Variants of Concern SARS-CoV-2 track updated with \(mm39\)](#)

[More news...](#)[Subscribe](#)

Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG

Databases

Tools

Auto annotation

Kanehisa Lab



KEGG



Search

Help

» Japanese

KEGG Home

[Release notes](#)

[Current statistics](#)

KEGG Database

[KEGG overview](#)

[KEGG mapping](#)

[Color codes](#)

KEGG Objects

[KEGG Weblinks](#)

[Entry format](#)

KEGG Software

[KEGG API](#)

[KGML](#)

KEGG FTP

[Subscription](#)

[Background info](#)

[GenomeNet](#)

[DBGET/LinkDB](#)

[Feedback](#)

[Copyright request](#)

[Kanehisa Labs](#)

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (April 1, 2024) for new and updated features.

New article [KEGG tools for classification and analysis of viral proteins](#)

● Main entry point to the KEGG web service

KEGG2

[KEGG Table of Contents](#) [[Update notes](#) | [Release history](#)]

● Data-oriented entry points

KEGG PATHWAY

[KEGG pathway maps](#)

KEGG BRITE

[BRITE hierarchies and tables](#)

KEGG MODULE

[KEGG modules](#)

KEGG ORTHOLOGY

[KO functional orthologs](#)

KEGG GENES

[Genes and proteins](#) [[SeqData](#)]

KEGG GENOME

[Genomes](#) [[KEGG Virus](#) | [Syntax](#)]

KEGG COMPOUND

[Small molecules](#)

KEGG GLYCAN

[Glycans](#)

KEGG REACTION

[Biochemical reactions](#) [[RModule](#)]

KEGG ENZYME

[Enzyme nomenclature](#)

KEGG NETWORK

[Disease-related network variations](#)

KEGG DISEASE

[Human diseases](#)

KEGG DRUG

[Drugs](#) [[New drug approvals](#)]

KEGG MEDICUS

[Health information resource](#) [[Drug labels search](#)]

● Organism-specific entry points

KEGG Organisms

[Enter org code\(s\)](#)

[Go](#)

[hsa](#)

[hsa eco](#)

[Pathway](#)

[Brite](#)

[Brite table](#)

[Module](#)

[Network](#)

[KO \(Function\)](#)

[Organism](#)

[Virus](#)

[Compound](#)

[Disease \(ICD\)](#)

[Drug \(ATC\)](#)

[Drug \(Target\)](#)

[Antimicrobials](#)

<https://www.genome.jp/kegg/>

The Cancer Genome Atlas (TCGA)



General information

URL:	https://portal.gdc.cancer.gov
Full name:	The Cancer Genome Atlas
Description:	A landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types.
Year founded:	2013
Last update:	
Version:	
Accessibility:	Manual: Accessible Real time : Accessible
Country/Region:	United States

Classification & Tag

Data type:	DNA Protein RNA
Data object:	NA
Database category:	Expression Genotype phenotype and variation Health and medicine Modification Structure
Major species:	Homo sapiens
Keywords:	cancer human

Ranking

All databases:	16/6000 (99.75%)
Genotype phenotype and variation:	5/852 (99.531%)
Expression:	4/1143 (99.738%)
Structure:	5/841 (99.524%)
Health and medicine:	2/1394 (99.928%)
Modification:	2/287 (99.652%)

16
TOTAL RANK

6,448
CITATIONS

586.182
Z-INDEX

Community reviews

Not Rated

Data quality & quantity:

Content organization & presentation

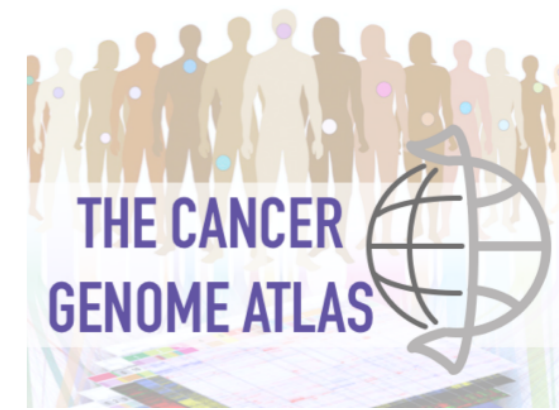
System accessibility & reliability:

[Submit a review](#)

Word cloud

nu

The Cancer Genome Atlas (TCGA)



Genomic Data Commons Data Portal

Harmonized Cancer Datasets

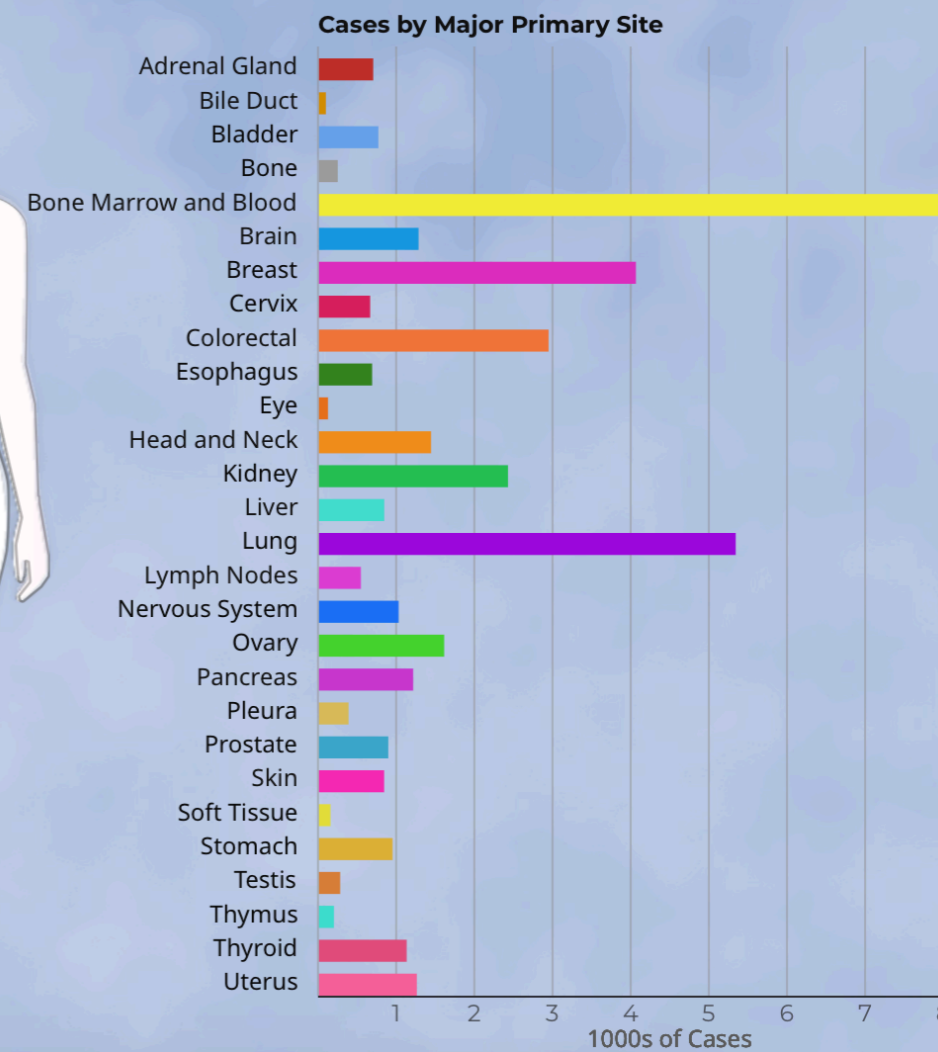
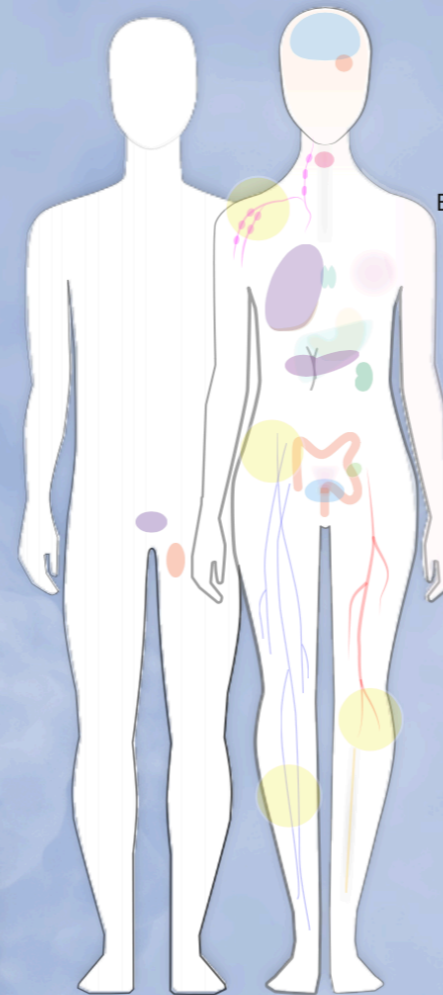
A repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy.

Explore Our Cancer Datasets

Data Portal Summary

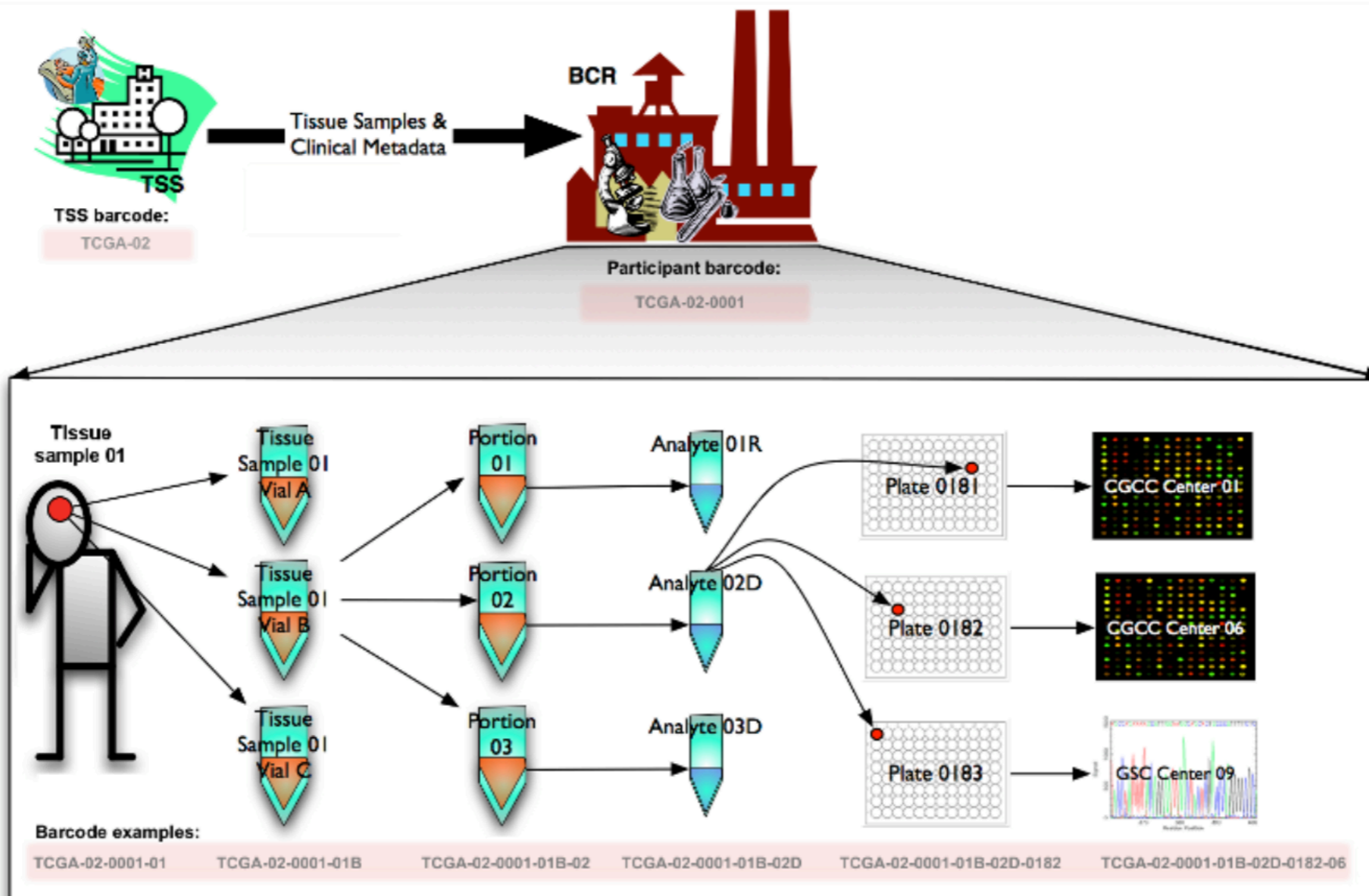
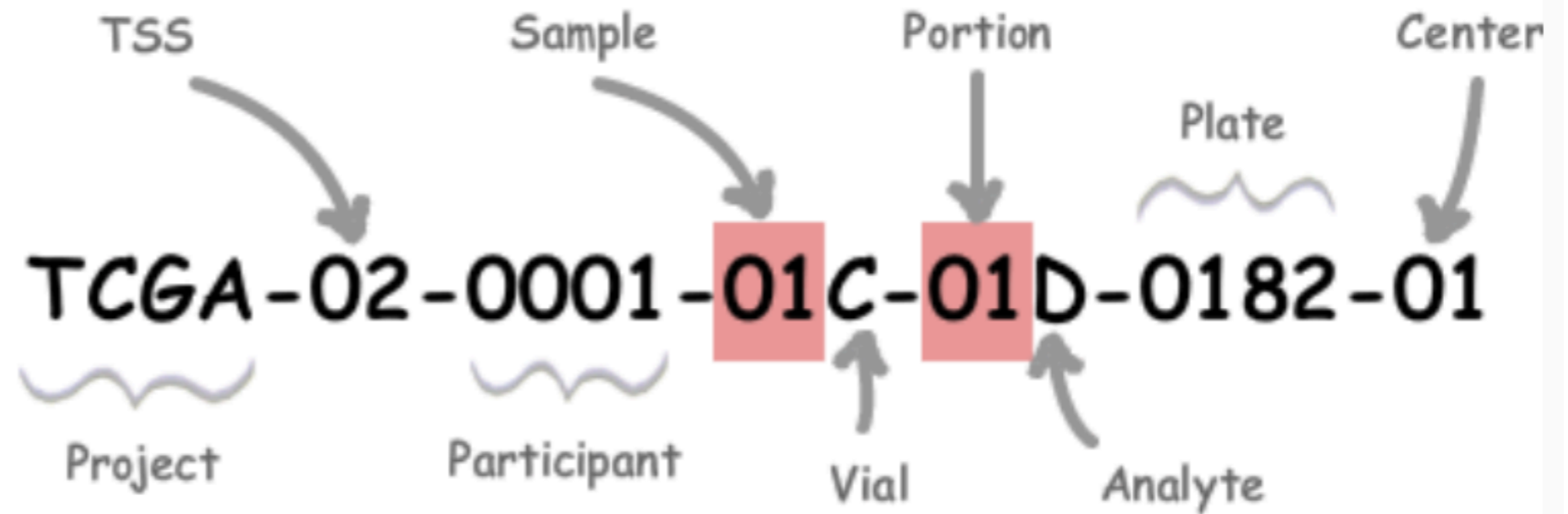
Data Release 39.0 - December 04, 2023

79 Projects	69 Primary Sites	44,451 Cases	986,114 Files	22,534 Genes	2,930,136 Mutations



TCGA barcodes

Each sample (i.e. sequencing) presents a unique barcode that describe the patient, tissue, molecule etc.



TCGA barcodes

Label	Identifier for	Value	Value Description	Possible Values
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	182	The 182nd plate	4-digit alphanumeric value
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	1	The first portion of the sample	01-99
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Project	Project name	TCGA	TCGA project	TCGA
Sample	Sample type	1	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Center	Sequencing or characterization center that will receive the aliquot for analysis	1	The Broad Institute GCC	See Code Tables Report
Participant	Study participant	1	The first participant from MD Anderson for GBM study	Any alpha-numeric value
TSS	Tissue source site	2	GBM (brain tumor) sample from MD Anderson	See Code Tables Report

<https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations>

Study Abbreviation	Study Name
LAML	Acute Myeloid Leukemia
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
LGG	Brain Lower Grade Glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
LCML	Chronic Myelogenous Leukemia
COAD	Colon adenocarcinoma
CNTL	Controls
ESCA	Esophageal carcinoma
FPPP	FFPE Pilot Phase II
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
MESO	Mesothelioma
MISC	Miscellaneous
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THYM	Thymoma
THCA	Thyroid carcinoma
UCS	Uterine Carcinosarcoma
UCEC	Uterine Corpus Endometrial Carcinoma
UVM	Uveal Melanoma

<https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes>

Code	Definition	Short Letter Code
01	Primary Solid Tumor	TP
02	Recurrent Solid Tumor	TR
03	Primary Blood Derived Cancer - Peripheral Blood	TB
04	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
05	Additional - New Primary	TAP
06	Metastatic	TM
07	Additional Metastatic	TAM
08	Human Tumor Original Cells	THOC
09	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
15	sample type 15	15SH
16	sample type 16	16SH
20	Control Analyte	CELLC
40	Recurrent Blood Derived Cancer - Peripheral Blood	TRB
50	Cell Lines	CELL
60	Primary Xenograft Tissue	XP
61	Cell Line Derived Xenograft Tissue	XCL
99	sample type 99	99SH

General information

URL: <http://cbioportal.org>

Full name: cBio Cancer Genomics Portal

Description: cBio Cancer Genomics Portal is an open-access resource for interactive exploration of multidimensional cancer genomics data sets, currently providing access to data from more than 5,000 tumor samples from 20 cancer studies.

Year founded: 2012

Last update: 2021-10-05

Version: 3.7.9

Accessibility: Manual: **Accessible** Real time : **Accessible**

Country/Region: [United States](#)

Classification & Tag

Data type: [DNA](#) | [Protein](#) | [RNA](#) |

Data object: [Animal](#) |

Database category: [Expression](#) | [Genotype phenotype and variation](#) | [Health and medicine](#) | [Modification](#) | [Structure](#) |

Major species: [Homo sapiens](#) |

Keywords: [tumor](#) | [cancer genome](#) | [TCGA](#) |

Ranking

All databases: **3/6000 (99.967%)**

Genotype phenotype and variation: **1/852 (100%)**

Expression: **1/1143 (100%)**

Modification: **1/287 (100%)**

Structure: **1/841 (100%)**

Health and medicine: **1/1394 (100%)**

3

TOTAL RANK

17,287
CITATIONS

1,440.58
Z-INDEX

Community reviews

Not Rated

Data quality & quantity:

Content organization & presentation

System accessibility & reliability:

[Submit a review](#)

Word cloud

platforms survival queries graphical gene-level er multi UN

Query

Quick Search **Beta!**

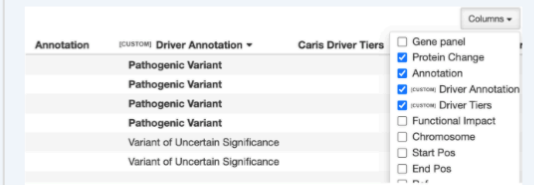
Please cite [cBioPortal](#)

What's New

@cbioportal

Mar 5, 2024

Local Installations Feature: Make representation of custom driver annotation configurable:



Feb 7, 2024

New feature: The *Datasets Page* now lists what studies have samples profiled for Structural Variants:

Sign up for low-volume email news alerts

Subscribe

Example Queries

- Primary vs. metastatic prostate cancer
- RAS/RAF alterations in colorectal cancer
- BRCA1 and BRCA2 mutations in ovarian cancer
- POLE hotspot mutations in endometrial cancer
- TP53 and MDM2/4 alterations in GBM
- PTEN mutations in GBM in text format
- Patient view of an endometrial cancer case
- All TCGA Pan-Cancer
- MSK-IMPACT clinical cohort, Zehir et al. 2017
- Histone mutations across cancer types

Local Installations

Host your own

Select Studies for Visualization & Analysis:

0 studies selected (0 samples)

Data type

Search...

Help

PanCancer Studies 10

Pediatric Cancer Studies 14

Immunogenomic Studies 8

Cell lines 3

Adrenal Gland 2

Ampulla of Vater 1

Biliary Tract 15

Bladder/Urinary Tract 21

Bone 3

Bowel 19

Breast 26

CNS/Brain 26

Cervix 3

Quick select:

TCGA PanCancer Atlas Studies

Curated set of non-redundant studies

Looking for **AACR Project GENIE**, the largest public clinicogenomic cancer dataset? [It's available here.](#)

PanCancer Studies

- MSK-IMPACT Clinical Sequencing Cohort (MSK, Nat Med 2017) 10945 samples
- Metastatic Solid Cancers (UMich, Nature 2017) 500 samples
- MSS Mixed Solid Tumors (Broad/Dana-Farber, Nat Genet 2018) 249 samples
- SUMMIT - Neratinib Basket Study (Multi-Institute, Nature 2018) 141 samples
- TMB and Immunotherapy (MSK, Nat Genet 2019) 1661 samples
- Tumors with TRK fusions (MSK, Clin Cancer Res 2020) 106 samples
- Cancer Therapy and Clonal Hematopoiesis (MSK, Nat Genet 2020) 24146 samples
- China Pan-cancer (Origimed, Nature 2022) 10194 samples
- Pan-cancer analysis of whole genomes (ICGC/TCGA, Nature 2020) 2922 samples
- MSK MetTropism (MSK, Cell 2021) 25775 samples

Pediatric Cancer Studies

- Pediatric Preclinical Testing Consortium (CHOP, Cell Rep 2019) 261 samples
- Pediatric Acute Lymphoid Leukemia - Phase II (TARGET, 2018) 1978 samples
- Pediatric Rhabdoid Tumor (TARGET, 2018) 72 samples
- Pediatric Wilms' Tumor (TARGET, 2018) 657 samples
- Pediatric Acute Myeloid Leukemia (TARGET, 2018) 1025 samples
- Pediatric Neuroblastoma (TARGET, 2018) 1089 samples

0 studies selected (0 samples)

Query By Gene

OR

Explore Selected Studies

The Human Protein Atlas

The Human Protein Atlas

General information

URL: <http://www.proteinatlas.org>

Full name:

Description: The Human Protein Atlas project creates a map of protein expression patterns in normal cells, tissues and cancer. In this database, 11,200 unique proteins corresponding to over 50% of all human protein-encoding genes have been analysed.

Year founded: 2011

Last update:

Version:

Accessibility: Manual: **Accessible** Real time  : **Accessible**

Country/Region: [Sweden](#)

Classification & Tag

Data type: [Protein](#)

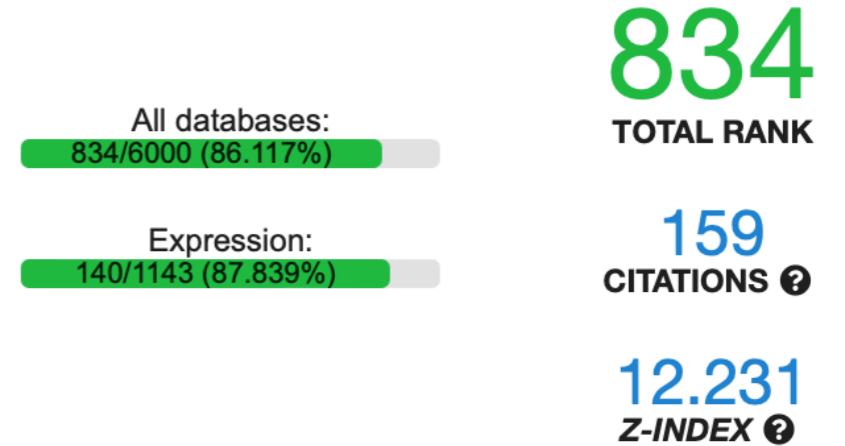
Data object: [Animal](#)

Database category: [Expression](#)

Major species: [Homo sapiens](#)

Keywords: [protein expression patterns](#) | [biomarker discovery for cancer](#)

Ranking



Community reviews

★ ★ ★ ★ ★ Not Rated

Data quality & quantity: ★ ★ ★ ★ ★

Content organization & presentation ★ ★ ★ ★ ★

System accessibility & reliability: ★ ★ ★ ★ ★

 [Submit a review](#)

Word cloud



THE HUMAN PROTEIN ATLAS

SECTIONS ABOUT NEWS LEARN DATA HELP

The open access resource for human proteins

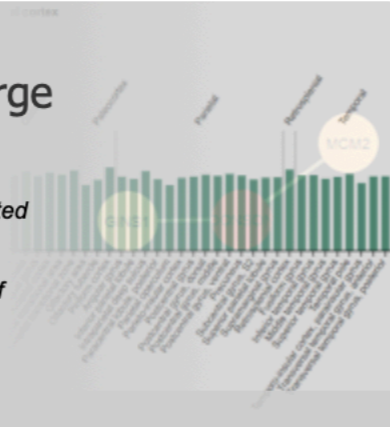
Search for specific genes/proteins or explore the 12 different sections

 [Fields »](#)
e.g. ACE2, GFAP, EGFR Search helpⁱ

News webinar Young researcher awarded large grant for FTD research

The postdoc Emma Gerrits in Jan Mulders HPA related research group at KI has recently, as the youngest person ever, received a 1.2 MSEK grant from Alzheimerfonden for research on the development of Frontotemporal dementia (FTD).... [Read more](#)

read the latest article - published Wed, 13 Mar 2024



TMEM253 - a "Gene Doe" of the intestine

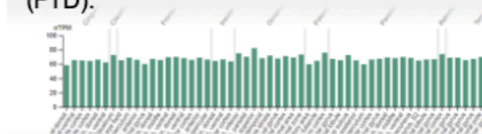
Here we want to draw some attention to one of the rather unknown but interesting genes that we have encountered, the TMEM253 gene specifically expressed in the intestine.



Wed, 6 Mar 2024

Young researcher awarded large grant for FTD research

The postdoc Emma Gerrits in Jan Mulders HPA related research group at KI has recently, as the youngest person ever, received a 1.2 MSEK grant from Alzheimerfonden for research on the development of Frontotemporal dementia (FTD).



Wed, 13 Mar 2024

[all news articles](#)

TISSUE

Protein and RNA profiles in tissues based on antibodies and transcriptomics

BRAIN

Protein and RNA profiles in brain based on microdissected regions

SINGLE CELL TYPE

Single cell RNA profiles in tissues based on single cell RNA analysis

TISSUE CELL TYPE

Cell type profiles in tissues based on deconvolution of bulk transcriptomics

PATHOLOGY

Protein and RNA profiles in human cancers based on antibodies and transcriptomics

DISEASE

Protein levels in blood in patients with diseases

IMMUNE CELL

BLOOD PROTEIN

SUBCELLULAR

CELL LINE

STRUCTURE

INTERACTION

```
graph TD; SRS3 --- FAM120A; SRS3 --- KLRC4; SRS3 --- RADA; SRS3 --- C2F3; SRS3 --- PAP12
```


STRING



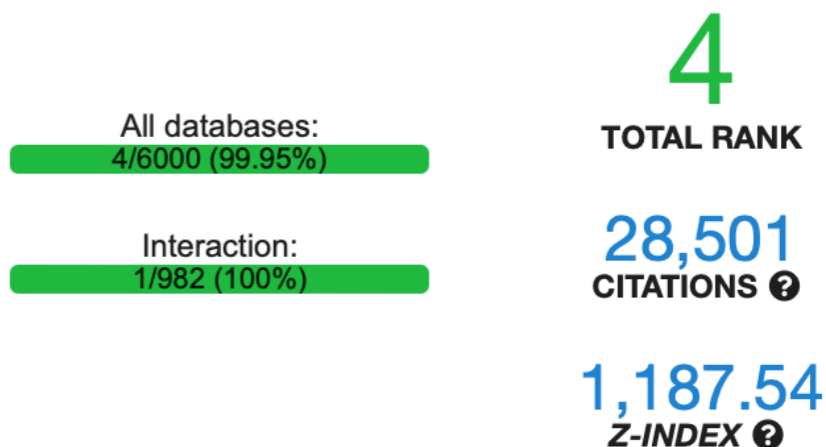
General information

URL:	https://string-db.org/
Full name:	Known and Predicted Protein-Protein Interactions
Description:	STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations
Year founded:	2000
Last update:	
Version:	v11.5
Accessibility:	Manual: Accessible Real time ? :
Country/Region:	Switzerland

Classification & Tag

Data type:	Protein
Data object:	Animal Archaea Bacteria Fungi Plant
Database category:	Interaction
Major species:	Homo sapiens Methanococcus voltae Arabidopsis thaliana
Keywords:	protein-protein interaction (PPI)

Ranking



Community reviews



Data quality & quantity: ★★★★★
Content organization & presentation: ★★★★★
System accessibility & reliability: ★★★★★

[Submit a review](#)

Word cloud



STRING

1. Go to:

**[https://string-db.org/cgi/input?
sessionId=buUIHtE4Ehj7&input_page_active_form=multiple_identifiers](https://string-db.org/cgi/input?sessionId=buUIHtE4Ehj7&input_page_active_form=multiple_identifiers)**

2. Copy and paste the following gene names in the “List of names” panel:

FOXA1

AR

MYC

HOXB13

KRAS

TP53

NRAS

HRAS

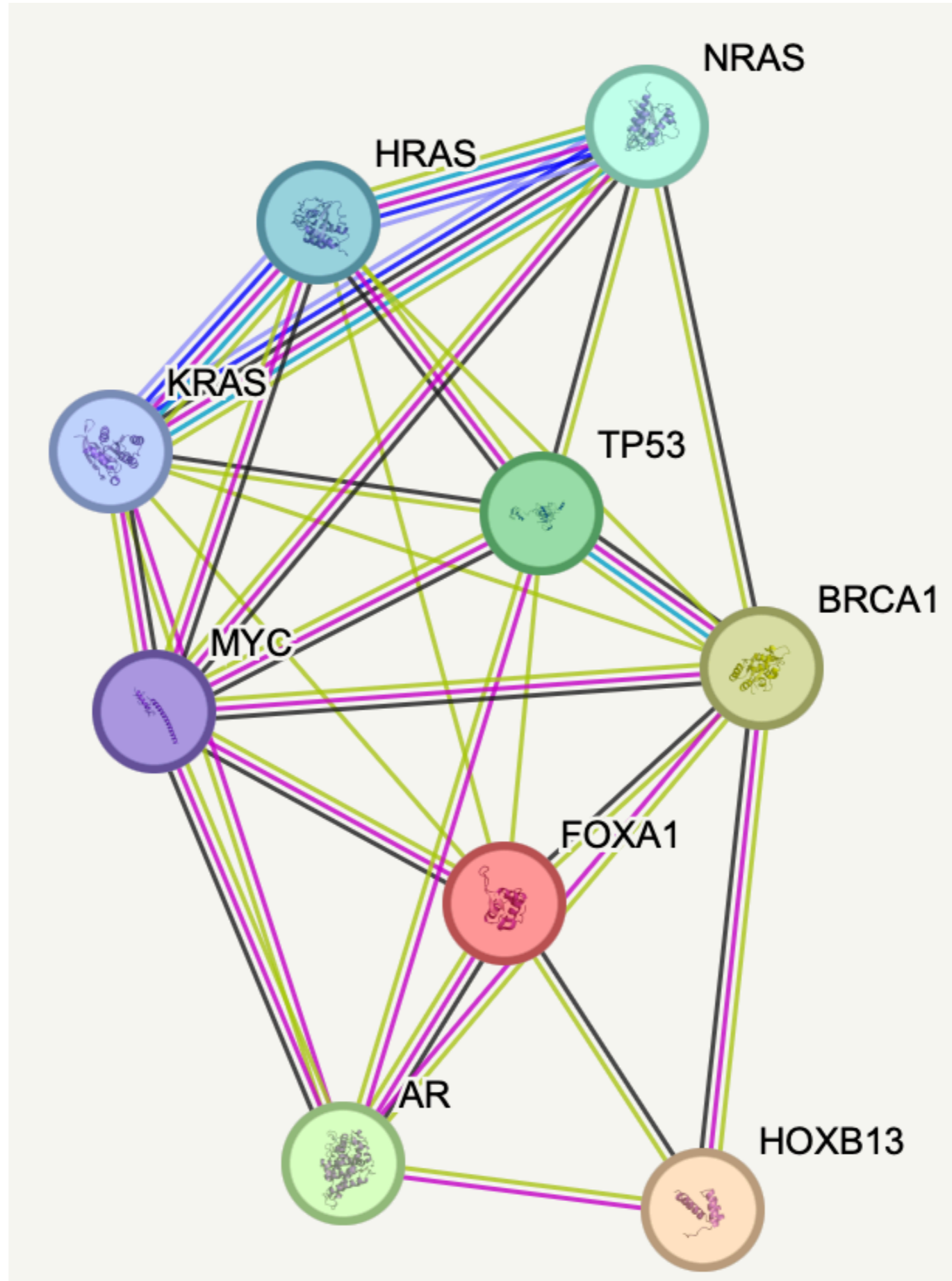
NCOR

BRCA1

3. Choose the organism “Homo sapiens”

STRING

4. Explore the result



R packages associated to databases

<https://bioconductor.org/packages/release/data/annotation/html/UCSCRepeatMasker.html>

<https://bioconductor.org/packages/release/data/experiment/html/msigdb.html>

<https://bioconductor.org/packages/release/bioc/html/cBioPortalData.html>

<https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html>

<https://bioconductor.org/packages/release/bioc/html/hpar.html>

<https://www.bioconductor.org/packages/release/bioc/html/STRINGdb.html>

...