# BCB 5200 Introduction to Bioinformatics

**Lecture 03: NCBI Databases**

Bioinformatics and Computational Biology
Saint Louis University

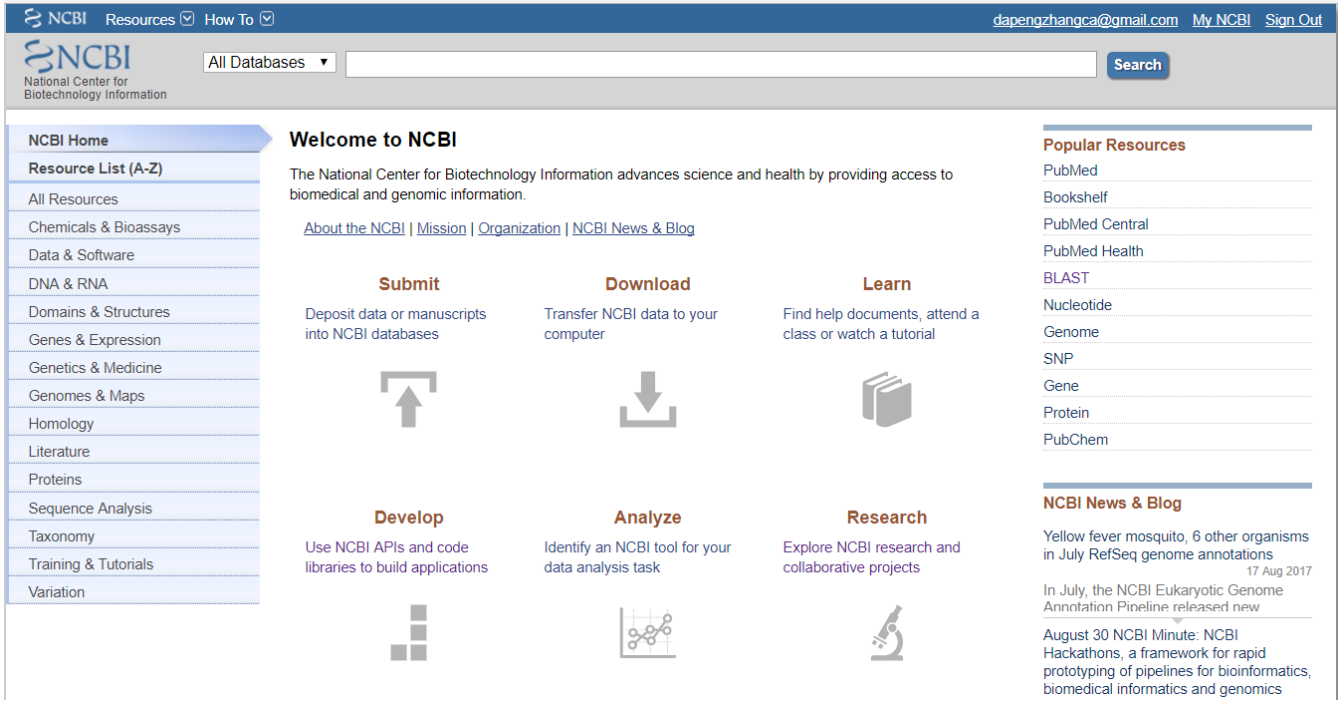
1. How to quickly identify the gene expression patterns of a particular gene?

Maelstrom

C10orf99

Variations of the data

# NCBI: https://www.ncbi.nlm.nih.gov/




- Develop and maintain molecular and bibliographic databases.
- Develop software for searching, and analysis of these data.
- Provide Web access point for data and software.


# Molecular Databases in NCBI

- Sequences: Nucleotide, Protein
- Gene annotation: Gene, RefSeq
- Genomes: Assembly, Genome
- Expression: GEO, EST
- Protein Domains: CDD
- Homologous Genes
- Genetic Variation: SNP, ClinVar, dbVar
- Taxonomy
- 3D Structures
- Pathways
- Literature: PubMed, PMC
- Small molecules: PubChem


## NCBI Search Services and Tools

- Entrez integrated literature and molecular databases
- BLAST sequence similarity search service
- Graphical Sequence Viewer annotation viewer and analysis tool
- Genome Workbench standalone sequence analysis annotation platform
- SRA Utilities
  - SRA Run Browser: web access for viewing, searching and downloading next generation reads
  - SRA toolkit: standalone SRA manipulator and client

# Entrez is NCBI's primary text search and retrieval system

**Search NCBI databases**

all[sb] Search

**About 1,353,092,513 search results for "all[sb]"**

## Literature

| | | |
|---|---|---|
| Books | 334,634 | books and reports |
| MeSH | 253,991 | ontology used for PubMed indexing |
| NLM Catalog | 1,511,375 | books, journals and more in the NLM Collections |
| PubMed | 24,324,879 | scientific & medical abstracts/citations |
| PubMed Central | 3,264,562 | full-text journal articles |

## Health

| | | |
|---|---|---|
| ClinVar | 125,404 | human variations of clinical significance |
| dbGaP | 166,753 | genotype/phenotype interaction studies |
| GTR | 35,977 | genetic testing registry |
| MedGen | 260,910 | medical genetics literature and links |
| OMIM | 23,795 | online mendelian inheritance in man |
| PubMed Health | 51,412 | clinical effectiveness, disease and drug reports |

## Genomes

| | | |
|---|---|---|
| Assembly | 33,680 | genomic assembly information |
| BioProject | 138,999 | biological projects providing data to NCBI |
| BioSample | 2,865,750 | descriptions of biological source materials |
| Clone | 37,024,042 | genomic and cDNA clones |
| dbVar | 4,305,990 | genome structural variation studies |
| Epigenomics | 6,635 | epigenomic studies and display tools |
| Genome | 10,777 | genome sequencing projects by organism |
| GSS | 37,646,655 | genome survey sequences |
| Nucleotide | 155,458,773 | DNA and RNA sequences |
| Probe | 31,890,151 | sequence-based probes and primers |
| SNP | 444,458,769 | short genetic variations |
| SRA | 1,064,488 | high-throughput DNA and RNA sequence read archive |
| Taxonomy | 1,310,804 | taxonomic classification and nomenclature catalog |

## Genes

| | | |
|---|---|---|
| EST | 75,736,497 | expressed sequence tag sequences |
| Gene | 18,350,280 | collected information about gene loci |
| GEO DataSets | 1,332,018 | functional genomics studies |
| GEO Profiles | 108,708,851 | gene expression and molecular abundance profiles |
| HomoloGene | 141,268 | homologous gene sets for selected organisms |
| PopSet | 212,126 | sequence sets from phylogenetic and population studies |
| UniGene | 6,473,284 | clusters of expressed transcripts |

## Proteins

| | | |
|---|---|---|
| Conserved Domains | 49,955 | conserved protein domains |
| Protein | 153,454,195 | protein sequences |
| Protein Clusters | 820,546 | sequence similarity-based protein clusters |
| Structure | 103,644 | experimentally-determined biomolecular structures |

## Chemicals

| | | |
|---|---|---|
| BioSystems | 653,443 | molecular pathways with links to genes, proteins and chemicals |
| PubChem BioAssay | 1,112,105 | bioactivity screening studies |
| PubChem Compound | 62,041,347 | chemical information with structures, information and links |
| PubChem Substance | 177,330,151 | deposited substance and chemical information |

The Entrez system: 39 (and counting) integrated databases

# The Syntax …

- **Key words search**
- **Boolean operators**: AND, OR, NOT must be entered in **UPPERCASE** (e.g., promoters OR response elements). The default is AND.
- **Parentheses:** Entrez processes all Boolean operators in a **left-to-right** sequence. You can change the order by enclosing individual concepts in **parentheses**. The terms inside the parentheses are processed first.
  - For example, the search statement: g1p3 OR (response AND element AND promoter).
- **Quotation marks:** The term inside the quotation marks is read as one phrase (e.g. "public health" is different than public health, which will also include articles on public latrines and their effect on health workers).
- **Asterisk**: Extends the search to all terms that start with the letters before the asterisk.
  - cilio* will include such terms as ciliopathy, ciliopathies, and ciliogenesis.

# The easiest Entrez search in Gene database

**Specific gene:**

XXX[Symbol] AND YYY[Organism]

```
MAEL[symbol] AND human[organism]
apt[sym] AND Escherichia coli[orgn]
```

**All genes:**

YYY[organism]

```
zebrafish[orgn]
Zea mays[orgn]
```

# NCBI-Nucleotide/Protein

- MAEL[All Fields] AND "Homo sapiens"[Organism]

NCBI Resources How To dapengzhangca@gmail.com My NCBI Sign Out

Protein | Protein | MAEL AND human[organism] | Search

Create alert Advanced Help

**Species**
Animals (23)
Customize ...

**Source databases**
RefSeq (8)
UniProtKB / Swiss-Prot (7)
Customize ...

**Sequence length**
Custom range...

**Molecular weight**
Custom range...

**Release date**
Custom range...

**Revision date**
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

See the results of this search (2 items) in our new Identical Protein Groups database.

**Items: 1 to 20 of 23**

<< First < Prev Page 1 of 2 Next > Last >>

1. MAEL protein [Homo sapiens]
   156 aa protein
   Accession: AAH34310.1 GI: 71296887
   GenPept Identical Proteins FASTA Graphics
2. protein maelstrom homolog isoform 2 [Homo sapiens]
   403 aa protein
   Accession: NP_001273306.1 GI: 556503424
   GenPept Identical Proteins FASTA Graphics
3. protein maelstrom homolog isoform 3 [Homo sapiens]
   378 aa protein
   Accession: NP_001273307.1 GI: 556503335
   GenPept Identical Proteins FASTA Graphics
4. protein maelstrom homolog isoform 1 [Homo sapiens]
   434 aa protein
   Accession: NP_116247.1 GI: 14249590
   GenPept Identical Proteins FASTA Graphics
5. RecName: Full=Protein maelstrom homolog
   434 aa protein
   Accession: Q96JY0.1 GI: 74760900
   GenPept Identical Proteins FASTA Graphics

**Filters:** Manage Filters

**Find related data**
Database: Select
Find items

**Search details**
MAEL[All Fields] AND "Homo sapiens" [Organism]
Search
See more...

**Recent activity**
Turn Off Clear
MAEL AND human[organism] (23) Protein
protein maelstrom homolog isoform 1 [Homo sapiens] Protein
MAEL[All Fields] AND human[organism] (23) Protein
MAEL maelstrom spermatogenic transposon silencer [Homo sapiens] Gene
MAEL[symbol] AND human[organism] AND (alive[prop]) (1) Gene

# NCBI-Nucleotide/Protein

https://www.ncbi.nlm.nih.gov/protein/NP_116247.1

## protein maelstrom homolog isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_116247.1

Identical Proteins FASTA Graphics

GenPept format

Go to:

```
LOCUS       NP_116247                434 aa            linear   PRI 03-JUN-2017
DEFINITION  protein maelstrom homolog isoform 1 [Homo sapiens].
ACCESSION   NP_116247
VERSION     NP_116247.1
DBSOURCE    REFSEQ: accession NM_032858.2
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 434)
  AUTHORS   Li Q, Wei P, Huang B, Xu Y, Li X, Li Y, Cai S and Li D.
  TITLE     MAEL expression links epithelial-mesenchymal transition and stem
            cell properties in colorectal cancer
  JOURNAL   Int. J. Cancer 139 (11), 2502-2511 (2016)
   PUBMED   27537253
  REMARK    GeneRIF: Study demonstrated that MAEL interacts with Snail and
            inhibit E-cadherin promoter activity. MAEL is an oncogene that
            plays an important role in the development and progression of colon
            cancer.
REFERENCE   2  (residues 1 to 434)
  AUTHORS   Li Q, Wojciechowski R, Simpson CL, Hysi PG, Verhoeven VJ, Ikram MK,
            Hohn R, Vitart V, Hewitt AW, Oexle K, Makela KM, MacGregor S,
            Pirastu M, Fan Q, Cheng CY, St Pourcain B, McMahon G, Kemp JP,
            Northstone K, Rahi JS, Cumberland PM, Martin NG, Sanfilippo PG, Lu
            Y, Wang YX, Hayward C, Polasek O, Campbell H, Bencic G, Wright AF,
            Wedenoja J, Zeller T, Schillert A, Mirshahi A, Lackner K, Yip SP,
```

FASTA

Fasta format

## protein maelstrom homolog isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_116247.1

GenPept Identical Proteins Graphics

```
>NP_116247.1 protein maelstrom homolog isoform 1 [Homo sapiens]
MPNRKASRNAYYFFVQEKIPELRRRGLPVARVADAIPYCSSDWALLREEEKEKYAEMAREWRAAQGKDPG
PSEKQKPVFTPLRRPGMLVPKQNVSPPDMSALSLKGDQALLGGIFYFLNIFSHGELPPHCEQRFLPCEIG
CVKYSLQEGIMADFHSFINPGEIPRGFRFHCQAASDSSHKIPISNFERGHNQATVLQNLYRFIHPNPGNW
PPIYCKSDDRTRVNWCLKHMAKASEIRQDLQLLTVEDLVVGIYQQKFLKEPSKTWIRSLLDVAMWDYSSN
TRCKWHEENDILFCALAVCKKIAYCISNSLATLFGIQLTEAHVPLQDYEASNSVTPKMVVLDAGRYQKLR
VGSSGFSHFNSSNEEQRSNTPIGDYPSRAKISGQNSSVRGRGITRLLESISNSSSNIHKFSNCDTSLSPY
MSQKDGYKSFSSLS
```

# NCBI-Genome (www.ncbi.nlm.nih.gov/genome/)

NCBI Resources How To dapengzhangca@gmail.com My NCBI Sign Out

Genome | Genome | human[orgn] | Search

Create alert Limits Advanced Help

**Homo sapiens (human)**
**Reference genome: Homo sapiens (assembly GRCh38.p11)**
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Homo sapiens **genome**
**All 61 genomes for species:**
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**

Display Settings: Overview Send to:

**Organism Overview** ; Genome Assembly and Annotation report [61] ; Organelle Annotation Report [19] ID: 51

## Homo sapiens (human)

Human genome projects have generated an unprecedented amount of knowledge about human genetics and health.

**Lineage: Eukaryota[2448]; Metazoa[818]; Chordata[348]; Craniata[340]; Vertebrata[340]; Euteleostomi[335]; Mammalia[143]; Eutheria[138]; Euarchontoglires[63]; Primates[28]; Haplorrhini[22]; Catarrhini[17]; Hominidae[5]; Homo[1]; Homo sapiens[1]**

Study of the human condition such as genetic and infectious disease, the intersection between genetics and the environment, and population variation is supported by a wealth of genome-scale data. These data sets include: a) numerous sequenced genomes including several which have been assembled; b) studies that examine transcript and protein existence, More...

### Summary

**Sequence data:** genome assemblies: 61; sequence reads: 387 (See Genome Assembly and Annotation report)
**Statistics:** median total length (Mb): 2996.43
median protein count: 79257
median GC%: 40.9
**NCBI Annotation Release:** 108

### Publications

1. Long-read sequencing and de novo assembly of a Chinese genome. Shi L, et al. Nat Commun 2016 Jun 30
2. De novo assembly and phasing of a Korean human genome. Seo JS, et al. Nature 2016 Oct 13
3. Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. Mohajeri K, et al. Genome Res 2016 Nov

More...

**NCBI Resources**
Genome Data Viewer
FTP Human annotation (GFF)
FTP Human chromosomes
Map Viewer

**Tools**
BLAST Genome

**Related information**
Assembly
BioProject
Gene
Components
Protein
PubMed
Taxonomy

**Search details**
"Homo sapiens"[Organism]

Search See more...

# NCBI-Genome (www.ncbi.nlm.nih.gov/genome/)

**Representative** (genome information for reference and representative genomes)

**Reference genome:** [see all organisms]

- *Homo sapiens GRCh38.p11*

**Submitter:** Genome Reference Consortium

| Loc | Type | Name | RefSeq | INSDC | Size (Mb) | GC% | Protein | rRNA | tRNA | Other RNA | Gene | Pseudogene |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nuc | Chr | 22 | NC_000022.11 | CM000684.2 | 50.82 | 47.7 | 2,493 | - | - | 965 | 1,172 | 348 |
| Nuc | Chr | 1 | NC_000001.11 | CM000663.2 | 248.96 | 42.3 | 11,046 | 17 | 90 | 4,350 | 5,078 | 1,372 |
| Nuc | Chr | 2 | NC_000002.12 | CM000664.2 | 242.19 | 40.3 | 8,054 | - | 8 | 3,638 | 3,862 | 1,166 |
| Nuc | Chr | 3 | NC_000003.12 | CM000665.2 | 198.3 | 39.7 | 6,790 | - | 4 | 2,723 | 2,971 | 887 |
| Nuc | Chr | 4 | NC_000004.12 | CM000666.2 | 190.22 | 38.3 | 4,374 | - | 1 | 2,209 | 2,441 | 799 |
| Nuc | Chr | 5 | NC_000005.10 | CM000667.2 | 181.54 | 39.5 | 4,599 | - | 17 | 2,225 | 2,578 | 766 |

**Homo sapiens (human)**

**Reference genome: Homo sapiens (assembly GRCh38.p11)**

Download sequences in FASTA format for **genome**, **transcript**, **protein**

Download genome annotation in **GFF**, **GenBank** or **tabular** format

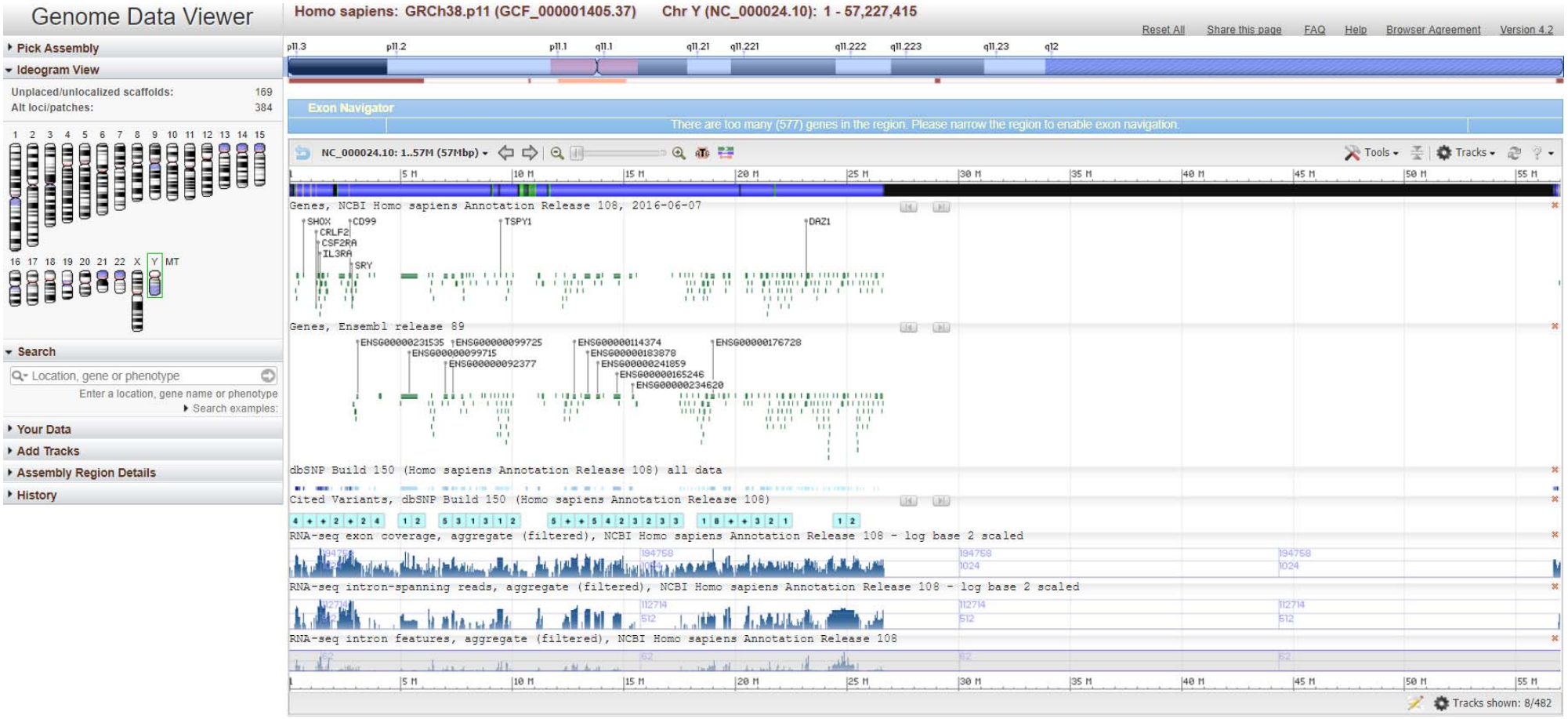
BLAST against Homo sapiens **genome**

**All 61 genomes for species:**

Browse the **list**

Download sequence and annotation from **RefSeq** or **GenBank**

# Genome Data Viewer



# NCBI-SRA (sequence read archive)

NCBI Resources How To dapengzhangca@gmail.com My NCBI Sign Out

SRA SRA Search

Advanced Help

## SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

### Getting Started

- How to Submit
- Log in to SRA (for updating and troubleshooting submissions)
- Log in to Submission Portal (for submitting sequence data)
- SRA Documentation
- Download Guide
- SRA Fact Sheet (.pdf)

### Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

### Related Resources

- Submission Portal
- Trace Archive
- dbGaP Home
- BioProject
- BioSample

To look for all human gut microbiome data:

(gut microbiome[All Fields]) AND "Homo sapiens"[Organism]

SRA | SRA | (gut microbiome[All Fields]) AND "Homo sapiens"[Organism] | Search

Create alert Advanced Help

**Access**
Controlled (470)
Public (1,403)

**Source** RNA/transcriptome
DNA (1,779)
RNA (36)

**Type**
genome (23)

Clear all
metagenome
Show additional filters

Summary 20 per page Send to:

View results as an expanded interactive table using the RunSelector. Send results to Run selector

**Search results**

**Items: 1 to 20 of 1873**

<< First < Prev Page 1 of 94 Next > Last >>

1. 454 GS FLX Titanium sequencing
   1 LS454 (454 GS FLX Titanium) run: 3,136 spots, 1.3M bases, 2.8Mb downloads
   Accession: ERX458806
2. 454 GS FLX Titanium sequencing
   1 LS454 (454 GS FLX Titanium) run: 8,572 spots, 3.4M bases, 7.7Mb downloads
   Accession: ERX458801
3. 454 GS FLX Titanium sequencing
   1 LS454 (454 GS FLX Titanium) run: 4,605 spots, 1.8M bases, 4.1Mb downloads
   Accession: ERX458800
4. 454 GS FLX Titanium sequencing
   1 LS454 (454 GS FLX Titanium) run: 15,212 spots, 6.1M bases, 13.6Mb downloads
   Accession: ERX458799

**Filters:** Manage Filters

**Search in related databases**

| Database | Access | | all |
|---|---|---|---|
| | public | controlled | |
| BioSample | 652 | 312 | 964 |
| BioProject | 22 | 2 | 24 |
| dbGaP | | 1 | 1 |
| GEO Datasets | | | |

**Find related data**

Database: Select

Find items

**Search details**

```
("gut metagenome"[Organism] OR gut
microbiome[All Fields]) AND "Homo
sapiens"[Organism]
```

# NCBI-Taxonomy

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC

Search for [ ] as complete name ☑ lock Go

Display 3 levels using filter: none

The "Token set" option returns longer names that include the search terms, e.g., hybrid taxa. See what happens if you query "Bos taurus" using the "Complete match" option versus the "Set of tokens when you are not sure about the exact spelling of a organism name. It tries to find the phonetically closest strings (try "Drozofila" as an example).

**This is the top level of the taxonomy database maintained by NCBI/GenBank. You can explore any of the taxa listed below by clicking it.**

- Archaea
- Bacteria
- Eukaryota
- Viroids
- Viruses
- Other
- Unclassified

---

**These are direct links to some of the organisms commonly used in molecular research projects:**

| | | |
|---|---|---|
| Arabidopsis thaliana | Escherichia coli | Pneumocystis carinii |
| Bos taurus | Hepatitis C virus | Rattus norvegicus |
| Caenorhabditis elegans | Homo sapiens | Saccharomyces cerevisiae |
| Chlamydomonas reinhardtii | Mus musculus | Schizosaccharomyces pombe |
| Danio rerio (zebrafish) | Mycoplasma pneumoniae | Takifugu rubripes |
| Dictyostelium discoideum | Oryza sativa | Xenopus laevis |
| Drosophila melanogaster | Plasmodium falciparum | Zea mays |

# Homo sapiens

Search for | as complete name | lock | Go

Display 3 levels using filter: none

## Homo sapiens

*Taxonomy ID:* 9606
*Genbank common name:* **human**
*Inherited blast name:* **primates**
*Rank:* species
*Genetic code:* Translation table 1 (Standard)
*Mitochondrial genetic code:* Translation table 2 (Vertebrate Mitochondrial)
*Other names:*
common name: **man**
authority: **Homo sapiens Linnaeus, 1758**

*Lineage( full )*
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

| Entrez records | | |
|---|---|---|
| Database name | Subtree links | Direct links |
| Nucleotide | 14,908,727 | 14,908,681 |
| Nucleotide EST | 8,705,106 | 8,705,106 |
| Nucleotide GSS | 1,783,249 | 1,781,923 |
| Protein | 1,175,370 | 1,175,052 |
| Structure | 35,858 | 35,858 |
| Genome | 1 | 1 |
| Popset | 23,768 | 23,767 |
| SNP | 336,828,091 | 336,828,091 |
| Domains | 43 | 43 |
| GEO Datasets | 1,256,551 | 1,256,551 |
| UniGene | 130,056 | 130,056 |
| PubMed Central | 24,983 | 24,918 |
| Gene | 221,647 | 221,574 |
| HomoloGene | 18,713 | 18,713 |
| SRA Experiments | 833,162 | 832,633 |
| Probe | 27,382,489 | 27,382,489 |
| Assembly | 108 | 108 |
| Bio Project | 36,789 | 36,776 |
| Bio Sample | 3,071,592 | 3,071,440 |
| Bio Systems | 3,077 | 3,077 |
| Clone DB | 17,630,270 | 17,630,270 |
| dbVar | 5,169,896 | 5,157,957 |
| GEO Profiles | 61,958,910 | 61,958,910 |
| PubChem BioAssay | 311,518 | 311,510 |
| Protein Clusters | 13 | 13 |
| Taxonomy | 3 | 1 |

# Capsaspora

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search for [ ] as complete name ▾ lock Go

Display 2 levels using filter: has genome sequences ▾

## Capsaspora owczarzaki ATCC 30864

*Taxonomy ID:* 595528
*Inherited blast name:* **eukaryotes**
*Rank:* no rank
*Genetic code:* Translation table 1 (Standard)
*Mitochondrial genetic code:* Translation table 4 (Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate Mitochondrial; Mycoplasma; Spiroplasma)

*Lineage( full )*
cellular organisms; Eukaryota; Opisthokonta; Opisthokonta incertae sedis; Ichthyosporea; Capsaspora; Capsaspora owczarzaki

| Entrez records | |
|---|---|
| Database name | Direct links |
| Nucleotide | 8,993 |
| Protein | 18,915 |
| Genome | 1 |
| PubMed Central | 3 |
| Gene | 9,504 |
| SRA Experiments | 11 |
| Assembly | 2 |
| Bio Project | 2 |
| Bio Sample | 12 |
| Taxonomy | 1 |

**External Information Resources (NCBI LinkOut)**

| LinkOut | Subject | LinkOut Provider |
|---|---|---|
| GOLD: Go0003455 | organism-specific | Genomes On Line Database |
| WebScipio: Capsaspora owczarzaki ATCC 30864 | organism-specific | WebScipio - eukaryotic gene identification |
| diArk: Capsaspora owczarzaki ATCC 30864 | organism-specific | diArk a resource for eukaryotic genome research |

# Capsaspora

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein

Search for [ ] as complete name ▼ lock Go

Display 2 levels using filter: has genome sequences ▼

Nucleotide | Nucleotide EST | Nucleotide GSS | Protein | Structure | Genome | Popset
Domains | GEO Datasets | UniGene | PubMed Central | Gene | HomoloGene | SRA Expe
Assembly | LinkOut | BLAST | TRACE | Host | Viral Host | Bio Project
Bio Systems | Clone DB | dbVar | GEO Profiles | PubChem BioAssay | Protein Clusters

**Lineage** (full): root; cellular organisms; Eukaryota

- **Opisthokonta** *Click on organism name to get more information.*
  - **Choanoflagellida**
    - **Craspedida**
  - **Fungi** (fungi)
    - **Blastocladiomycota**
    - **Chytridiomycota**
    - **Cryptomycota**
    - **Dikarya**
    - **Microsporidia**
    - **Mucoromycota**
    - **Neocallimastigomycota**
    - **Zoopagomycota**
    - **unclassified Fungi**
  - **Metazoa** (metazoans)
    - **Eumetazoa**
    - **Mesozoa**
    - **Placozoa** (placozoans)
    - **Porifera** (sponges)
  - **Nucleariidae and Fonticula group**
    - **Fonticula**
    - **Fonticula-like sp. SCN 57-25**
  - **Opisthokonta incertae sedis**
    - **Ichthyosporea**

# Entrez: Integrated Molecular and Sequence Databases

**Search NCBI databases**

all[sb] Search

About 1,353,092,513 search results for "all[sb]"

Central Resources / Databases

- **Taxonomy**
- **BioProject**
- **Assembly**
- **Gene**

Follow links to others when needed

**Nucleotide, Protein, SRA**

## Literature

| | | |
|---|---|---|
| Books | 334,634 | |
| MeSH | 253,991 | |
| NLM Catalog | 1,511,375 | |
| PubMed | 24,324,879 | |
| PubMed Central | 3,264,562 | |

## Health

| | | |
|---|---|---|
| ClinVar | 125,404 | |
| dbGaP | 166,753 | |
| GTR | 35,977 | |
| MedGen | 260,910 | |
| OMIM | 23,795 | online mendelian inheritance in man |
| PubMed Health | 51,412 | clinical effectiveness, disease and drug reports |

## Genomes

| | | |
|---|---|---|
| Assembly | 33,680 | genomic assembly information |
| BioProject | 138,999 | biological projects providing data to NCBI |
| BioSample | 2,865,750 | descriptions of biological source materials |
| Clone | 37,024,042 | genomic and cDNA clones |
| dbVar | 4,305,990 | genome structural variation studies |
| Epigenomics | 6,635 | epigenomic studies and display tools |
| Genome | 10,777 | genome sequencing projects by organism |
| GSS | 37,646,655 | genome survey sequences |
| Nucleotide | 155,458,773 | DNA and RNA sequences |
| Probe | 31,890,151 | sequence-based probes and primers |
| SNP | 444,458,769 | short genetic variations |
| SRA | 1,064,488 | high-throughput DNA and RNA sequence read archive |
| Taxonomy | 1,310,804 | taxonomic classification and nomenclature catalog |

| | | |
|---|---|---|
| | | expressed sequence tag sequences |
| | | collected information about gene loci |
| | | functional genomics studies |
| | | gene expression and molecular abundance profiles |
| | | homologous gene sets for selected organisms |
| | | sequence sets from phylogenetic and population studies |
| | | clusters of expressed transcripts |
| Domains | | conserved protein domains |
| Protein | 153,454,195 | protein sequences |
| Protein Clusters | 820,546 | sequence similarity-based protein clusters |
| Structure | 103,644 | experimentally-determined biomolecular structures |

## Chemicals

| | | |
|---|---|---|
| BioSystems | 653,443 | molecular pathways with links to genes, proteins and chemicals |
| PubChem BioAssay | 1,112,105 | bioactivity screening studies |
| PubChem Compound | 62,041,347 | chemical information with structures, information and links |
| PubChem Substance | 177,330,151 | deposited substance and chemical information |

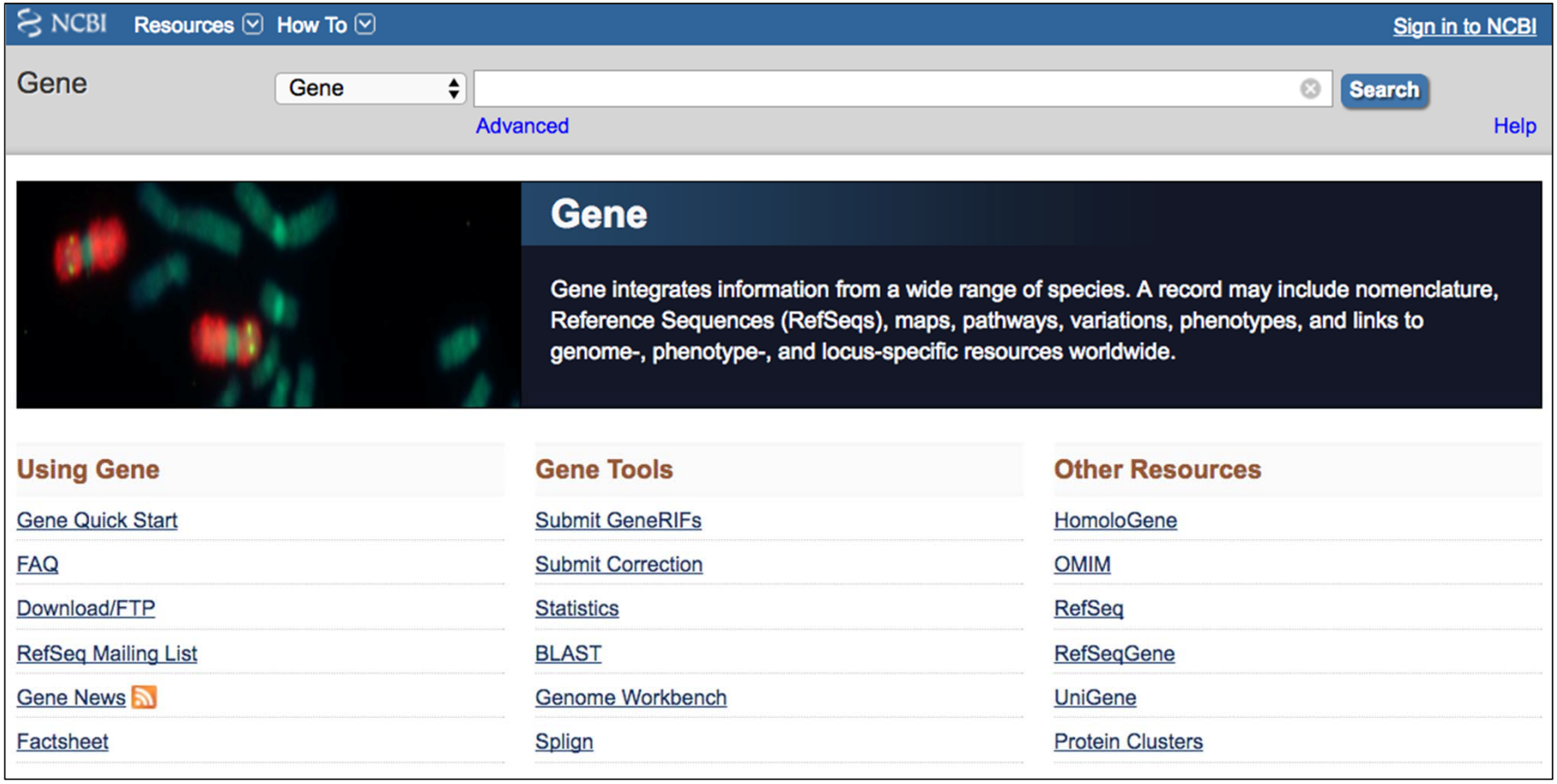
The Entrez system: 39 (and counting) integrated databases

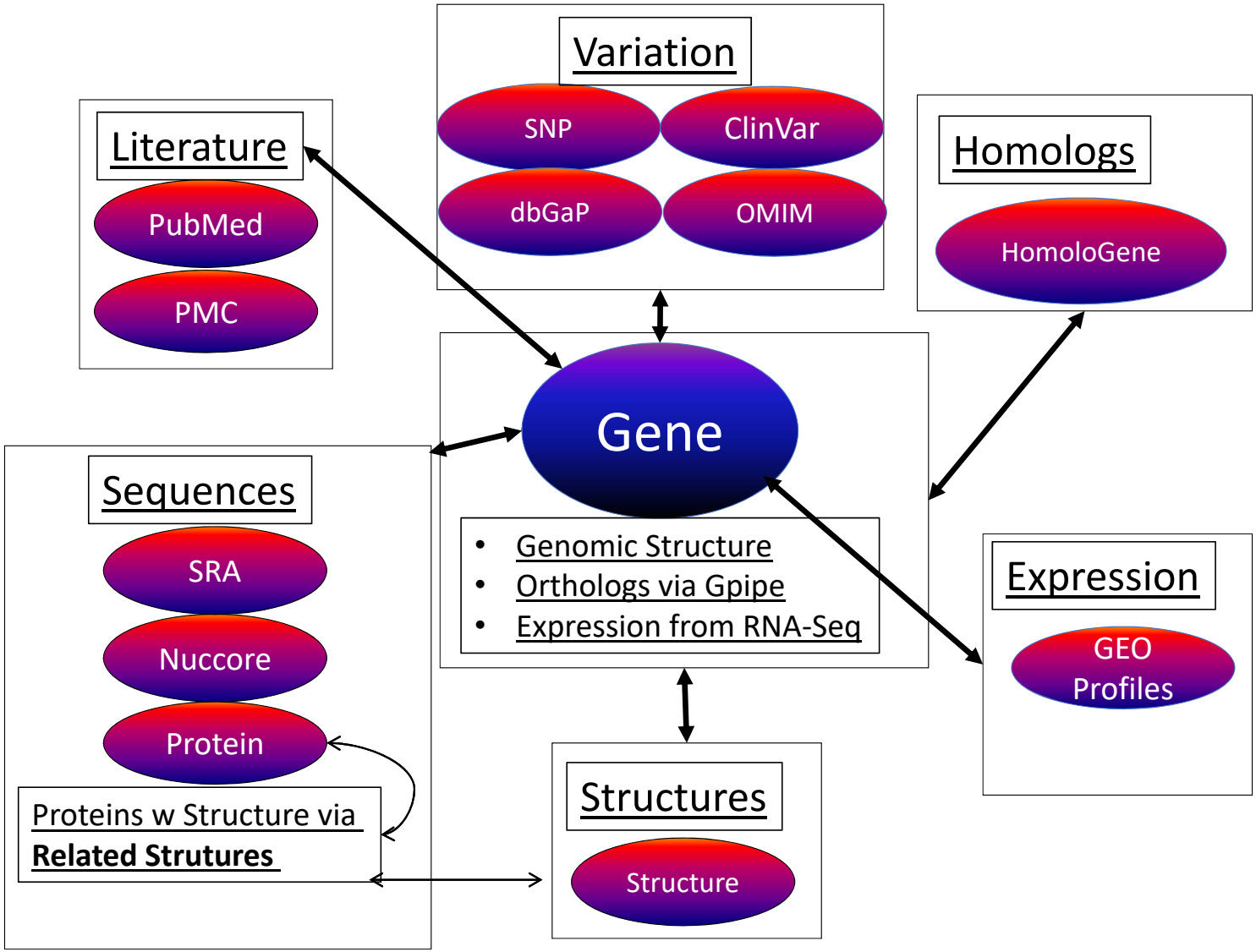
# Start in High Level Resources

If your question is about data for ...

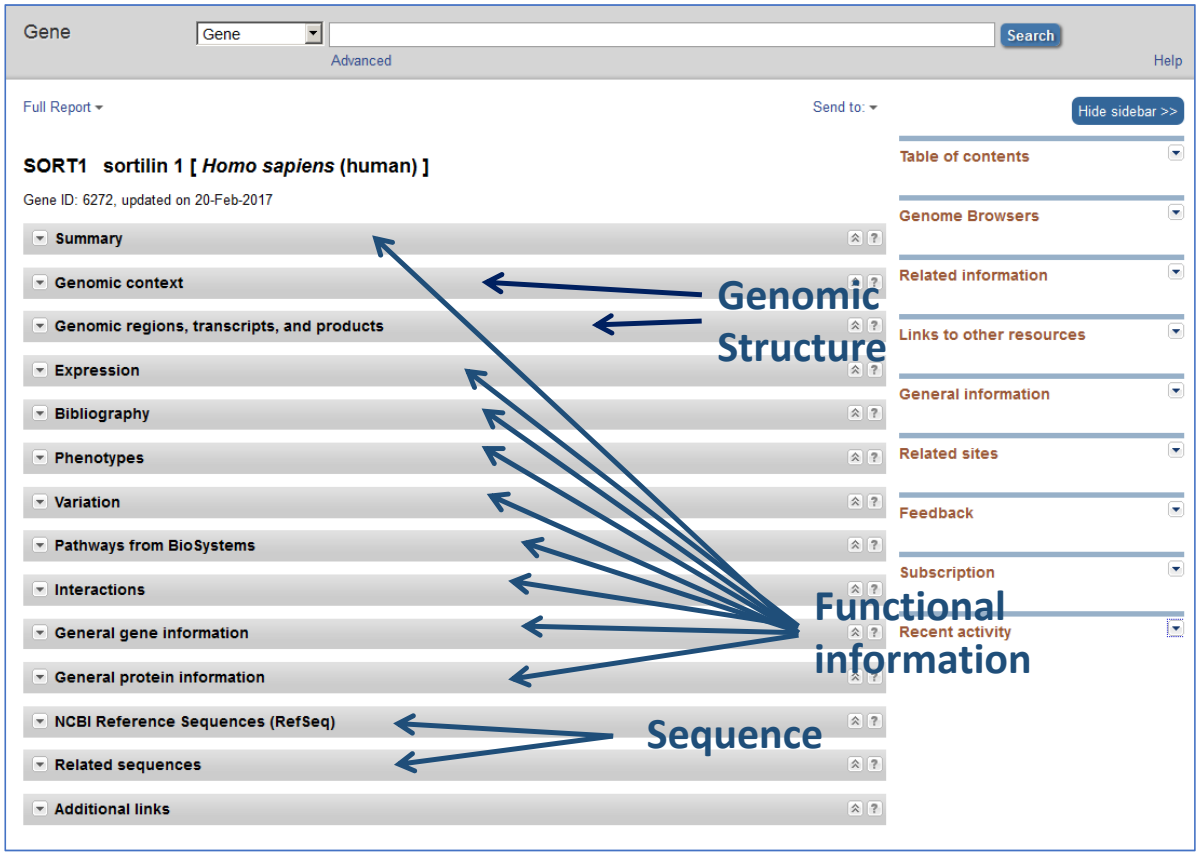
- an organism -> Taxonomy
- a gene name -> Gene (common organisms)
- a large-scale project -> BioProject
- a bacterial genome -> Genome
- a genome sequence -> Assembly

# NCBI-Gene (www.ncbi.nlm.nih.gov/gene)




Variation
SNP
ClinVar
dbGaP
OMIM
Literature
PubMed
PMC
Homologs
HomoloGene
Gene
Genomic Structure
Orthologs via Gpipe
Expression from RNA-Seq
Sequences
SRA
Nuccore
Protein
Proteins w Structure via
Related Strutures
Expression
GEO
Profiles
Structures
Structure

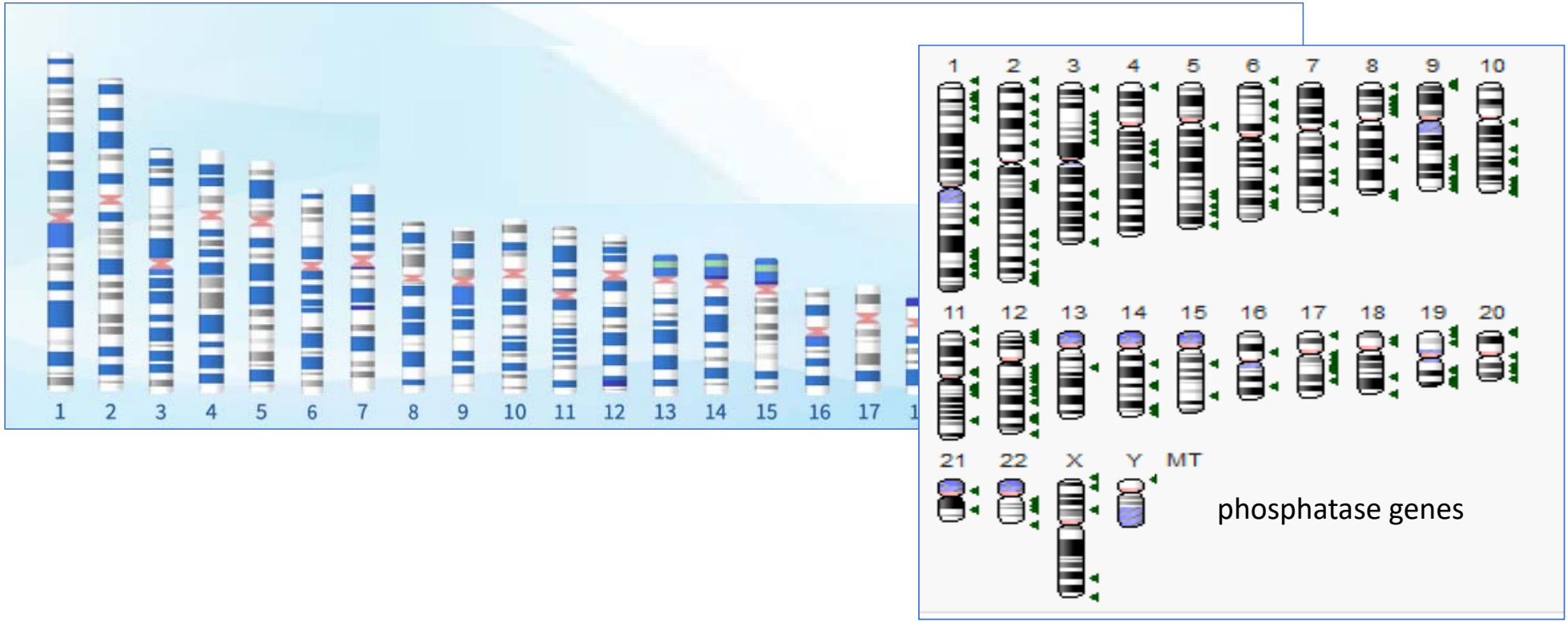
# A gene record: SORT1 (https://www.ncbi.nlm.nih.gov/gene/6272)



# Some Examples

1. Where is the gene located (chromosome and position) in the genome?
2. What are the Reference genomic, transcript and protein sequences for the gene?
3. What variations are present in the gene and are they associated with disease?
4. What are the equivalent genes (homologs) in other species?

# 1. Where is the gene located in the genome?



# Genome context and structure:

DSG2[All Fields] AND "Homo sapiens"[Organism]

**DSG2 desmoglein 2 [ *Homo sapiens* (human) ]**

Gene ID: 1829, updated on 20-Feb-2017

**Summary**

**Official Symbol** DSG2 provided by HGNC

Table of contents

Summary

Genomic context

Genomic regions, transcripts, and products

Expression

**Genomic context**

**Location:** 18q12.1

**Exon count:** 15

| Annotation release | Status | Assembly |
|---|---|---|
| 108 | current | GRCh38.p7 (GCF_ |
| 105 | previous assembly | GRCh37.p13 (GCF_ |

Chromosome 1

[ 31343368

DSG1-AS1 DSG2 DSG4 DSG3 DSG2-

**Genomic regions, transcripts, and products**

Go to reference sequence details

**Genomic Sequence:** NC_000018.10 Chromosome 18 Reference GRCh38.p7 Primary Assembly

**Go to nucleotide:** Graphics FASTA GenBank

NC_000018.10: 31M..32M (66Kbp) Tools Tracks

31,490 K 31,500 K 31,510 K 31,520 K 31,530 K 31,540 K 31,550 K

Genes, NCBI Homo sapiens Annotation Release 108, 20...

DSG2

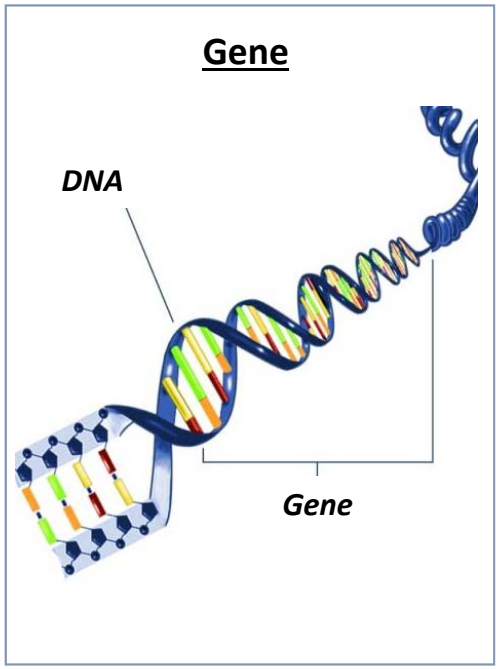
NM_001943.4 NP_001934.2

DSG2-AS1

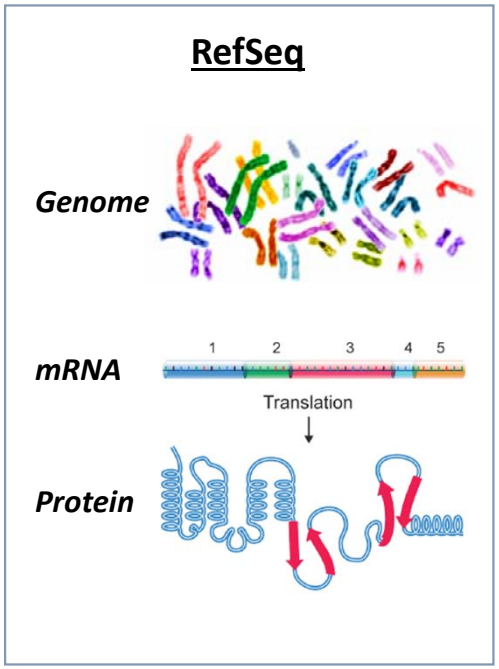
A lab report has shown that a patient has a deletion of chromosome 5p which corresponds to nts 204,700-5,500,000. How do I find out what genes are involved?

A recent research article has shown that a genomic region is associated with a disease. How can I find what genes are there?

# 2. What are the Reference genomic, transcript and protein sequences for the gene?

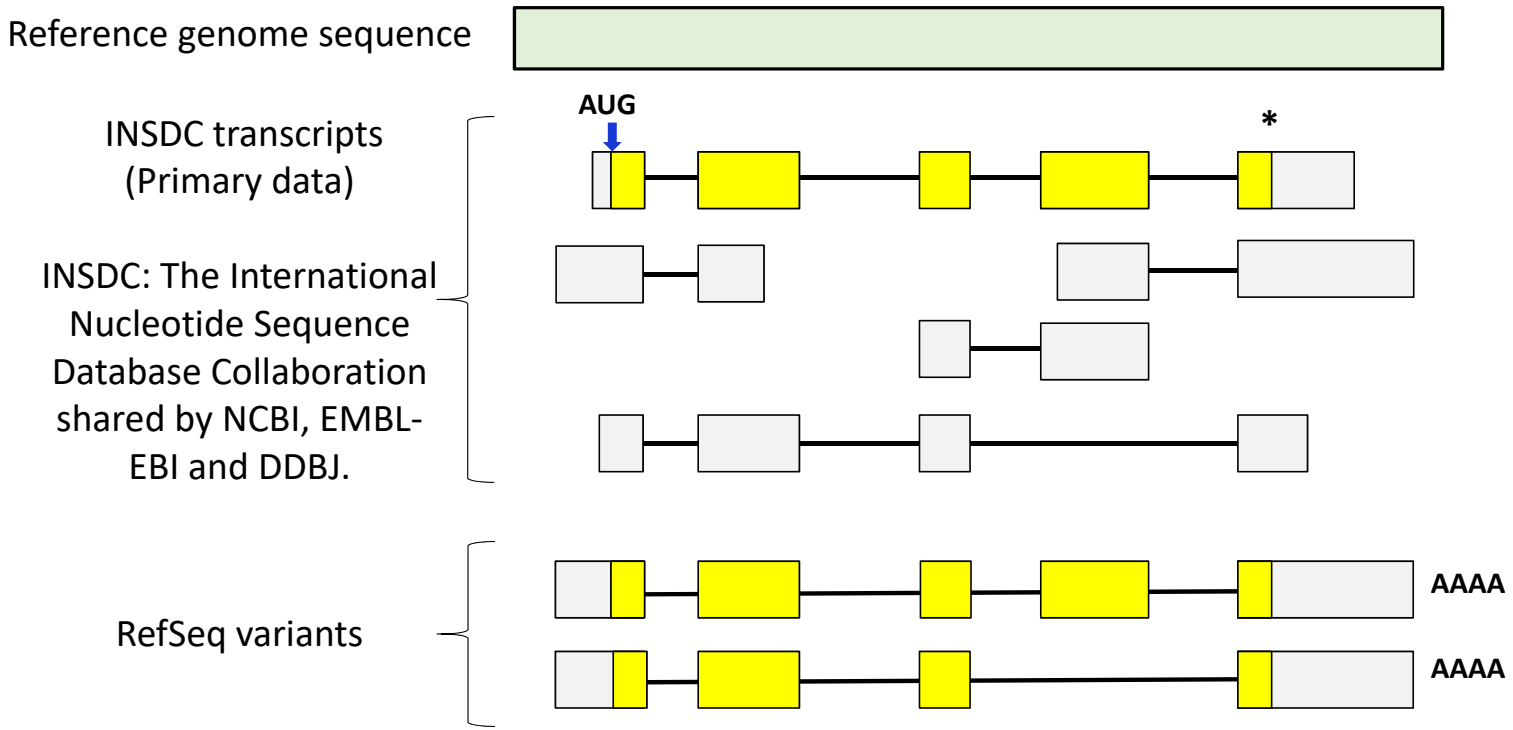



*Gene level*




*Transcript and protein level*

# NCBI Reference Sequences (RefSeq) Project



# RefSeq

- Created by NCBI data curators and computational algorithms
- Used by researchers & collaborators as a reference standard
- For selected eukaryotes, represent all molecules in the central dogma
  - Genomic (DNA), Transcripts (mRNA), Proteins
- Distinct accession series – with an underscore!
  - Genomic: NC_, AC_, NG_, NT_, NW_
  - Transcripts: NM_, NR_, XM_, XR_
  - Proteins: NP_, XP_

# RefSeq

**DSG2 desmoglein 2 [ *Homo sapiens* (human) ]**

Gene ID: 1829, updated on 5-Mar-2017

**Summary**

| | |
|---|---|
| **Official Symbol** | DSG2 provided by HGNC |
| **Official Full Name** | desmoglein 2 provided by HGNC |
| **Primary source** | HGNC:HGNC:3049 |
| **See related** | Ensembl:ENSG00000046604 MIM:125671; Vega:OTTHUMG0000 |

a; Euteleostom
Hominidae; Homo

mily and cadherin cell adhesion molecule superfamily of
embrane glycoprotein components of desmosomes,
nd other cell types. The encoded preproprotein is
ycoprotein. This gene is present in a gene cluster with
some 18. Mutations in this gene have been associated
hilial, 10. [provided by RefSeq, Jan 2016]

Genomic context

Ontology

General protein information

NCBI Reference Sequences (RefSeq)

Related sequences

Additional links

Locus-specific Databases

**RefSeqs of Annotated Genomes: Homo sapiens Annotation Release 108** details...

The following sections contain reference sequences that belong to a specific genome build. Explain

**Reference GRCh38.p7 Primary Assembly**

***Genomic***

1. **NC_000018.10 Reference GRCh38.p7 Primary Assembly**

| | |
|---|---|
| Range | 31498004..31549008 |
| Download | GenBank, FASTA, Sequence Viewer (Graphics) |

**NCBI Reference Sequences (RefSeq)**

**RefSeqs maintained independently of Annotated Genomes**

These reference sequences exist independently of genome builds. Explain

***Genomic***

1. **NG_007072.3 RefSeqGene**

| | |
|---|---|
| Range | 4823..55610 |
| Download | GenBank, FASTA, Sequence Viewer (Graphics), LRG_397 |

***mRNA and Protein(s)***

1. NM_001943.4 → NP_001934.2 **desmoglein-2 preproprotein**

See identical proteins and their annotated locations for NP_001934.2

I am studying the globin gene cluster on chromosome 11 and I need to get the protein sequences for all gene family members.

I am interested in the promoter region of genes that don’t code for proteins. How can I get this information?

# 3. What variations are present in the gene and are they associated with disease?

**C9orf72 chromosome 9 open reading frame 72 [ *Homo sapiens* (human) ]**

Gene ID: 203228, updated on 3-Sep-2017

**Summary**

| | |
|---|---|
| **Official Symbol** | C9orf72 provided by HGNC |
| **Official Full Name** | chromosome 9 open reading frame 72 provided by HGNC |
| **Primary source** | HGNC:HGNC:28337 |
| **See related** | Ensembl:ENSG00000147894 MIM:614260; Vega:OTTHUMG00000019716 |
| **Gene type** | protein coding |
| **RefSeq status** | REVIEWED |
| **Organism** | Homo sapiens |
| **Lineage** | Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo |
| **Also known as** | ALSFTD; FTDALS; DENNL72; FTDALS1 |
| **Summary** | The protein encoded by this gene plays an important role in the regulation of endosomal trafficking, and has been shown to interact with Rab proteins that are involved in autophagy and endocytic transport. Expansion of a GGGGCC repeat from 2-22 copies to 700-1600 copies in the intronic sequence between alternate 5' exons in transcripts from this gene is associated with 9p-linked ALS (amyotrophic lateral sclerosis) and FTD (frontotemporal dementia) (PMID: 21944778, 21944779). Studies suggest that hexanucleotide expansions could result in the selective stabilization of repeat-containing pre-mRNA, and the accumulation of insoluble dipeptide repeat protein aggregates that could be pathogenic in FTD-ALS patients (PMID: 23393093). Alternative splicing results in multiple transcript variants encoding different isoforms. [provided by RefSeq, Jul 2016] |
| **Orthologs** | mouse all |

**Genomic context**

**Table of contents**


Summary
Genomic context
Genomic regions, transcripts, and products
Expression
Bibliography
Phenotypes
Variation
Pathways from BioSystems
Interactions
General gene information
Markers, Clone Names, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links
Locus-specific Databases

# dbSNP, dbVar, ClinVar and their overlap

- dbSNP: Database of Short Genetic Variations (Single-nucleotide polymorphism); 53 organisms

- dbVar: Database of Genomic Structural Variations (insertions and deletions, tri-/ hexa- nucleotide repeat expansion)

- ClinVar: Database of relationships between human variations and phenotypes (observed health status)

# Variation and Disease

## Phenotypes

Find tests for this gene in the NIH Genetic Testing Registry (GTR)

Review eQTL and phenotype association data in this region using PheGenI

Associated conditions

| Description |
| --- |
| Amyotrophic lateral sclerosis<br>MedGen: C0002736, GeneReviews: Amyotrophic Lateral Sclerosis Overview |
| Amyotrophic lateral sclerosis and/or frontotemporal dementia 1<br>MedGen: C1862937, OMIM: 105550, GeneReviews: Amyotrophic Lateral Sclerosis Overview, C9orf72-Related Amyotrophic Lateral Sclerosis and Frontotemporal Dementia |

## Variation

See variants in ClinVar

See studies and variants in dbVar

See Variation Viewer (GRCh37.p13)

See Variation Viewer (GRCh38)

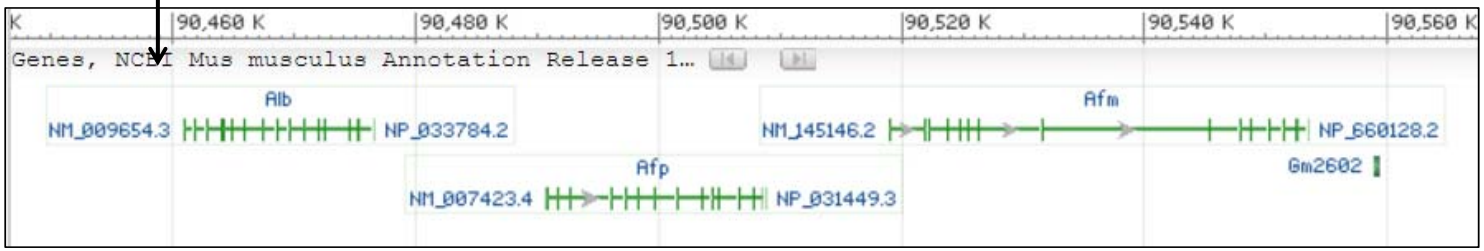
What diseases can be caused by variations in the tyrosine hydroxylase gene? Can I get a list of all disease causing single nucleotide variants that affect the coding regions with their positions? Are there any common protein variants in this gene?

4. What are the equivalent genes (homologs) in other species?

# Orthologs and Paralogs

Many gene families have multiple members; Serum albumin gene family: albumin (Alb), alpha-fetoprotein (AFP) and afamin (Afm).


- Human ALB, AFP and AFM are **paralogous genes**.
- **Paralogs are derived by gene duplication within a species (or ancestral species).**




- ALB (human) and Alb (mouse) are **orthologous** genes.
- **Orthologs are derived by speciation events (homologs between species).**

# 4. What are the equivalent genes (homologs) in other species?

**C9orf72 chromosome 9 open reading frame 72 [ *Homo sapiens* (human) ]**

Gene ID: 203228, updated on 3-Sep-2017

**Summary**

| | |
|---|---|
| **Official Symbol** | C9orf72 provided by HGNC |
| **Official Full Name** | chromosome 9 open reading frame 72 provided by HGNC |
| **Primary source** | HGNC:HGNC:28337 |
| **See related** | Ensembl:ENSG00000147894 MIM:614260; Vega:OTTHUMG00000019716 |
| **Gene type** | protein coding |
| **RefSeq status** | REVIEWED |
| **Organism** | Homo sapiens |
| **Lineage** | Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo |
| **Also known as** | ALSFTD; FTDALS; DENNL72; FTDALS1 |
| **Summary** | The protein encoded by this gene plays an important role in the regulation of endosomal trafficking, and has been shown to interact with Rab proteins that are involved in autophagy and endocytic transport. Expansion of a GGGGCC repeat from 2-22 copies to 700-1600 copies in the intronic sequence between alternate 5' exons in transcripts from this gene is associated with 9p-linked ALS (amyotrophic lateral sclerosis) and FTD (frontotemporal dementia) (PMID: 21944778, 21944779). Studies suggest that hexanucleotide expansions could result in the selective stabilization of repeat-containing pre-mRNA, and the accumulation of insoluble dipeptide repeat protein aggregates that could be pathogenic in FTD-ALS patients (PMID: 23393093). Alternative splicing results in multiple transcript variants encoding different isoforms. [provided by RefSeq, Jul 2016] |
| **Orthologs** | mouse all |

Table of contents


Summary
Genomic context
Genomic regions, transcripts, and products
Expression
Bibliography
Phenotypes
Variation
Pathways from BioSystems
Interactions
General gene information
Markers, Clone Names, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links


Homology

Homologs of the C9orf72 gene: The C9orf72 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, and zebrafish.
Orthologs from Annotation Pipeline: 216 organisms have orthologs with human gene C9orf72
The Hierarchical Catalog of Orthologs

# 4. What are the equivalent genes (homologs) in other species?

**HomoloGene:10137. Gene conserved in Euteleostomi**

## Genes

*Genes identified as putative homologs of one another during the construction of HomoloGene.*

C9orf72, *H.sapiens*
chromosome 9 open reading frame 72
C9H9orf72, *P.troglodytes*
chromosome 9 open reading frame, human C9orf72
C15H9orf72, *M.mulatta*
chromosome 9 open reading frame 72 ortholog
C11H9orf72, *C.lupus*
chromosome 11 open reading frame, human C9orf72
C8H9orf72, *B.taurus*
chromosome 8 open reading frame, human C9orf72
3110043O21Rik, *M.musculus*
RIKEN cDNA 3110043O21 gene
RGD1359108, *R.norvegicus*
similar to RIKEN cDNA 3110043O21
C9ORF72, *G.gallus*
chromosome Z open reading frame, human C9orf72
zgc:100846, *D.rerio*
zgc:100846

## Proteins

*Proteins used in sequence comparisons and their conserved domain architectures.*

NP_001242983.1
481 aa
XP_003951454.1
481 aa
XP_002800118.1
481 aa
XP_005626755.1
481 aa
NP_001096558.1
481 aa
NP_001074812.1
481 aa
NP_001007703.1
481 aa
XP_424945.2
481 aa
NP_991166.1
462 aa

## Protein Alignments

*Protein multiple alignment, pairwise similarity scores and evolutionary distances.*

Show Multiple Alignment

## Conserved Domains

*Conserved Domains from CDD found in protein sequences by rpsblast searching.*

C9orf72-like (pfam15019)
C9orf72-like protein family.

1. A lab report has shown that a patient has a deletion of chromosome 5p which corresponds to nts 204,700-5,500,000. How do I find out what genes are involved?

2. I am studying the globin gene cluster on 11 and I need to get the protein sequences for all gene family members.

3. I am interested in the promoter region of genes that don’t code for proteins. How can I get this information?

4. What diseases can be caused by variations in the tyrosine hydroxylase gene? Can I get a list of all disease causing single nucleotide variants that affect the coding regions with their positions? Are there any common protein variants in this gene?

# Data formats

- GenBank DNA sequence entry
- EMBL sequence entry
- FASTA sequence format
- ASN.1: Abstract Syntax Notation
- XML
- PDB
- Phylip: Phylogenetic Inference package
- Clustal
- MSF: multiple sequence format

# One format can be converted to another

- READSEQ: http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi
  - http://iubio.bio.indiana.edu/soft/molbio/readseq/java/Readseq2-help.html

- EMBOSS Seqret: www.ebi.ac.uk/Tools/sfc/emboss_seqret/

- SEQIO: http://search.cpan.org/dist/BioPerl/Bio/SeqIO.pm