

# **EBI web resources II: Ensembl and InterPro**

Yanbin Yin

Fall 2015

<http://www.ebi.ac.uk/training/online/course/>

# Homework 3

- Go to <http://www.ebi.ac.uk/interpro/training.html> and finish the second online training course “Introduction to protein classification at the EBI” and then answer the following questions:
  - What is the difference between a protein family and a protein domain?
  - Can a protein belong to multiple families or contain multiple domains?
  - What are protein sequence features? Examples?
  - What is a protein signature? What is it used for?
  - What are the major signature types?
  - Is PROSITE a sequence pattern database or a profile database? What about Pfam?
  - What is the definition of “annotation”?
- In your report, answer these questions and also include the screen shot of the page(s) that support your answer.

**Due on 10/8** (send by email)

Office hour:

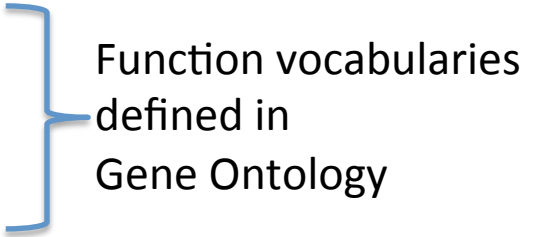
**Tue, Thu and Fri 2-4pm, MO325A**

**Or email: [yyin@niu.edu](mailto:yyin@niu.edu)**

# Outline

- Intro to genome annotation
- Protein family/domain databases
  - InterPro, Pfam, Superfamily etc.
- Genome browser
  - Ensembl
- Hands on Practice

# Genome annotation

- Predict genes (where are the genes?)
    - protein coding
    - RNA coding
  
  - Function annotation (What are these genes?)
    - Search against UniProt or NCBI-nr (GenPept)
    - Search against protein family/domain databases
    - Search against Pathway databases
- 
- Function vocabularies defined in Gene Ontology

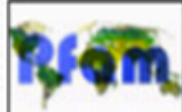
Proteins can be classified into groups according to sequence or structural similarity. These groups often contain well characterized proteins whose function is known. Thus, when a novel protein is identified, its functional properties can be proposed based on the group to which it is predicted to belong.

## Hidden Markov Models



PIRSF

TIGR  
tigr fams



SMART

## Finger-Prints



## Profiles



## Patterns



Structural domains

Functional annotation of families/domains

Protein features (sites)

Superfamily  
Gene3D

SCOP  
CATH

PDB

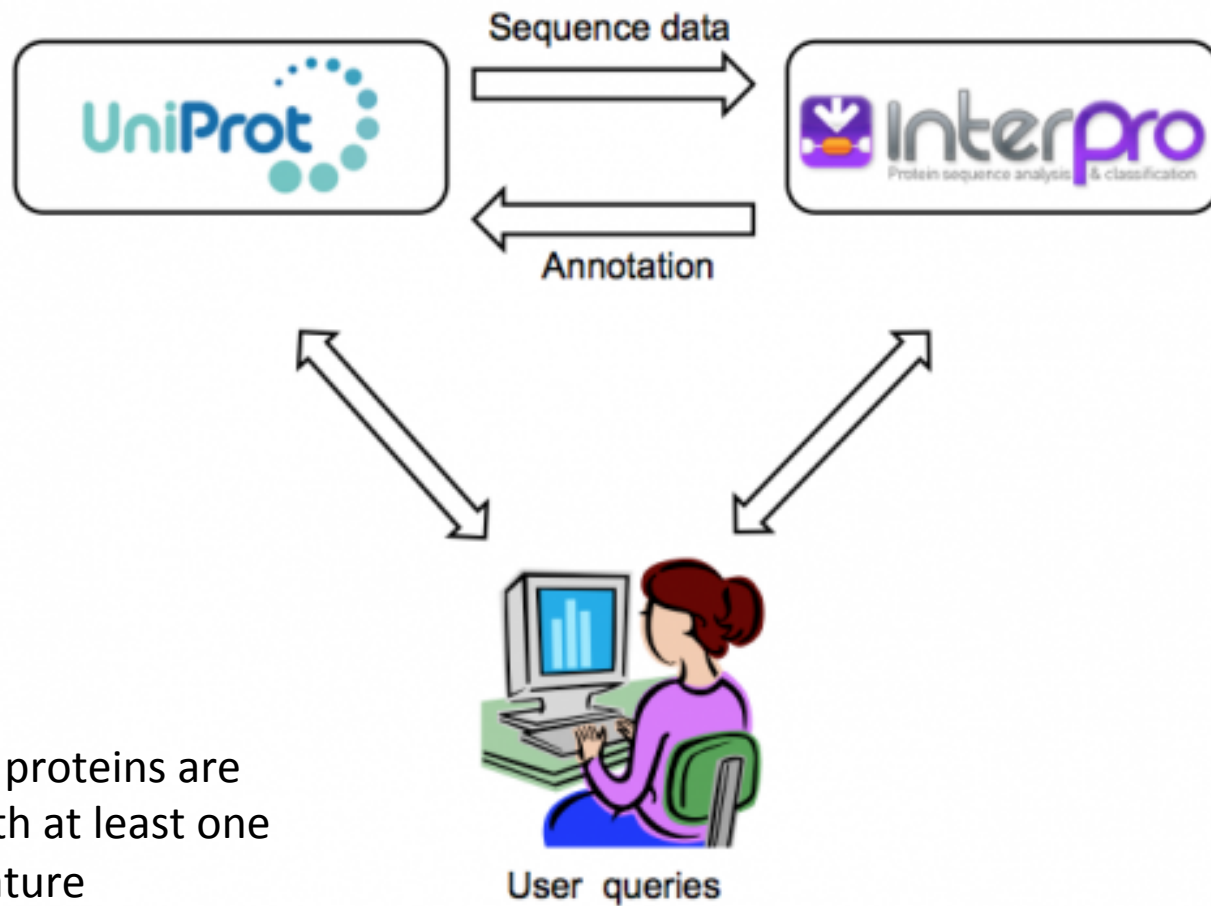


## InterPro components

1. CATH/Gene3D      University College, London, UK
2. PANTHER      University of Southern California, CA, USA
3. PIRSF      Protein Information Resource, Georgetown University, USA
- ★ 4. Pfam      Wellcome Trust Sanger Institute, Hinxton, UK
5. PRINTS      University of Manchester, UK
6. ProDom      PRABI Villeurbanne, France
7. PROSITE      Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland
- ★ 8. SMART      EMBL, Heidelberg, Germany
9. SUPERFAMILY      University of Bristol, UK
- ★ 10. TIGRFAMs      J. Craig Venter Institute, Rockville, MD, US
11. HAMAP      Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland

## CDD components

Pfam, SMART, TIGRFAM,  
COG, KOG, PRK, CD, LOAD



Most UniProt proteins are annotated with at least one InterPro signature

Sequence database	Version	Count	Count of proteins matching	
			any signature	integrated signatures
UniProtKB	2014_07	80370243	71766615 (89.3%)	67116794 (83.5%)
UniProtKB/TrEMBL	2014_07	79824243	71234772 (89.2%)	66591418 (83.4%)
UniProtKB/Swiss-Prot	2014_07	546000	531843 (97.4%)	525376 (96.2%)

Each InterPro entry is assigned one of a number of types which tell you what you can infer when a protein matches the entry.

The entry types are:



## Family

A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.



## Domain

Domains are distinct functional, structural or sequence units that may exist in a variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain.



## Repeat

A match to an InterPro entry of this type identifies a short sequence that is typically repeated within a protein.

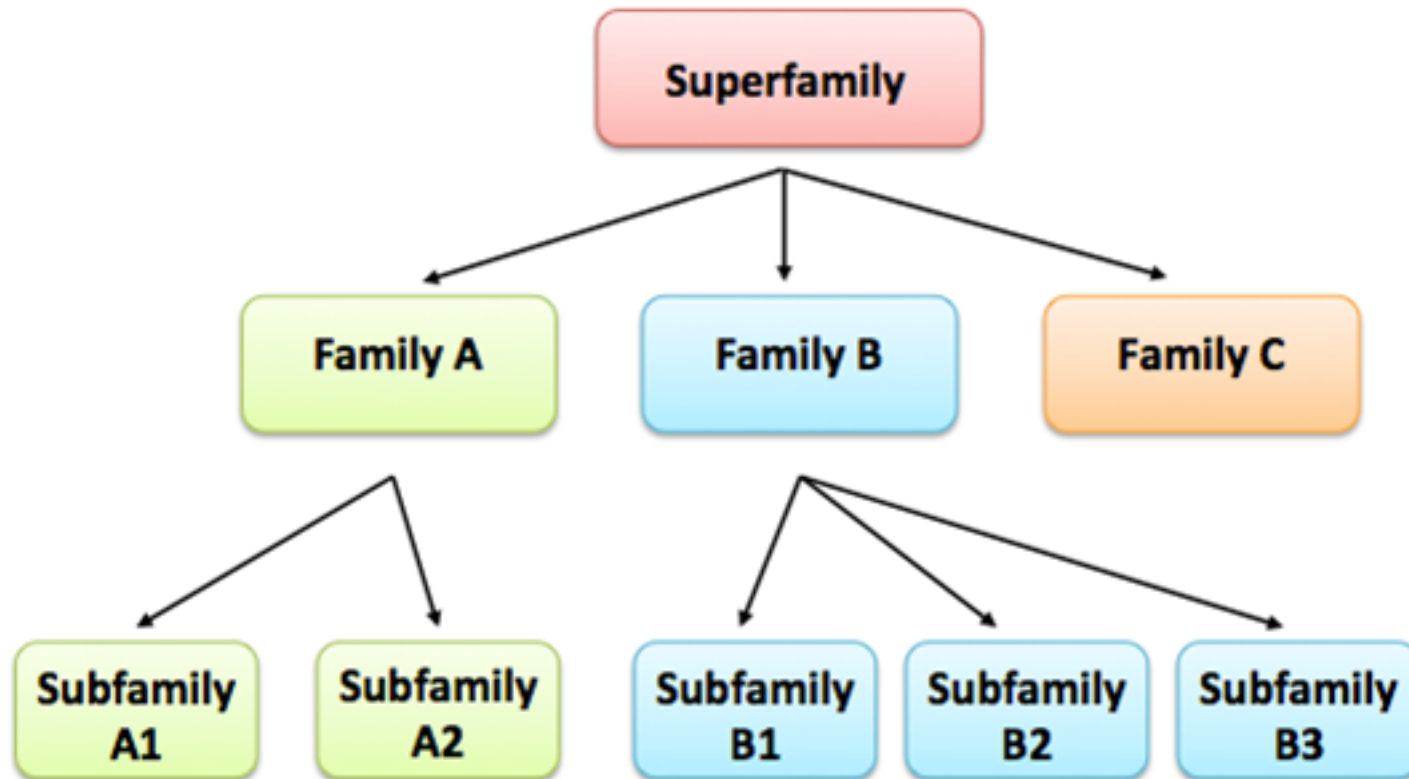


## Site

A match to an InterPro entry of this type indicates a short sequence that contains one or more conserved residues. The type of sites covered by InterPro are active sites, binding sites, post-translational modification sites and conserved sites.



Protein families are often arranged into hierarchies, with proteins that share a common ancestor subdivided into smaller, more closely related groups. The terms superfamily (describing a large group of distantly related proteins) and subfamily (describing a small group of closely related proteins) are sometimes used in this context



# Protein Classification

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are family, superfamily and fold, described below.

## **Family: Clear evolutionarily relationship**

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater.

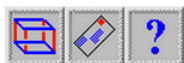
## **Superfamily: Probable common evolutionary origin**

Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies.

## **Fold: Major structural similarity**

Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

<http://scop.mrc-lmb.cam.ac.uk/scop/intro.html>



Welcome to **SCOP**: Structural Classification of Proteins.

**1.75 release** (June 2009)

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).

Folds, superfamilies, and families [statistics here](#).

[New folds](#) [superfamilies](#) [families](#).

[List of obsolete entries and their replacements](#).

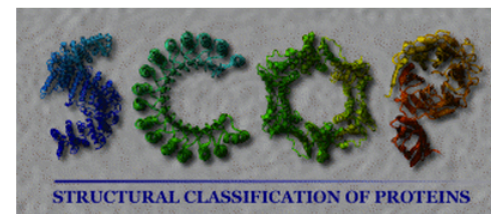
**Authors.** Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. [scop@mrc-lmb.cam.ac.uk](mailto:scop@mrc-lmb.cam.ac.uk)

**Reference:** Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [\[PDF\]](#)

**Recent changes** are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [\[PDF\]](#),

Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [\[PDF\]](#), and

Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* 2008 36: D419-D425; doi:10.1093/nar/gkm993 [\[PDF\]](#).

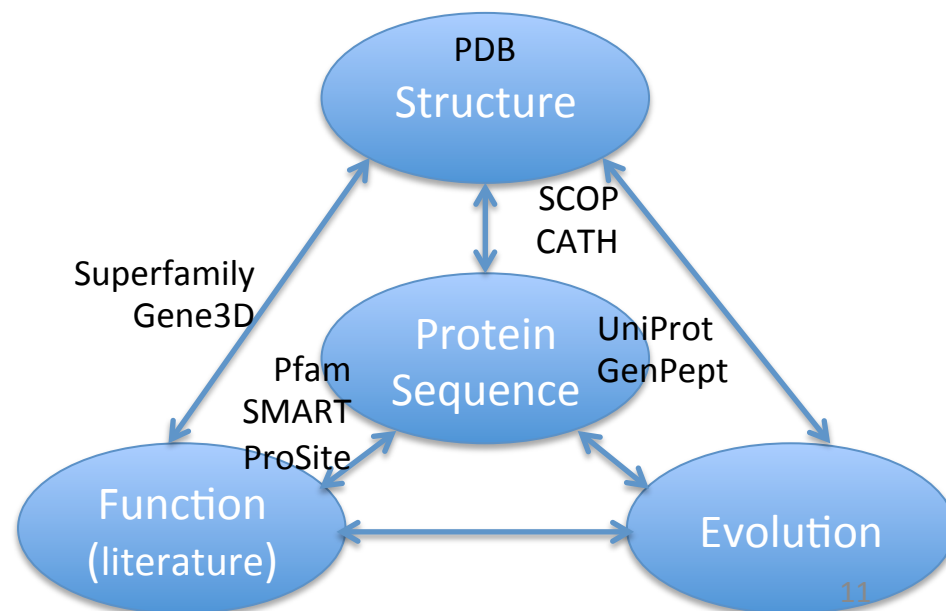


## Postdoc Wanted

- Want to help us design and build the next generation of SCOP and ASTRAL?  
[Get more details and apply here.](#)

## Access methods

- Enter scop at the [top of the hierarchy](#)
- [Keyword search of SCOP entries](#)
- [SCOP parseable files](#)
- [All SCOP releases and reclassified entry history](#)
- [pre-SCOP - preview of the next release](#)
- SCOP domain sequences and pdb-style coordinate files ([ASTRAL](#))
- Hidden Markov Model library for SCOP superfamilies ([SUPERFAMILY](#))
- Structural alignments for proteins with non-trivial relationships ([SISYPHUS](#))



# CATH / Gene3D

26 million protein domains classified into 2,738 superfamilies

[Browse »](#)

[Search »](#)

[Download »](#)

[Ta](#)

## What is CATH?

**CATH is a classification of protein structures downloaded from the Protein Data Bank.**

We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- [Search CATH by text, ID or keyword](#)
- [Search CATH by protein sequence \(FASTA\)](#)
- [Search CATH by PDB structure](#)
- [Browse CATH Hierarchy](#)
- [CATH Release Notes](#)
- [CATH Tutorials](#)

## Example pages

- [PDB "2bop"](#)
- [Domain "1cukA01"](#)
- [Relatives of "1cukA01"](#)
- [Superfamily "HUPs"](#)
- [Functional Family](#)
- [FunFam Alignment](#)
- [Search for "enolase"](#)
- [Superfamily Comparison](#)

## Citing CATH

If you find this resource useful, please consider citing the reference that describes this work:










## Latest Release Statistics

**CATH v4.0** based on PDB dated March 26, 2013

235,858	<a href="#">CATH Domains</a>
2,738	<a href="#">CATH Superfamilies</a>
69,058	<a href="#">Annotated PDBs</a>

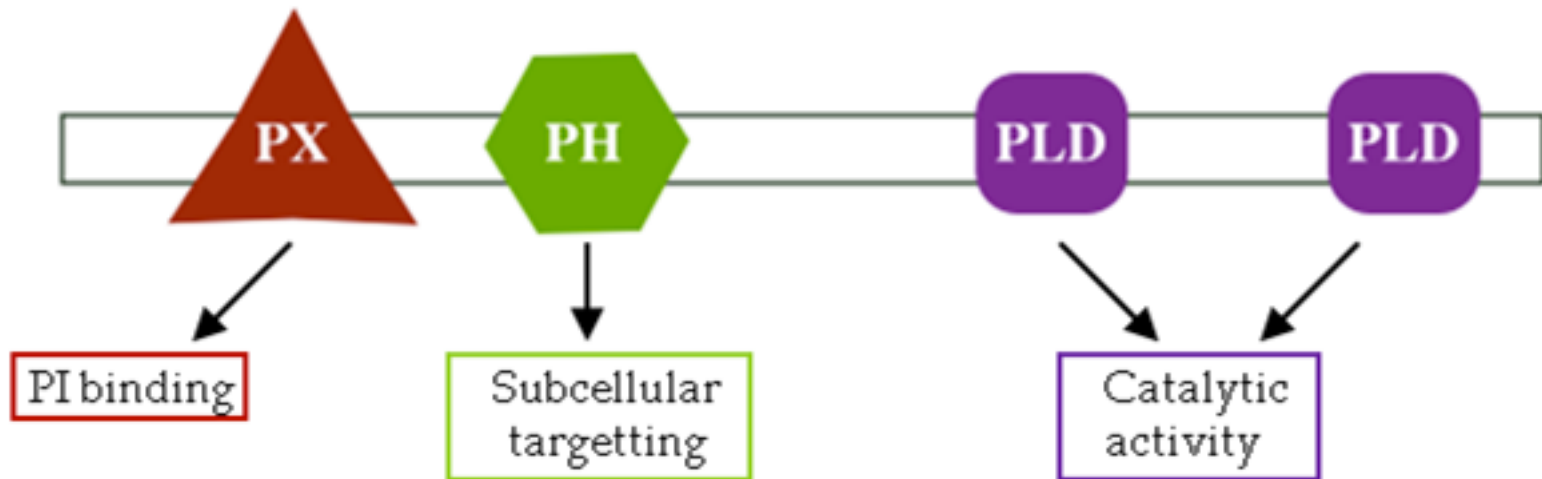
**Gene3D v12** released March 18, 2012

6,131	Cellular Genomes
21,662,155	Protein Sequences
25,615,754	CATH Domain Predictions

Depth	Letter	Name	Clustering criteria
1		Class	Secondary structure content
2		Architecture	General spatial arrangement of secondary structures
3		Topology	Spatial arrangement and connectivity of secondary structures (fold)
4		Homologous Superfamily	Manual curation of evidence of evolutionary relationship (at least two criteria)
5		Sequence Family (S35)	>= 35% sequence similarity
6		Orthologous Family (S60) *	>= 60% sequence similarity
7		âLikeâ domain (S95) *	>= 95% sequence similarity
8		Identical domain (S100)	100% sequence similarity
9		Domain counter	Unique domains

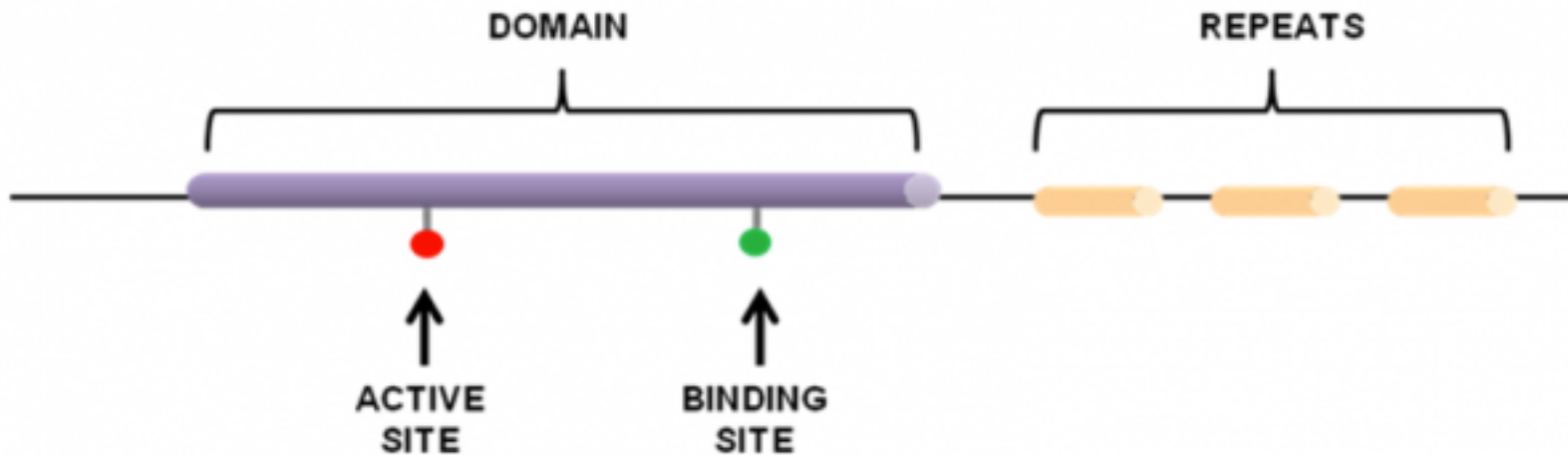
fold ~ class – superfamily ~ clan – family – subfamily – domain sequence

Family- and domain-based classifications are not always straightforward and can overlap, since proteins are sometimes assigned to families by virtue of the domain(s) they contain. An example of this kind of complexity is outlined below



Domain composition of phospholipase D1, which is an enzyme that breaks down phosphatidylcholine. The protein contains a PX (phox) domain that is involved in binding phosphatidylinositol, a PH (pleckstrin homology) domain that has a role in targeting the enzyme to particular locations within the cell, and two PLD (phospholipase D) domains responsible for the protein's catalytic activity

Sequence features differ from domains in that they are usually quite small (often only a few amino acids long), whereas domains represent entire structural or functional units of the protein (see Figure). Sequence features are often nested within domains – a protein kinase domain, for example, usually contains a protein kinase active site



Sequence features are groups of amino acids that confer certain characteristics upon a protein, and may be important for its overall function. Such features include:

active sites, which contain amino acids involved in catalytic activity.

binding sites, containing amino acids that are directly involved in binding molecules or ions.

post-translational modification (PTM) sites, which contain residues known to be chemically modified (phosphorylated, palmitoylated, acetylated, etc) after the process of protein translation.

repeats, which are typically short amino acid sequences that are repeated within a protein, and may confer binding or structural properties upon it.

# Hands on exercise 1: search against protein family databases



http://www.ebi.ac.uk/interpro/

<http://cys.bios.niu.edu/yyin/teach/PBB/csl-pr.fa>, put the first sequence in the search box

Hit Search; take about 1 min

Read more about InterPro

The screenshot shows the InterPro website interface. At the top, there is a search bar with the text "Search InterPro..." and a "Search" button. Below the search bar, there are navigation links: Home, Search, Release notes, Download, About InterPro, Help, and Contact. The main heading is "InterPro: protein sequence analysis & classification". Below this, there is a paragraph describing InterPro's functionality. A search box labeled "Analyse your protein sequence" contains the sequence: `>AT5G22740.1|AT5G22740.1|cs|A  
MDGVSPKFLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLVKLFWK  
KPKDKRYKFEPIHDDEELGSSNFPVVLVQIPMFNEREVYKLSIGAACGLSWPDRDLVIQVLDSDTPTVKQMVE  
VECQRWASKGINIRYQIRENRVGYKAGALKEGLKRSYVKHCEYVVFADDFQPEPDFLRRSIPFLMHNPNIALV  
QARWRVNSDECLTRMQEMSLDYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRRTTVEDMD  
LAVRASLRGWKFLYLGLDQVKSELSTPSTFRFRFQQHRWSCGPANLFRKMMVEIVRNKVKVRFWKKVYIYSF`. Below the search box are buttons for "Search", "Clear", and "Example protein sequence". On the right side, there is a "Read more about InterPro" link with a red arrow pointing to it. Below the search box, there are three columns: "Documentation" with links for "About InterPro" and "FAQs"; "Protein focus" with a link for "Dionysian mysteries - the aldehyde dehydrogenase (ALDH) family" and a red arrow pointing to the search box; and "Publications" with a link for "InterProScan 5: genome-scale protein function classification". On the far right, there are three promotional banners: "InterPro v48.0 17th July 2014" with "Download" and "Read more" buttons; "IDA Domain organisation search" with a "Search >>" button; and "Interproscan 5" with a "Learn more >>" button. At the bottom right, there is a "Tweets" section with a tweet from InterPro (@InterProDB) dated 18 Aug, mentioning "InterProScan 5 (version 5.7-48.0) is now available." and a "Follow" button.

# Release notes

## Latest release note

[http://www.ebi.ac.uk/interpro/release\\_notes.html](http://www.ebi.ac.uk/interpro/release_notes.html)



**InterPro 48.0**  
**17th July 2014**

New features include:

- Integration of 294 new methods from the CATH-Gene3D, PANTHER, Pfam, ProDom and SUPERFAMILY databases.

[Previous release notes](#)

## Contents and coverage of InterPro 48.0

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. The following statistics are for all UniProtKB proteins. InterPro release 48.0 contains 26238 entries (last entry: [IPR029787](#)), representing:

**F** Family (17620)

**D** Domain (7497)

**R** Repeat (277)

**S** Sites

    i. Active site (108)

    i. Binding site (73)

    i. Conserved site (647)

    i. PTM (16)

InterPro cites 41206 publications in PubMed.

## Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	3.5.0	<a href="#">2626</a>	<a href="#">1718</a>
HAMAP	201311.27	<a href="#">1916</a>	<a href="#">1912</a>
PANTHER	9.0	<a href="#">59948</a>	<a href="#">3673</a>

Click to link to InterPro page of this domain

Click to link to individual database website

The screenshot shows a protein entry page for AT5G22740.1|AT5G22740.1|CSLA. The protein length is 534 amino acids. The page is annotated with a domain from InterPro (IPR029044) and several unintegrated signatures. A filter sidebar on the left allows filtering by entry type (Family, Domains, Repeats, Site) and status (Unintegrated). A 'GO term prediction' section is also visible. Red arrows point from the text above to the domain link and the unintegrated signatures.

**Overview**

Similar proteins  
Structures

Filter view on

Entry type

- Family
- Domains
- Repeats
- Site

Status

- Unintegrated

Colour by [help](#)

- domain relationship
- source database

**Protein**

AT5G22740.1|AT5G22740.1|CSLA

Length 534 amino acids

Export [↓](#) Select format [↓](#)

**Protein family membership**

None predicted.

**Domains and repeats**

Domain

1 50 100 150 200 250 300 350 400 450 500 534

**Detailed signature matches**

**IPR029044** Nucleotide-diphospho-sugar transferases

- ▶ SSF53448 (Nucleotid...)
- ▶ G3DSA:3.90.55...


**no IPR** Unintegrated signatures


- ▶ CYTOPLASMIC\_D... (C...)
- ▶ NON\_CYTOPLASM... (N...)
- ▶ PF13641 (Glyco\_tran...)
- ▶ PTHR32044 (FAMILY N...)
- ▶ PTHR32044:SF6 (GLUC...)
- ▶ TMhelix
- ▶ TRANSMEMBRANE (Tran...)

**GO term prediction**

These are individual family/domain matches not integrated in InterPro

This is linked from the previous page: the InterPro page to describe IPR029044

EMBL-EBI  Services Research Training Abo

 **InterPro**  
Protein sequence analysis & classification

Search InterPro...   
Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

Home Search Release notes Download About InterPro Help Contact

**Overview**

- Proteins matched (427828)
- Domain organisations (2216)
- Pathways & Interactions
- Species
- Structures
- Literature (1)
- Cross-references (1)





## D Domain

### Nucleotide-diphospho-sugar transferases (IPR029044)

Short name: *Nucleotide-diphossugar\_trans*

### Domain relationships


#### **Nucleotide-diphospho-sugar transferases (IPR029044)**

-  **2-C-methyl-D-erythritol 4-phosphate cytidyltransferase (IPR001228)**
-  **Glycosyltransferase 2-like (IPR001173)**
-  **MobA-like NTP transferase domain (IPR025877)**
-  **Nucleotidyl transferase (IPR005835)**

### Description



This entry represents a domain with a Rossmann like fold and can be found in diverse glycosyltransferases.

The biosynthesis of disaccharides, oligosaccharides and polysaccharides involves the action of hundreds of different glycosyltransferases. These enzymes catalyse the transfer of sugar moieties from activated donor molecules to specific acceptor molecules, forming glycosidic bonds. A classification of glycosyltransferases using nucleotide diphospho-sugar, nucleotide monophospho-sugar and sugar phosphates ([EC:2.4.1.-](#)) and related proteins into distinct sequence based families has been described ([PMID: 9334165](#)). This classification is available on the CAZy (CARbohydrate-Active EnZymes) web site. The same three-dimensional fold is expected to occur within each of the families. Because 3-D structures are better conserved than sequences, several of the families defined on the basis of sequence similarities may have similar 3-D structures and therefore form 'clans'.

 Add your annotation

### Contributing signatures

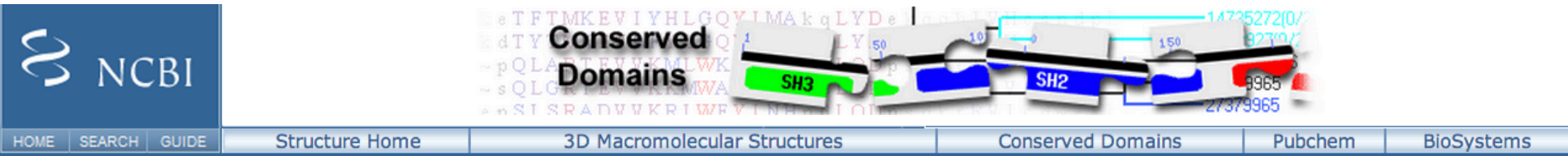
Signatures from InterPro member databases are used to construct an entry.

-  **GENE3D** ⓘ  
[G3DSA:3.90.550.10](#)  
(G3DSA:3.90.550.10)
-  **SUPERFAMILY** ⓘ  
[SSF53448](#) (SSF53448)

Scientific literature for this IPR family

<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>

NCBI's Conserved Domain Database (CDD): equivalent to InterPro of EBI, much faster, but integrate less member databases



NCBI

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems

Search for Conserved Domains within a protein or coding nucleotide sequence

**NEW!** Use **Batch CD-search** to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in [FASTA format](#) ?

Submit Reset

**OPTIONS**

Search against database ?  CDD v3.11 - 45746 PSSMs  
 Pfam v27.0 - 14831 PSSMs  
 SMART v6.0 - 1013 PSSMs  
 KOG v1.0 - 4825 PSSMs  
 COG v1.0 - 4873 PSSMs  
 PRK v6.9 - 10885 PSSMs  
 TIGR v13.0 - 4284 PSSMs

Expect Value ? threshold:

Apply low-complexity filter

Composition based statistic

Force live search ?

Maximum number of hits ?

Result mode  Concise ?  Standard ?  Full ?

Search for Conserved Domains within a protein or coding nucleotide sequence

**NEW!** Use **Batch CD-search** to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in [FASTA format](#) ?

Submit

Reset

#### OPTIONS

Search against database ?

- CDD v3.11 - 45746 PSSMs
- Pfam v27.0 - 14831 PSSMs
- SMART v6.0 - 1013 PSSMs
- KOG v1.0 - 4825 PSSMs
- COG v1.0 - 4873 PSSMs
- PRK v6.9 - 10885 PSSMs
- TIGR v13.0 - 4284 PSSMs

Expect Value ? threshold:

Apply low-complexity filter

Composition based statistic

Force live search ?




Maximum number of hits ?

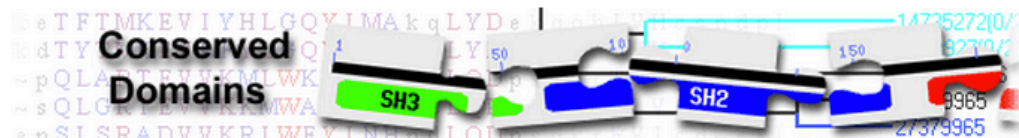
Result mode  Concise ?  Standard ?  Full ?

#### Retrieve previous CD-search result

Request ID:   ?

#### References:

-  Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", **Nucleic Acids Res.**39(D)225-9.
-  Marchler-Bauer A et al. (2009), "CDD: specific functional annotation with the Conserved Domain Database.", **Nucleic Acids Res.**37(D)205-10.
-  Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

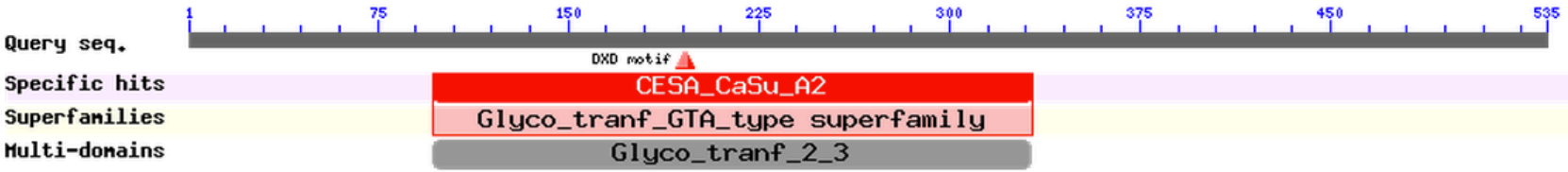


## Conserved domains on [AT5G22740.1|AT5G22740.1]

[View full result](#)

csIA

### Graphical summary [show options](#)



[Search for similar domain architectures](#) [Refine search](#)

### List of domain hits

	Description	PssmId	Multi-dom	E-value
CESA_CaSu_A2[cd06437], Cellulose synthase catalytic subunit A2 (CESA2) is a catalytic subunit or a catalytic subunit substitute of the cellulose synt		133059	yes	2.38e-142
Glyco_tranf_2_3[pfam13641], Glycosyltransferase like family 2; Members of this family of prokaryotic proteins include putative glucosyltransferase, ...		205818	yes	1.53e-25

### References:

- Marchler-Bauer A et al. (2013), "CDD: conserved domains and protein three-dimensional structure.", **Nucleic Acids Res.**41(D1)348-52.
- Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", **Nucleic Acids Res.**39(D)225-9.
- Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

[Help](#) | [Disclaimer](#) | [Write to the Help Desk](#)  
 NCBI | NLM | NIH

# Genome browser: ENSEMBL

<http://www.ensembl.org/>

**The Ensembl project aims to automatically *annotate* genome sequences, *integrate* these data with other biological information and to make the results freely available to geneticists, molecular biologists, bioinformaticians and the wider research community. Ensembl is jointly headed by Dr Stephen Searle at the Wellcome Trust *Sanger Institute* and Dr Paul Flicek at the European Bioinformatics Institute (*EBI*).**

Ensembl <sup>east</sup> | BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors | Login/Register

Search: All species for  Go  
e.g. BRCA2 or rat 5:62797383-63627669 or coronary heart disease

### Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

#### Popular genomes

- Human** GRCh38
- Mouse** GRCm38
- Zebrafish** Zv9

★ [Log in to customize this list](#)

#### All genomes

-- Select a species --

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Pre!](#) and [EnsemblGenomes](#)

#### ENCODE data in Ensembl

#### Variant Effect Predictor

#### Gene expression in different tissues

#### Find SNPs and other variants for my gene

```
GTATATACATT  
CTRAAAGTCTT  
CTTCTAATTCT  
...  
GCTGACTTCGGGTGG  
GGCTTGTGGCGGAGC  
GGCCCTCTGCTGGCCT  
AGGGACAGATTTGTGA  
CACCTCTGGAGCGGTT  
CCAGTCCAGCGTGGC
```

#### Retrieve gene sequence

#### Compare genes across species

#### Use my own data in Ensembl

#### Learn about a disease or phenotype

### What's New in Release 76 (August 2014)

- [Updated human assembly to GRCh38](#)
- [New BLAST/BLAT interface](#)
- [New regulation displays](#)
- New species: [Amazon molly](#) and [Olive baboon](#)

[Full details of this release](#)  
[All web updates by release](#)  
[More release news on our blog](#)

#### Latest blog posts

- 27 Aug 2014: [What's coming in Ensembl release 77](#)
- 20 Aug 2014: [The Ensembl Regulatory Build – the Track Hub](#)
- 19 Aug 2014: [Ensembl Genomes release 23 is out!](#)

[Go to Ensembl blog](#)

#### Did you know...?

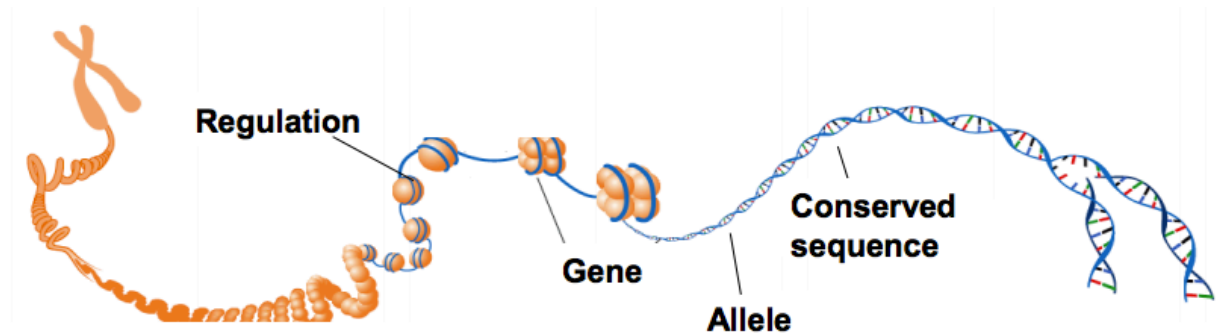
Have some data? [View it in Ensembl](#) by uploading a file or attaching a url.



## What do we need in genome browsers?

To make the bare DNA **sequence**, its properties, and the associated **annotations** more accessible through graphical interface.

Genome browsers provide access to large amounts of sequence data via a graphical user interface. They use a **visual, high-level overview of complex data** in a form that can be grasped at a glance and provide the means to **explore the data in increasing resolution** from megabase scales down to the level of individual elements of the DNA sequence.

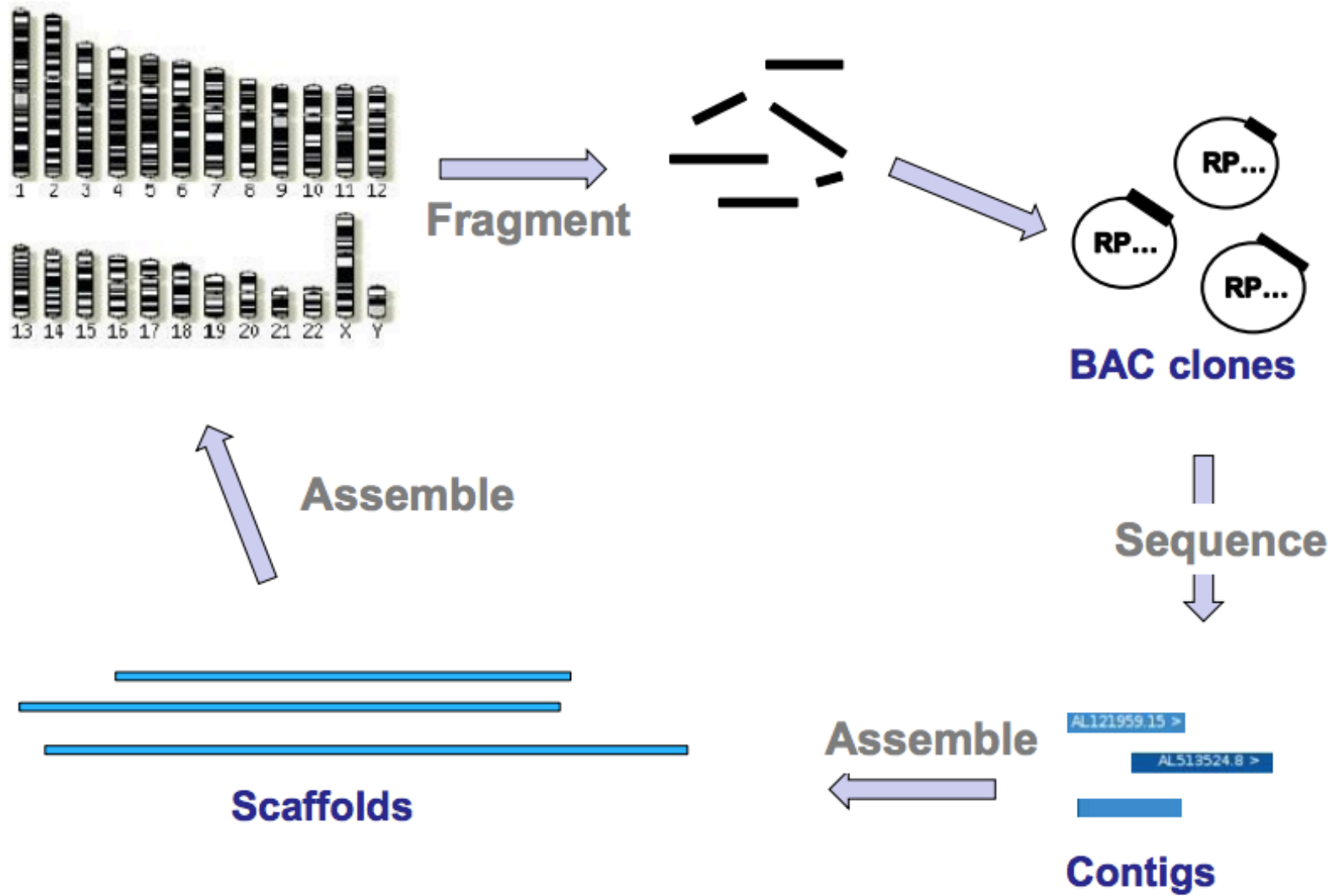


- **Splice variants, proteins, non-coding RNA**
- **Small and large scale sequence variation, phenotype associations**
- **Whole genome alignments, protein trees**
- **Potential promoters and enhancers, DNA methylation**
- **User upload, custom data**

Short tutorial videos introducing ENSEMBL

<http://useast.ensembl.org/info/website/tutorials/index.html>

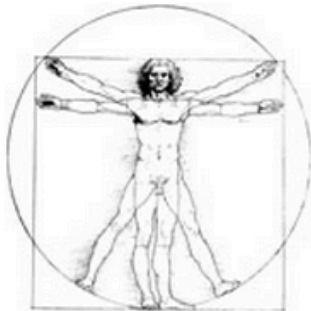
# Genome Sequencing



<http://useast.ensembl.org/info/website/tutorials/index.html>

## Genome Assemblies

The GRC has built tools to facilitate the curation of genome assemblies based on the sequence overlaps of long, high quality sequences (Clones and PCR products, not currently supports production of assemblies for human, mouse or zebrafish. If your assembly data fits this model and you are interested in using these tools please contact [Subscribe](#) to the grc-announce email list to receive email notification for all GRC assembly updates.



### Human

The human genome assembly was produced as part of the [Human Genome Project \(HGP\)](#). The previous assembly (NCBI36) was the last one produced by the HGP and was described in 2004 ([PMID: 15496913](#)); this was the starting point for the GRC. The assembly is based largely on assembling overlapping clone sequences.

#### Human assembly information

Current Major Assembly	GRCh38
Regions with Alternate Loci	178
Assembly N50	67,794,873 bp
Remaining Gaps	875

[More human assembly statistics...](#)

The Genome Reference Consortium consists of:



# 1000 Genomes

A Deep Catalog of Human Genetic Variation

*Nature 491, 56-65 ( 01 November 2012 )*

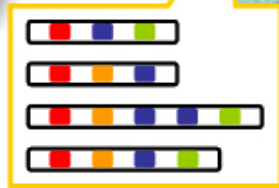
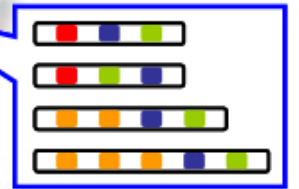
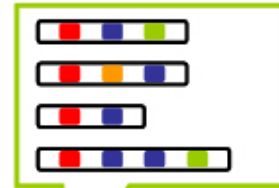
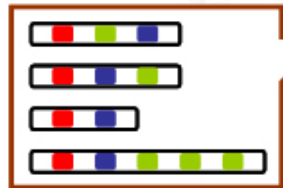
Insertion 

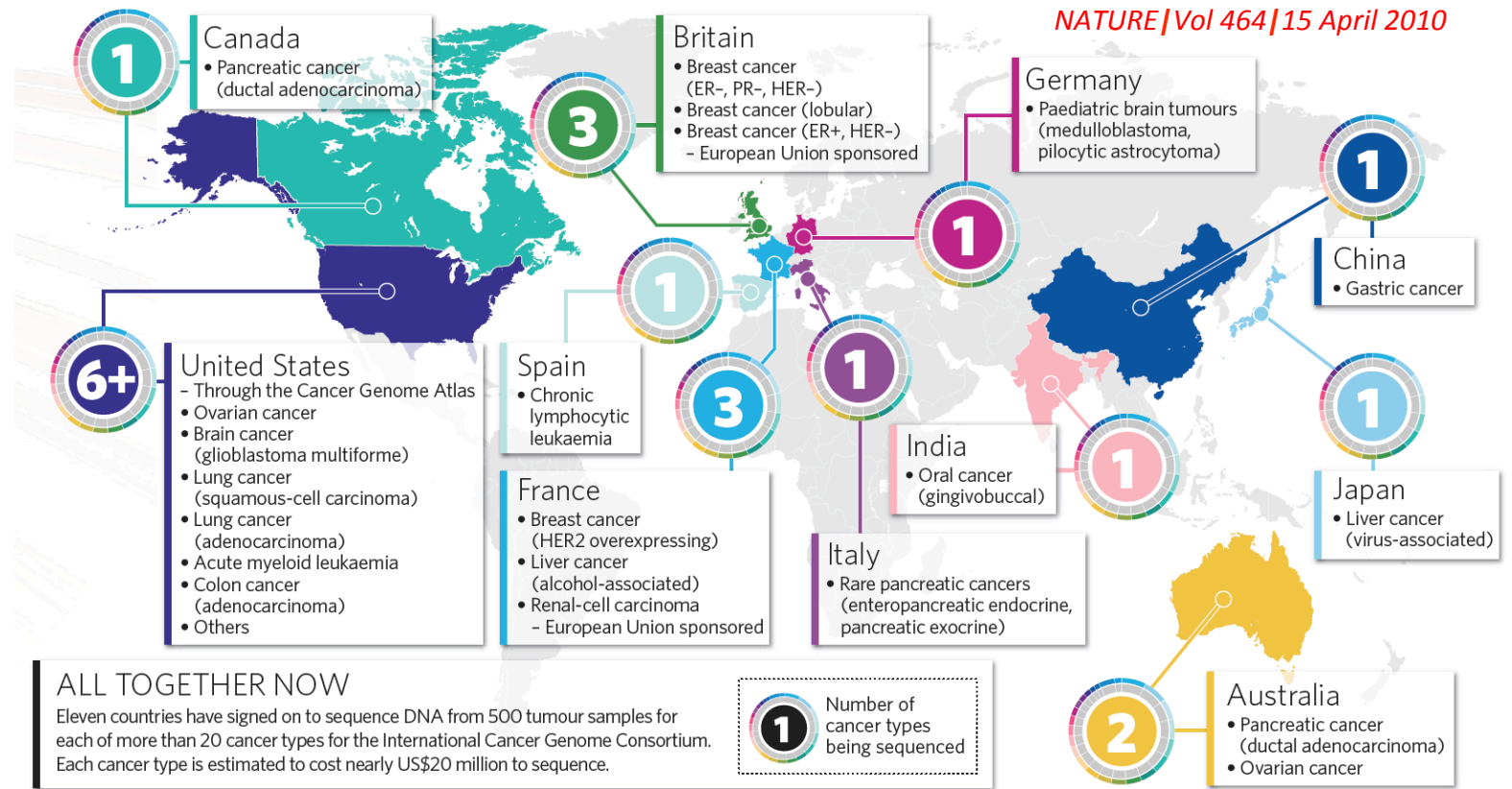
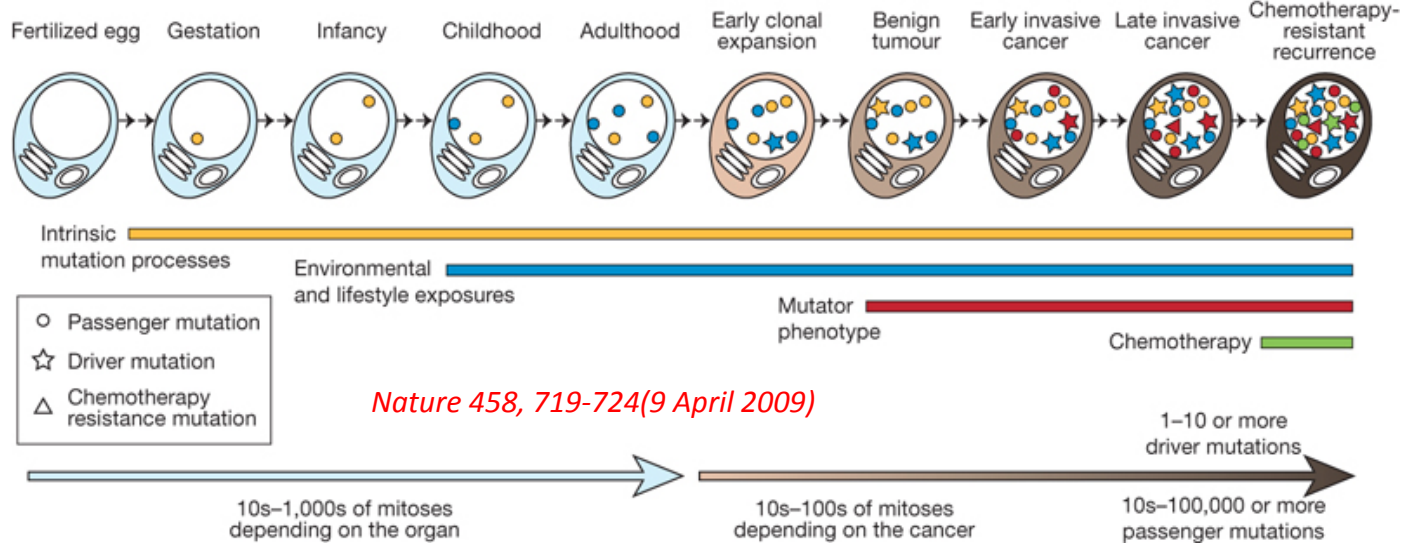
Deletion 

Copy Number Variant 

Inversion 

Reference 





While a user may start browsing for a **particular gene**, the user interface will display the area of the genome containing the gene, along with a broader context of other information available in the region of the chromosome occupied by the gene.

This information is shown in “**tracks**,” with each track **showing either the genomic sequence from a particular species or a particular kind of annotation on the gene**. The tracks are aligned so that the information about a particular base in the sequence is lined up and can be viewed easily.

In modern browsers, the abundance of **contextual information linked to a genomic region** not only helps to satisfy the most directed search, but also makes available a depth of content that facilitates **integration of knowledge about genes, gene expression, regulatory sequences, sequence conservation between species, and many other classes of data**.

- Ensembl Genome Browsers: <http://www.ensemblgenomes.org>
- NCBI Map Viewer: <http://www.ncbi.nlm.nih.gov/mapview/>
- UCSC Genome Browser: <http://genome.ucsc.edu>

Each uses a centralized model, where the web site provides access to a large public database of genome data for many species and also integrates specialized tools, such as BLAST at NCBI and Ensembl and BLAT at UCSC.

The public browsers provide a valuable service to the research community by providing **tools** for free access to whole genome data and by supporting the complex and robust **informatics infrastructure** required to make the data accessible



# Hands on exercise 2: Ensembl gene search

Click to link to human page

**Ensembl**<sup>east</sup> | BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search:  for

e.g. **BRCA2** or **rat 5:62797383-63627669** or **coronary heart disease**

### Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

#### Popular genomes

- Human**  
GRCh38
- Mouse**  
GRCm38
- Zebrafish**  
Zv9

★ [Log in to customize this list](#)

#### All genomes

-- Select a species --

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Pre!](#) and [EnsemblGenomes](#)

Ensembl supports data from external projects through [Datahubs](#)

#### ENCODE data in Ensembl

#### Variant Effect Predictor

#### Gene expression in different tissues

#### Find SNPs and other variants for my gene

#### Retrieve gene sequence

```
GCCTGACTTCGGGATGG:  
GGGCTTGTGGGCGGAGCC  
GGGCTCTGCTGGCCCT:  
AGGGACAGATTGTGGA:  
CACCTCTGGAGCGGTT:  
CCCACTCCAGCGTGCCG:
```

#### Compare genes across species

#### Use my own data in Ensembl

#### Learn about a disease or phenotype

#### What's New in Release 76 (Aug 2014)

- Updated human assembly to GF
- New BLAST/BLAT interface
- New regulation displays
- New species: Amazon molly and

[Full details of this release](#)

[All web updates, by release](#)

[More release news on our blog ->](#)

#### Latest blog posts

- 27 Aug 2014: [What's coming in](#)
- 20 Aug 2014: [The Ensembl Reg](#)
- 19 Aug 2014: [Ensembl Genome](#)

[Go to Ensembl blog ->](#)

#### Did you know...?

Lost ... try our [tutorials](#)



Ensembl is a joint project between [EMBL - EBI](#) and the [Wellcome Trust Sanger Institute](#) to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl receives major funding from the Wellcome Trust. Our [acknowledgements page](#) includes a list of additional current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

Put "liver cancer" in the search box and Go

Human (GRCh38) ▼



# Human

*Homo sapiens*

Search all categories ▼

liver cancer

Go

e.g. [BRCA2](#) or [17:63973115-64437414](#) or [osteoarthritis](#)

## Genome assembly: GRCh38 (GCA\_000001405.15)



More information and statistics



Download DNA sequence (FASTA)



Convert your data to GRCh38 coordinates



Display your data in Ensembl

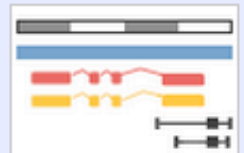
### Other assemblies

GRCh37 (Long-term archive with BLAST, VEP and BioMart) ▼

Go



View karyotype



Example region

## Comparative genomics



This keyword search gives everything that contains “liver cancer”

Human (GRCh38) ▾

Current selection: < all Species  
Only searching Human

Only searching Human ▾ liver cancer 🔍

259952 results match liver cancer when restricted to species: Human ✕

**Restrict category to:**

Gene	341
Transcript	941
Variation	29830
Phenotype	224
Somatic Mutation	69274
GenomicAlignment	159234
Protein Domain	11
Protein Family	97

Per page: 10 25 50 100

Layout: Standard **Table**

Tip: Help and Documentation can be searched from the homepage! Just type in a term you want to know more about, like non-synonymous

**HULC (Human Gene)**  
ENSG00000276019 6:8653558-8653797:1  
Highly up-regulated in liver cancer conserved region [Source:RFAM;Acc:RF02101] HULC (RFAM record with a description of Highly up-regulated in liver cancer conserved region) is associated with Gene ENSG00000276019  
Variation table • Location • Regulation • Orthologues • Gene tree

**DLC1-013 (Human Transcript)**  
ENST00000517333 8:13499822-13515658:-1  
Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897] .  
Location • cDNA seq. • Variation table • Population

**DLC1-018 (Human Transcript)**  
ENST00000521730 8:13086232-13088739:-1  
Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897] .  
Location • cDNA seq. • Variation table • Population

**DLC1-010 (Human Transcript)**  
ENST00000506171 8:13214267-13276310:-1  
Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897] .  
Location • cDNA seq. • Variation table • Population

**HULC (Human Gene)**  
ENSG00000251164 6:8652137-8653846:1  
Hepatocellular carcinoma up-regulated long non-coding RNA [Source:HGNC Symbol;Acc:HGNC:34232] HIGHLY UPREGULATED IN LIVER CANCER [\*612210] (MIM gene record with a description of HIGHLY UPREGULATED IN LIVER CANCER; HULC) is associated with Gene ENSG00000251164  
Variation table • Location • Regulation • Orthologues • Gene tree

**DLC1-019 (Human Transcript)**  
ENST00000517868 8:13498859-13499241:-1  
Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897] .  
Location • cDNA seq. • Variation table • Protein seq. • Population • Protein

Click on Table to have a table view

This col tells the category of the entry

Current selection:

< all Species  
Only searching Human

Restrict category to:

- Gene 341
- Transcript 341
- Variation 29830
- Phenotype 224
- Somatic Mutation 69274
- GenomicAlignment 159234
- Protein Domain 11
- Protein Family 97

Per page:

10 25 50 100

Layout:

Standard Table

Show 10 entries

ID	Name	Species	Category	Description
<a href="#">ENSG00000276019</a>	HULC	Human	Gene	Highly up-regulated in liver cancer conserved region [Source:RFAM; description of Highly up-regulated in <i>liver cancer</i> conserved region]
<a href="#">ENST00000517393</a>	DLC1-013	Human	Transcript	Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897]
<a href="#">ENST00000521730</a>	DLC1-018	Human	Transcript	Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897]
<a href="#">ENST00000506171</a>	DLC1-010	Human	Transcript	Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897]
<a href="#">ENSG00000251164</a>	HULC	Human	Gene	Hepatocellular carcinoma up-regulated long non-coding RNA [Source:UPREGULATED IN <i>LIVER CANCER</i> [*612210] (MIM gene record with <i>LIVER CANCER</i> ; HULC) is associated with Gene ENSG00000251164]
<a href="#">ENST00000517868</a>	DLC1-019	Human	Transcript	Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897]
<a href="#">ENST00000513883</a>	DLC1-007	Human	Transcript	Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897]
<a href="#">ENST00000612720</a>	HULC.1-201	Human	Transcript	Highly up-regulated in liver cancer conserved region [Source:RFAM; description of Highly up-regulated in <i>liver cancer</i> conserved region]
<a href="#">ENST00000510250</a>	DLC1-017	Human	Transcript	Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897]
<a href="#">ENST00000510318</a>	DLC1-006	Human	Transcript	Deleted in liver cancer 1 [Source:HGNC Symbol;Acc:HGNC:2897]

<< < 1 2 3 4 5 6 7 8 ... 25995 25996 > >>

Click on the numbers to only show gene entries

Click here to show the list and select Location and Score to show chromosome location info and score respectively

This is the list of genes

Current selection: < all Species Only searching Human < all Categories Only searching Gene

Per page: 10 25 50 100

Layout: Standard Table

Tip: You can choose which results appear near the top of your search by updating your favourite species.

Show 10 entries

ID	Name	Location	Species	Category	ID	Description	Score
<a href="#">ENSG00000276019</a>	HULC	6:8653558-8653797:1	Human	Gene	<input checked="" type="checkbox"/>	Highly up-regulated in liver cancer conserved region [Source:RFAM;Acc:RF02101] HULC FAM record with a description of Highly up-regulated in <b>LIVER CANCER</b> conserved region) is associated with Gene ENSG00000276019	0.42055306
<a href="#">ENSG00000251164</a>	HULC	6:8652137-8653846:1	Human	Gene	<input checked="" type="checkbox"/>	Epitaxial carcinoma up-regulated long non-coding RNA [Source:HGNC Symbol;Acc:HGNC:34232] HIGHLY UPREGULATED IN <b>LIVER CANCER</b> [*612210] (MIM gene record with a description of HIGHLY UPREGULATED IN <b>LIVER CANCER</b> ; HULC) is associated with Gene ENSG00000251164	0.38804042
<a href="#">ENSG00000226023</a>	CT47A6	X:120943561-120961487:-1	Human	Gene	<input type="checkbox"/>	Cancer/testis antigen family 47, member A6 [Source:HGNC Symbol;Acc:HGNC:33287] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000226023	0.35215402
<a href="#">ENSG00000237957</a>	CT47A5	X:120963026-120966348:-1	Human	Gene	<input checked="" type="checkbox"/>	Cancer/testis antigen family 47, member A5 [Source:HGNC Symbol;Acc:HGNC:33286] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000237957	0.35100457
<a href="#">ENSG00000242362</a>	CT47A2	X:120977606-120980928:-1	Human	Gene	<input checked="" type="checkbox"/>	Cancer/testis antigen family 47, member A2 [Source:HGNC Symbol;Acc:HGNC:33283] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000242362	0.35100457
<a href="#">ENSG00000228517</a>	CT47A7	X:120953288-120956600:-1	Human	Gene	<input checked="" type="checkbox"/>	Cancer/testis antigen family 47, member A7 [Source:HGNC Symbol;Acc:HGNC:33288] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000228517	0.35100457
<a href="#">ENSG00000230347</a>	CT47A8	X:120948422-120951744:-1	Human	Gene	<input checked="" type="checkbox"/>	Cancer/testis antigen family 47, member A8 [Source:HGNC Symbol;Acc:HGNC:33289] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000230347	0.35100457
<a href="#">ENSG00000230594</a>	CT47A4	X:120967886-120971208:-1	Human	Gene	<input checked="" type="checkbox"/>	Cancer/testis antigen family 47, member A4 [Source:HGNC Symbol;Acc:HGNC:33285] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000230594	0.35100457
<a href="#">ENSG00000236126</a>	CT47A3	X:120972746-120976068:-1	Human	Gene	<input checked="" type="checkbox"/>	Cancer/testis antigen family 47, member A3 [Source:HGNC Symbol;Acc:HGNC:33284] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000236126	0.35100457
<a href="#">ENSG00000226600</a>	CT47A9	X:120943561-120946883:-1	Human	Gene	<input checked="" type="checkbox"/>	Cancer/testis antigen family 47, member A9 [Source:HGNC Symbol;Acc:HGNC:33290] CT47A11 (EntrezGene record with a description of <b>cancer</b> /testis antigen family 47, member A11), with a synonym of CT47.11, is associated with Gene ENSG00000226600	0.35100457

<< < 1 2 3 4 5 6 7 8 ... 34 35 > >>

The first two entries in this page are ncRNA genes. Let's try the 2<sup>nd</sup> one

Score is calculated based on the query: how much the annotation description is similar to the searching keyword (liver cancer)

Now it's showing the Gene; there are also other tabs

Many things can be explored

Human (GRCh38) Location: 6:8,652,137-8,653,846 Gene: HULC Transcript: HULC-001

**Gene: HULC** ENSG00000251164 **This is ENSEMBL Gene ID**

**Description** hepatocellular carcinoma up-regulated long non-coding RNA [Source:HGNC Symbol;Acc:HGNC:34232]

**Synonyms** HCCAT1, LINC00078, NCRNA00078

**Location** [Chromosome 6: 8,652,137-8,653,846](#) forward strand.

**INSDC coordinates** chromosome:GRCh38:CM000668.2:8652137:8653846:1

**Transcripts** This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	Length	Protein	Biotype	CCDS	RefSeq	Flags
HULC-001	<a href="#">ENST00000503668</a>	556 bp	No protein product	LincRNA	-	<a href="#">NR_004605</a>	GENCODE basic

**Summary**

**Name** [HULC](#) (HGNC Symbol)

**RefSeq** Overlapping RefSeq Gene ID [728655](#) matches but different biotype of misc RNA

**Ensembl version** ENSG00000251164.1

**GRCh37 assembly** This gene maps to [8,652,370-8,654,079](#) in GRCh37 coordinates. Stable ID ENSG00000251164 not present in GRCh37.

**Gene type** Known lincRNA

**Prediction Method** Manual annotation (determined on a case-by-case basis) from the [Havana](#) project.

**Alternative genes** **This gene corresponds to the following database identifiers:**  
Havana gene: [OTTHUMG00000160417](#)

[Go to Region in Detail for more tracks and navigation options \(e.g. zooming\)](#)

- Gene-based displays
- Summary
- Splice variants (1)
- Transcript comparison
- Supporting evidence
- Sequence
  - Secondary Structure
- External references
- Regulation
- Expression
- Comparative Genomics
  - Genomic alignments
  - Gene tree (image)
  - Gene tree (text)
  - Gene tree (alignment)
  - Gene gain/loss tree
- Orthologues
- Paralogues
- Protein families
- Phenotype
- Genetic Variation
  - Variation table
  - Variation image
  - Structural variation
- External data
  - Personal annotation
  - ID History
  - Gene history

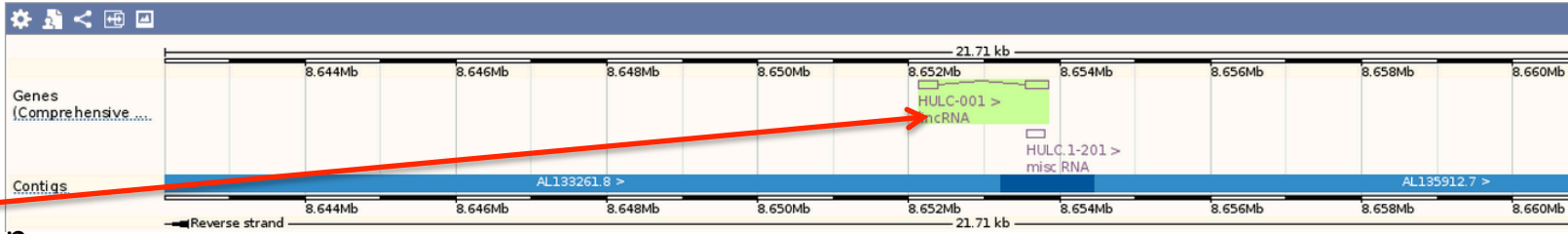
This is ENSEMBL Gene ID

Link to NCBI

This is ENSEMBL Transcript ID

This is a long intergenic non-coding RNA gene

Here is the graphical representation of the gene



Let's try a protein-coding gene: LAT1, also known as SLC7A5

Human (GRCh38) Location: 16:87,830,023-87,869,488 Gene: SLC7A5

 **Human**  
*Homo sapiens*

Search all categories ▾ SLC7A5

e.g. [BRCA2](#) or [17:63973115-64437414](#) or [osteoarthritis](#)

**Genome assembly: GRCh38 (GCA\_00001405.15)**

-  More information and statistics
-  Download DNA sequence (FASTA)
-  Convert your data to GRCh38 coordinates
-  Display your data in Ensembl

**Other assemblies**

GRCh37 (Long-term archive with BLAST, VEP and BioMart)

 View karyotype

 Example region



Click here



Human (GRCh38) ▾

Current selection:

< all Species

Only searching Human

Restrict category to:

Gene	4
Transcript	6
Variation	1390
Somatic Mutation	41
GeneTree	1
ProbeFeature	50
Protein Family	1

Per page:

10 25 50 100

Only searching Human ▾ SLC7A5



1493 results match SLC7A5 when restricted to species: Human ✕

### SLC7A5 (Human Gene)

[ENSG00000103257](#) 16:87830023-87869488:-1

Solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbol;Acc:HGNC:11063] **SLC7A5** (Vega gene) is associated with Gene ENSG00000103257  
[Variation table](#) • [Location](#) • [Regulation](#) • [Orthologues](#) • [Gene tree](#)

### SLC7A5-001 (Human Transcript)

[ENST00000261622](#) 16:87830023-87869488:-1

Solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbol;Acc:HGNC:11063] **SLC7A5-001** (Vega transcript) is associated with Transcript ENST00000261622

[Location](#) • [cDNA seq.](#) • [Variation table](#) • [Protein seq.](#) • [Population](#) • [Protein](#)

### SLC7A5-003 (Human Transcript)

[ENST00000563489](#) 16:87832732-87836805:-1

Solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbol;Acc:HGNC:11063] **SLC7A5-003** (Vega transcript) is associated with Transcript ENST00000563489

[Location](#) • [cDNA seq.](#) • [Variation table](#) • [Regulation](#)

Click to view the sequence page

Different names of the gene

Human (GRCh38) Location: 16:87,830,023-87,869,488 Gene: SLC7A5

### Gene-based displays

- Summary
- Splice variants (3)
- Transcript comparison
- Supporting evidence
- Sequence
- Secondary Structure
- External references
- Regulation
- Expression
- Comparative Genomics
- Genomic alignments
- Gene tree (image)
- Gene tree (text)
- Gene tree (alignment)
- Gene gain/loss tree
- Orthologues (68)
- Paralogues (7)
- Protein families (2)
- Phenotype
- Genetic Variation

## Gene: SLC7A5 ENSG00000103257

**Description** solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbol;Acc:HGNC:11063]

**Synonyms** CD98, D16S469E, E16, LAT1, MPE16

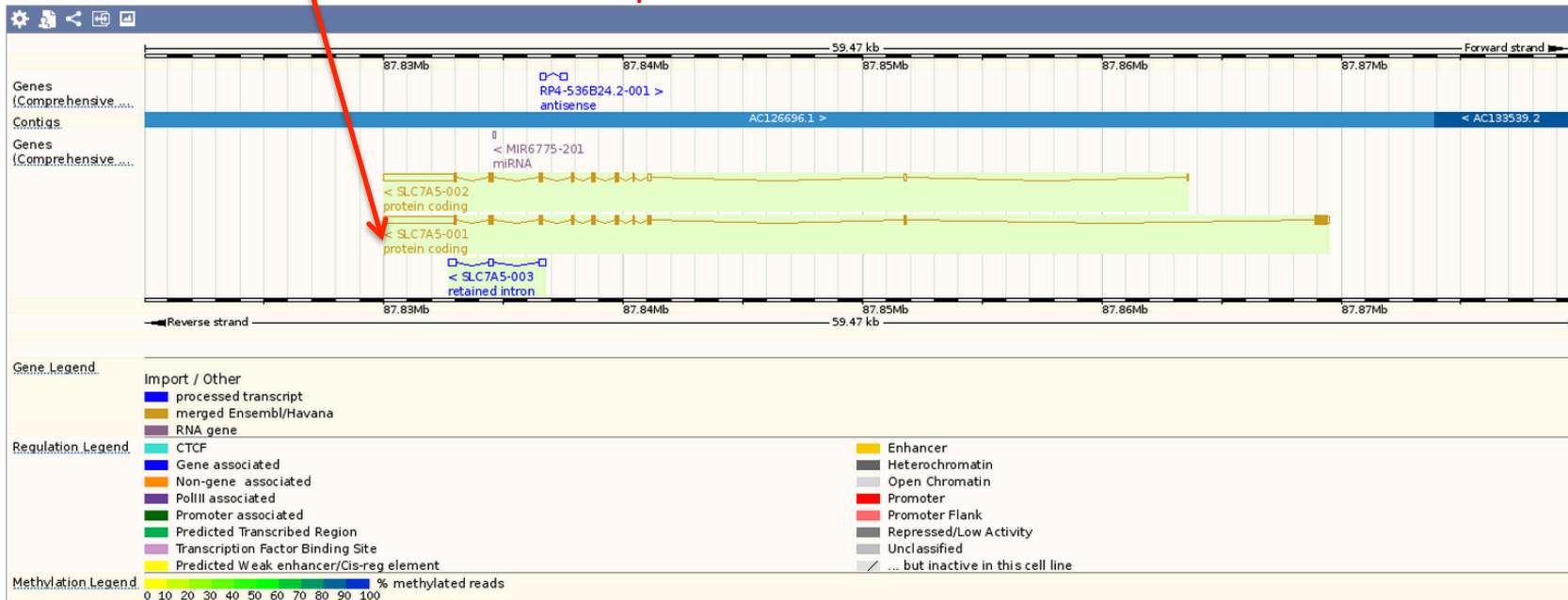
**Location** [Chromosome 16: 87,830,023-87,869,488](#) reverse strand.

**INSDC coordinates** chromosome:GRCh38:CM000678.2:87830023:87869488:1

**Transcripts** This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	Length	Protein	Biotype	CCDS	RefSeq	Flags
SLC7A5-001	<a href="#">ENST00000261622</a>	4537 bp	507 aa ( <a href="#">view</a> )	Protein coding	<a href="#">CCDS10964</a>	<a href="#">NM_003486</a> <a href="#">NP_003477</a>	Gencode basic
SLC7A5-003	<a href="#">ENST00000563489</a>	780 bp	No protein product	Retained intron	-	-	
SLC7A5-002	<a href="#">ENST00000565644</a>	3983 bp	241 aa ( <a href="#">view</a> )	Protein coding	-	-	Gencode basic

The three transcripts



Now check the expression

## Gene: SLC7A5 ENSG00000103257

**Description** solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbols]

**Synonyms** CD98, D16S469E, E16, LAT1, MPE16

**Location** [Chromosome 16: 87,830,023-87,869,488](#) reverse strand.

**INSDC coordinates** chromosome:GRCh38:CM000678.2:87830023:87869488:1

**Transcripts** This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Show/hide columns (1 hidden)						Filter	
Name	Transcript ID	Length	Protein	Biotype	CCDS	RefSeq	Flags
SLC7A5-001	<a href="#">ENST00000261622</a>	4537 bp	507 aa ( <a href="#">view</a> )	Protein coding	<a href="#">CCDS10964</a>	<a href="#">NM_003486</a> <a href="#">NP_003477</a>	Gencode basic
SLC7A5-003	<a href="#">ENST00000563489</a>	780 bp	No protein product	Retained intron	-	-	
SLC7A5-002	<a href="#">ENST00000565644</a>	3983 bp	241 aa ( <a href="#">view</a> )	Protein coding	-	-	Gencode basic

## Marked-up sequence ⓘ

[Download sequence](#) [BLAST this sequence](#)

### Key

Features [All exons in this region](#)

```
>chromosome:GRCh38:16:87829423:87870088:-1
GTTCTTCCCTCGTCCCAGTTCGCGGCTCACCAGCCCCACTGATGCAGCCCCCAGGCTGGA
AGGAGGCTGCAGGAGCTTCCCCTCAGGTCATCCTCTCATCCCTCCCCTGCCCCAGGAG
CTGGTTGTGGGGGCGGTTCATCCCTCGGCCATCCGGGACAGGAGCCTAGGTTCCCTT
CGGGGGTACCCAAATCCATCCTTGGCCTCAGCCAGCCCTGGTGCAGTCCCGCTCC
CAGGCTTGACGAGAGGCTGCGGGCCAGTGGGTGAAGGGGCGCCCTGACTGCCAGGCC
CGCCAGGCGCATCCGGGAGGACGGGCTGGGATGACGCGGGCCCGGGAGGGGGAGGTC
CGGAGGCCGGGGTCTCCATGGCGCAGGAGGACTGGGGCCTTCGAGGACCACGCGGGCCTG
GGAATAGCCCGCCAGGCTGGGCCGACGACGCACGTGCTCCGAGCTGGGCCAGGGGGCG
GGGCTGAGGGACGGGGCCGGGCCAGGGGCGGGGAGGAGCCGCGGACGGTGGGCGGGCC
GGCGGGCCGGGGCTAAAAGGCGGGCGGGCGGGGTTCTGACGCAGTGCAGGGCGGG
GCGGCGGCACACTGCTCGTGGGCGCGGCTCCCGGGTGTCCAGGCCCGGGCGGTGCG
CAGAGCATGGCGGGTGGGGCCGGAAGCGGCGCGGCTAGCGGCGCGGGCGGGCCAGGAG
AAGGAAGAGGCGGGGAGAAGATGCTGGCCGCCAAGAGCGCGGACGGCTCGGCGCCGGCA
GGCGAGGGCGAGGGCGTGACCCTGCAGCGGAACATCACGCTGCTCAACGGCGTGGCCATC
```

Click to open a help page to explain what these highlights mean

- Summary
- Splice variants (3)
- Transcript comparison
- Supporting evidence
- Sequence**
  - Secondary Structure
  - External references
  - Regulation
  - Expression
- Comparative Genomics
  - Genomic alignments
  - Gene tree (image)
    - Gene tree (text)
    - Gene tree (alignment)
    - Gene gain/loss tree
  - Orthologues (68)
  - Paralogues (7)
  - Protein families (2)
- Phenotype
- Genetic Variation
  - Variation table
  - Variation image
  - Structural variation
- External data
  - Personal annotation
- ID History
  - Gene history

- Configure this page
- Add your data
- Export data
- Bookmark this page
- Share this page

**Gene-based displays**

- Summary
- Splice variants (3)
- Transcript comparison
- Supporting evidence
- Sequence
  - Secondary Structure
- External references
- Regulation
- Expression**
- Comparative Genomics
  - Genomic alignments
  - Gene tree (image)
  - Gene tree (text)
  - Gene tree (alignment)
  - Gene gain/loss tree
- Orthologues (68)
- Paralogues (7)
- Protein families (2)
- Phenotype
- Genetic Variation
  - Variation table
  - Variation image
  - Structural variation
- External data
  - Personal annotation
- ID History
  - Gene history

Configure this page

Add your data

Export data

Bookmark this page

**Gene: SLC7A5** ENSG00000103257

**Description** solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbol;Acc:HGNC:11063]

**Synonyms** CD98, D16S469E, E16, LAT1, MPE16

**Location** [Chromosome 16: 87,830,023-87,869,488](#) reverse strand.

**INSDC coordinates** chromosome:GRCh38:CM000678.2:87830023:87869488:1

**Transcripts** This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	Length	Protein	Biotype	CCDS	RefSeq	Flags
SLC7A5-001	<a href="#">ENST00000261622</a>	4537 bp	507 aa ( <a href="#">view</a> )	Protein coding	<a href="#">CCDS10964</a>	<a href="#">NM_003486</a> <a href="#">NP_003477</a>	GENCODE basic
SLC7A5-003	<a href="#">ENST00000563489</a>	780 bp	No protein product	Retained intron	-	-	
SLC7A5-002	<a href="#">ENST00000565644</a>	3983 bp	241 aa ( <a href="#">view</a> )	Protein coding	-	-	GENCODE basic

**Expression**

Expression data is available for the following tissues:

Tissue	All data	RNASeq gene models	Intron-spanning reads	RNASeq align
Adipose	<a href="#">View in location</a>	Models built using Human adipose total RNA, lot 05060581, caucasian female, throat cancer, Illumina Human Bodymap 2.0 Data	Y	Y
Adrenal	<a href="#">View in location</a>	Models built using Human adrenal total RNA, lot 0812003, caucasian male, cerebral vascular accident, Illumina Human Bodymap 2.0 Data	Y	Y
Blood	<a href="#">View in location</a>	Models built using Human white blood cell, caucasian male, healthy, Illumina Human Bodymap 2.0 Data	Y	Y

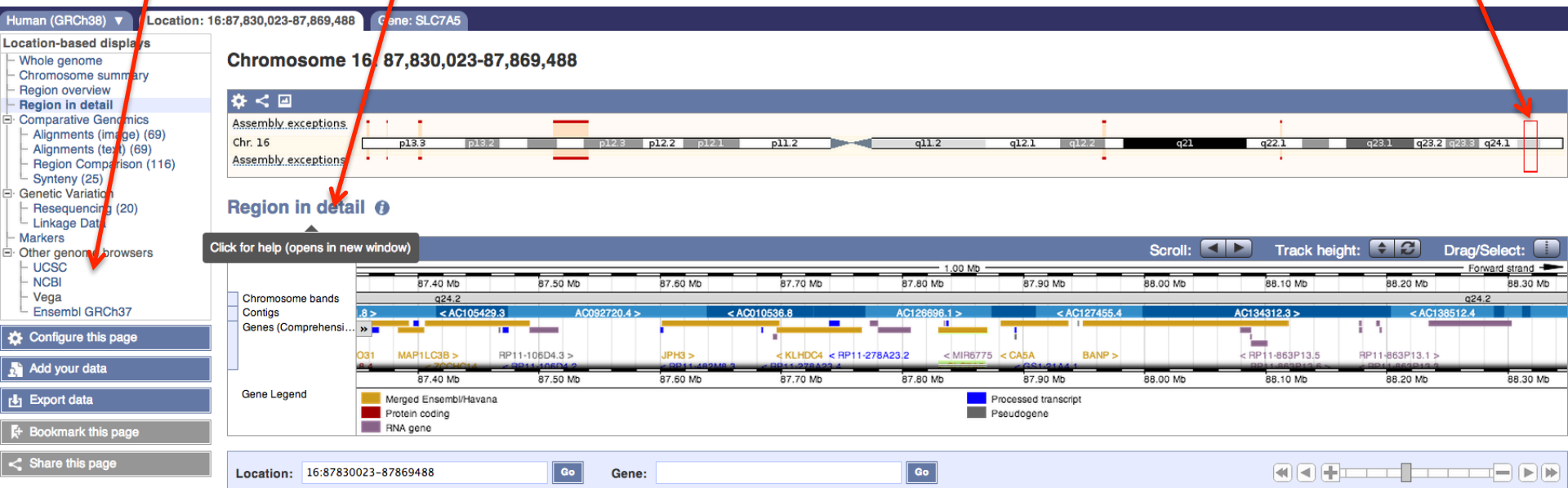


A long list, go further down to find liver and click “View in location”

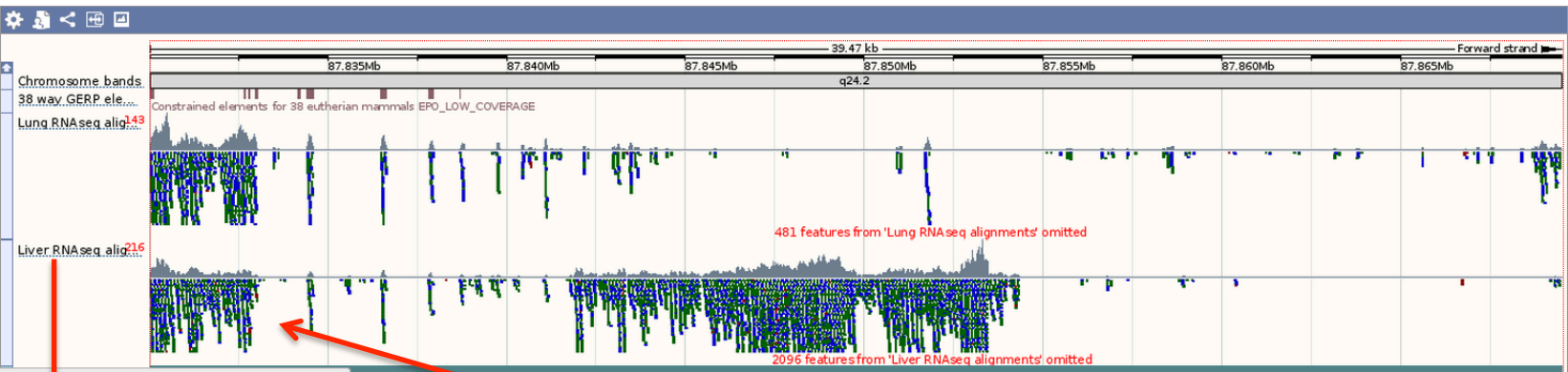
Links to other genome browsers

Zoomed in view

This is where the gene is located in the whole chromosome view



Further zoomed in view



A long page below

The RNA-seq read stack corresponding to exons

This is the same region in the UCSC browser

PS: much faster and easier to use/understand than ENSEMBL (richer info?)

# UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

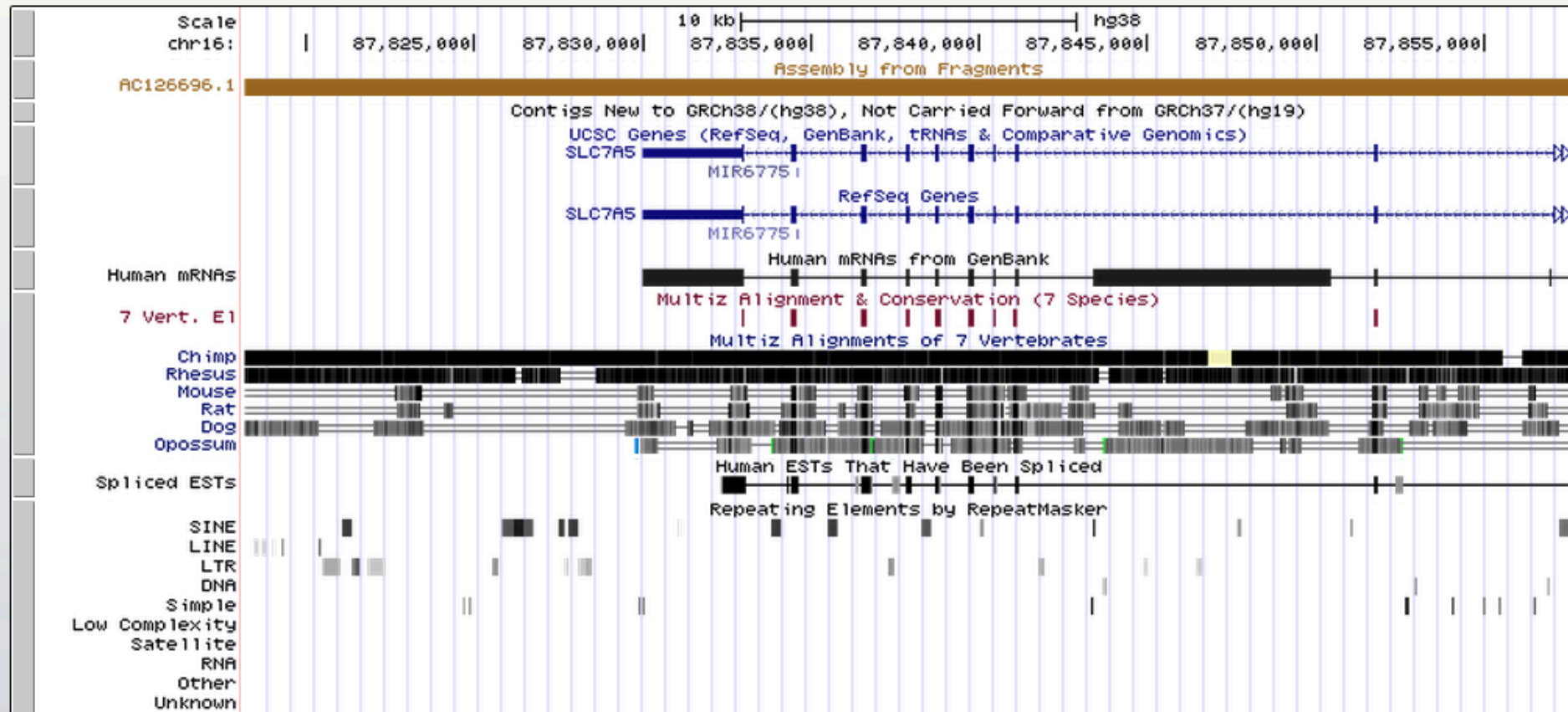
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr16:87,818,177-87,857,642 39,466 bp.

enter position, gene symbol or search terms

go

chr16 (q24.2) 16p13.3 12.3 12.1 p11.2 16q11.2 q12.1 16q21 22.1 q23.1



From the Gene tab click on Genome alignment will get you this page

Human (GRCh38) Location: 16:87,830,023-87,869,488 Gene: SLC7A5

### Gene: SLC7A5 ENSG00000103257

**Description** solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbol;Acc:HGNC:103257]

**Synonyms** CD98, D16S469E, E16, LAT1, MPE16

**Location** [Chromosome 16: 87,830,023-87,869,488](#) reverse strand.

**INSDC coordinates** chromosome:GRCh38:CM000678.2:87830023:87869488:1

**Transcripts** This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	Length	Protein	Biotype	CCDS	RefSeq	Flags
SLC7A5-001	<a href="#">ENST00000261622</a>	4537 bp	507 aa ( <a href="#">view</a> )	Protein coding	<a href="#">CCDS10964</a>	<a href="#">NM_003486</a> <a href="#">NP_003477</a>	GENCODE basic
SLC7A5-003	<a href="#">ENST00000563489</a>	780 bp	No protein product	Retained intron	-	-	
SLC7A5-002	<a href="#">ENST00000565644</a>	3983 bp	241 aa ( <a href="#">view</a> )	Protein coding	-	-	GENCODE basic

### Genomic alignments

Alignment:

**No alignment specified**

Please select the alignment you wish to display from the box above.

[Go to a graphical view of this alignment](#)

Key

Select 7 primates EPO and hit Go to see the whole genome alignment of 7 primates at this gene region

- Summary
- Splice variants (3)
- Transcript comparison
- Supporting evidence
- Sequence
  - Secondary Structure
- External references
- Regulation
- Expression
- Comparative Genomics
  - Genomic alignments**
  - Gene tree (image)
    - Gene tree (text)
    - Gene tree (alignment)
    - Gene gain/loss tree
  - Orthologues (68)
  - Paralogues (7)
  - Protein families (2)
- Phenotype
- Genetic Variation
  - Variation table
  - Variation image
  - Structural variation
- External data
  - Personal annotation
- ID History
  - Gene history

**Gene: SLC7A5 ENSG00000103257**

**Description** solute carrier family 7 (amino acid transporter light chain, L system), member 5 [Source:HGNC Symbols]

**Synonyms** CD98, D16S469E, E16, LAT1, MPE16

**Location** [Chromosome 16: 87,830,023-87,869,488](#) reverse strand.

**INSDC coordinates** chromosome:GRCh38:CM000678.2:87830023:87869488:1

**Transcripts** This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Show/hide columns (1 hidden)		Filter					
Name	Transcript ID	Length	Protein	Biotype	CCDS	RefSeq	Flags
SLC7A5-001	<a href="#">ENST00000261622</a>	4537 bp	507 aa ( <a href="#">view</a> )	Protein coding	<a href="#">CCDS10964</a>	<a href="#">NM_003486</a> <a href="#">NP_003477</a>	GENCODE basic
SLC7A5-003	<a href="#">ENST00000563489</a>	780 bp	No protein product	Retained intron	-	-	
SLC7A5-002	<a href="#">ENST00000565644</a>	3983 bp	241 aa ( <a href="#">view</a> )	Protein coding	-	-	GENCODE basic

**Genomic alignments** ⓘ

Alignment:

**Hidden and species** ⓘ

The following 1 species in the alignment are not shown in the image. Use the "Configure this page" on the left to show them.

- Ancestral sequence

The following 1 species have no alignment in this region:

- Marmoset (*Callithrix jacchus*)

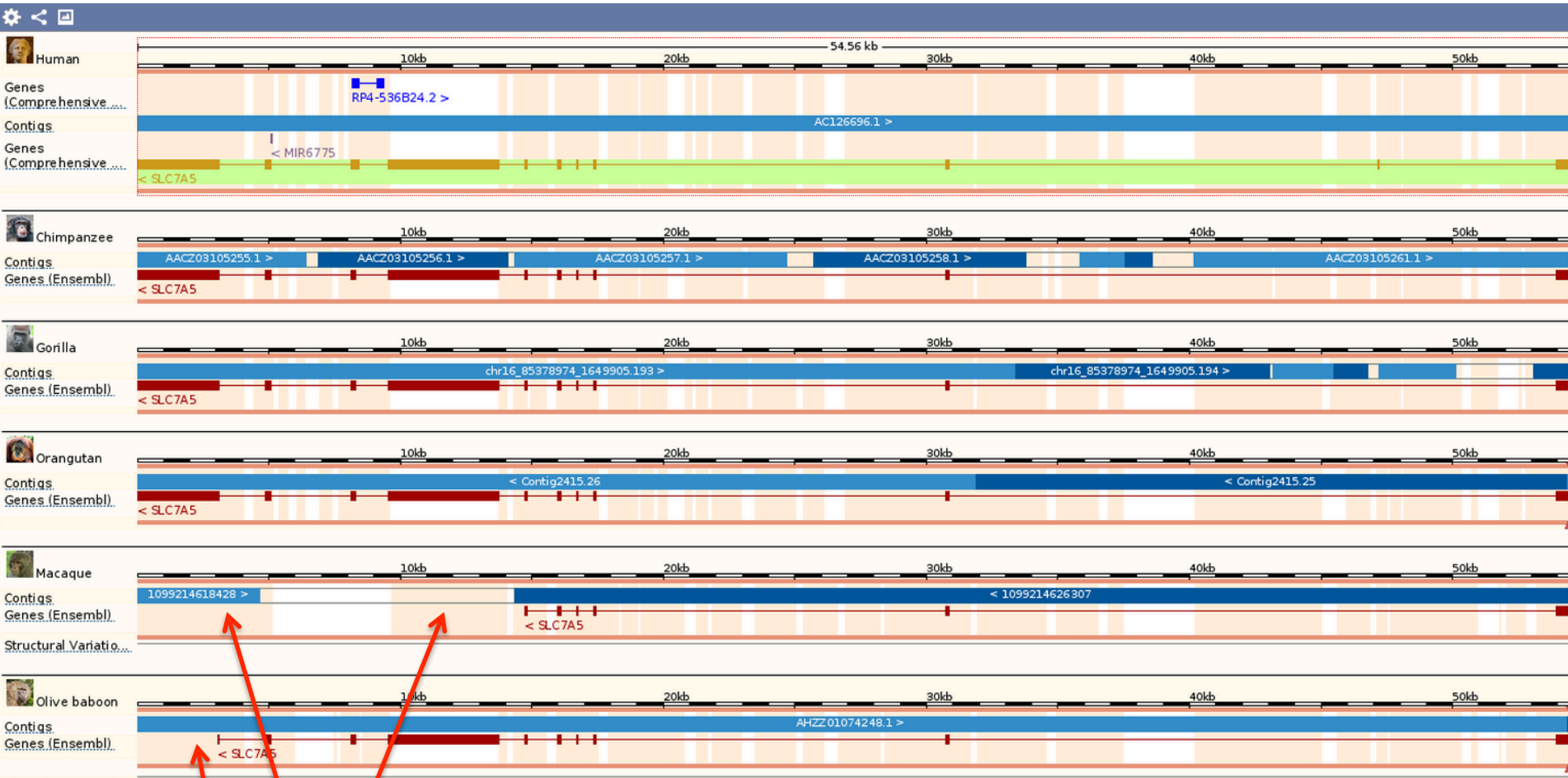
Hit here

[Go to a graphical view of this alignment](#) ←

A total of 2 alignment blocks have been found. Please select an alignment to view by selecting a Block from the Alignment column.



See how conserved this gene is across different primates



Some exons are missing in early primates

Search:  for

e.g. **Carboxy\*** or **chx28**

## Popular genomes



**Arabidopsis thaliana**

TAIR10



**Triticum aestivum**

IWGSP1



**Zea mays**

AGPv3



**Oryza sativa Japonica**

IRGSP-1.0



**Hordeum vulgare**

030312v2



**Physcomitrella patens**

ASM242v1

★ [Log in to customize this list](#)

## All genomes

-- Select a species --

[View full list of all Ensembl Plants species](#)

## What's New in Release 23

- New genomes
  - [Ostreococcus lucimarinus](#) (a green alga).
  - [Leersia perrieri](#) (a wild grass).
  - [Oryza rufipogon](#) (brownbeard rice).
  - [Theobroma cacao](#) (cocoa).
  - [Brassica oleracea](#) (the brassica C genome).
- Updated genomes
  - Updated assembly and gene build for [Oryza barthii](#) (wild rice).

### Did you know...?

The bread wheat A, B and D component genomes have been compared, allowing us to call orthology relationships between them, identifying the so called homoeologous genes. [Click here for example](#). Relationships between the component genomes can also now be browsed in our new region comparison view. [Click here for example](#).

## Variation data for bread wheat



Release 23 of Ensembl Plants includes three public variation datasets for bread wheat: the 80K iSelect array, and the KASP probe set. In total, ~724,000 SNPs have been added to the reference sequence performed at The Genome Analysis Centre (TGAC) infrastructure, allowing users search for SNPs by SNP ID (BA or BS nomenclature), find ne

The bread wheat genome in Ensembl Plants is the chromosome survey sequence for [Triticum aestivum](#) from the [Wheat Genome Sequencing Consortium](#). The gene models are provided by [MIPS \(version 4.0\)](#).

See also the wheat homepage at [ZURGI](#)

[Read more](#) about the [assembly](#), [annotation](#), [variation](#) and [analysis](#) of bread wheat

Ensembl Plants is developed in coordination with other plant genomics and bioinformatics transPLANT project is funded by the [European Commission](#) within its [7th Framework Programme](#) number [283496](#).



Wheat genomics resources are developed as part of our involvement in the consortium [Triti](#) [BBSRC](#), and led by [TGAC](#).



Databases are constructed in a direct collaboration with the [Gramene](#) resource, funded by the [US Department of Energy](#). More information about our collaboration with Gramene is available [here](#).



[Ensembl Genomes](#) is developed by [EMBL-EBI](#) and is powered by the [Ensembl](#) software system. Details of our funding please [click here](#).



# Next lecture: ExPASy and DTU tools