EBI web resources II: Ensembl and InterPro

Yanbin Yin

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 10/3 (send by email, if there are 2+ files, put them in a zip file; include your last name in the file name)

Outline

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

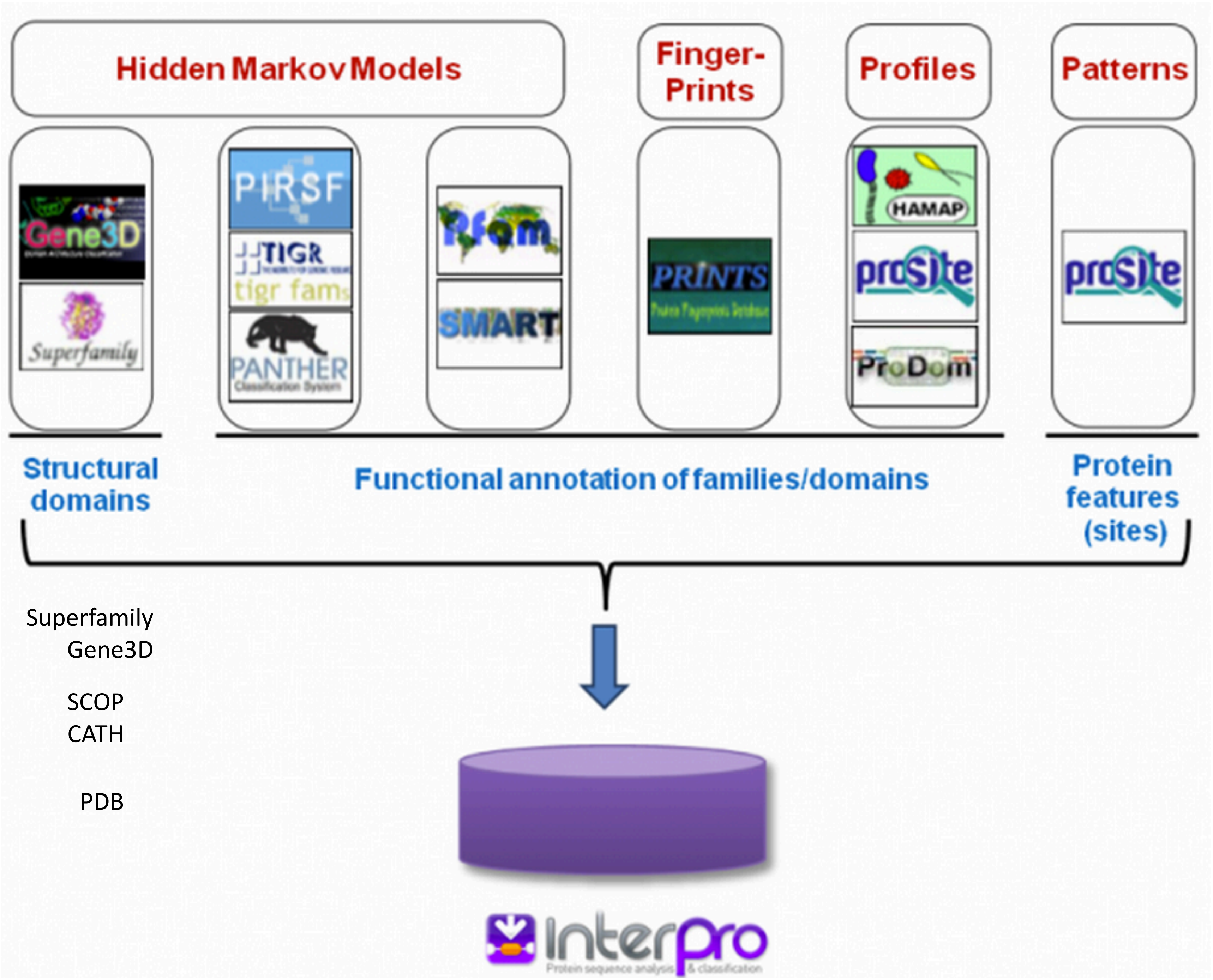
Genome annotation

- 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.



Superfamily
Gene3D

SCOP
CATH

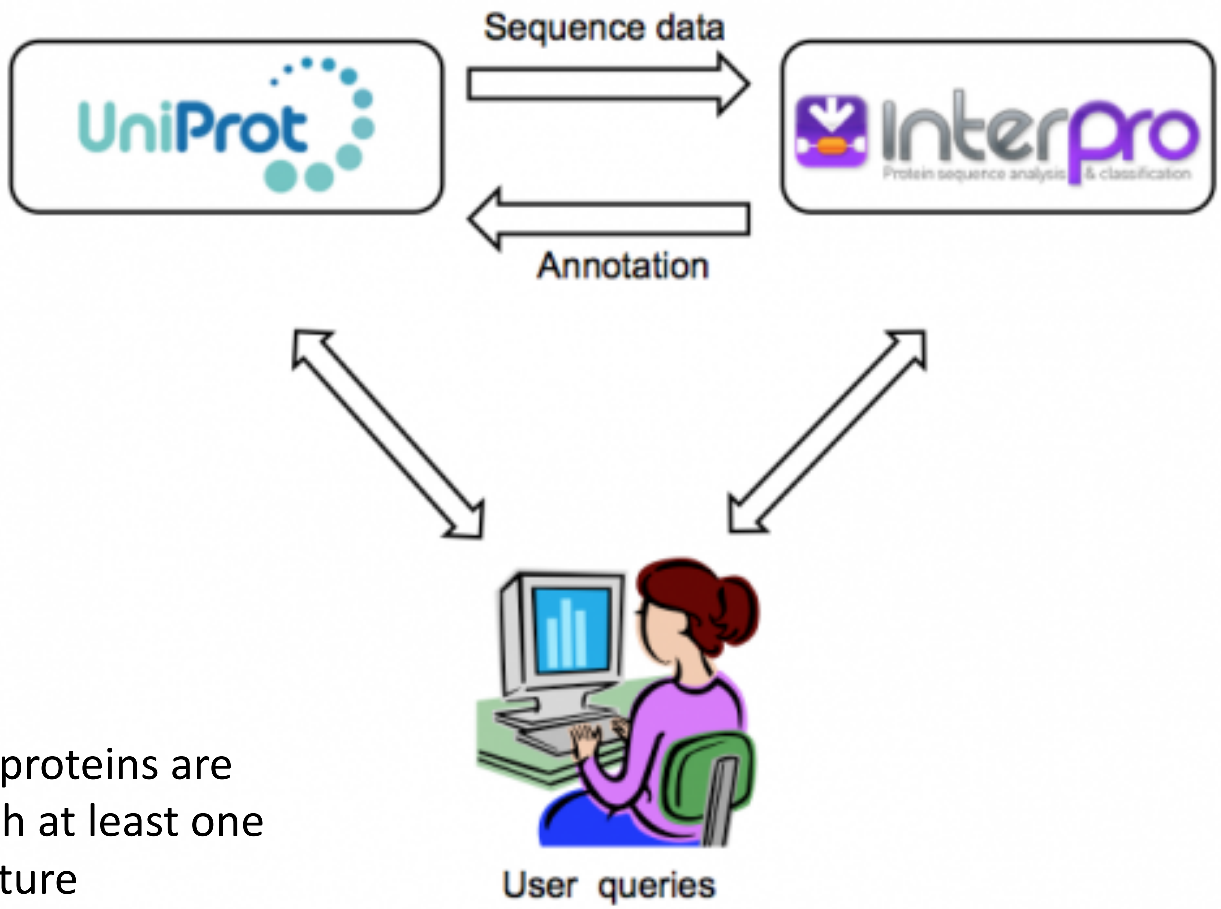
PDB

InterPro components

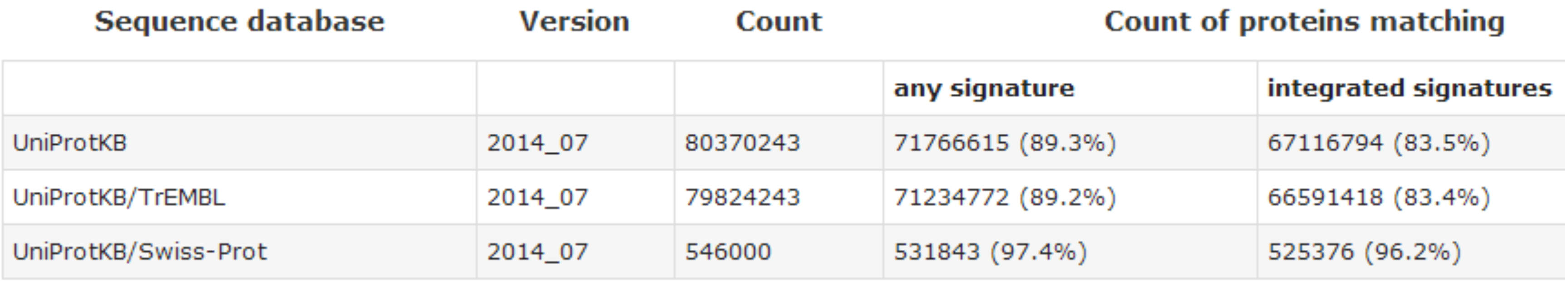
1.	CATH/Gene3D	University College, London, UK
2.	PANTHER	University of Southern California, CA, USA
3.	PIRSF	Protein Information Resource, Georgetown University, USA
✦4.	Pfam	Wellcome Trust Sanger Institute, Hinxton, UK
5.	PRINTS	University of Manchester, UK
6.	ProDom	PRABI Villeurbanne, France
7.	PROSITE	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

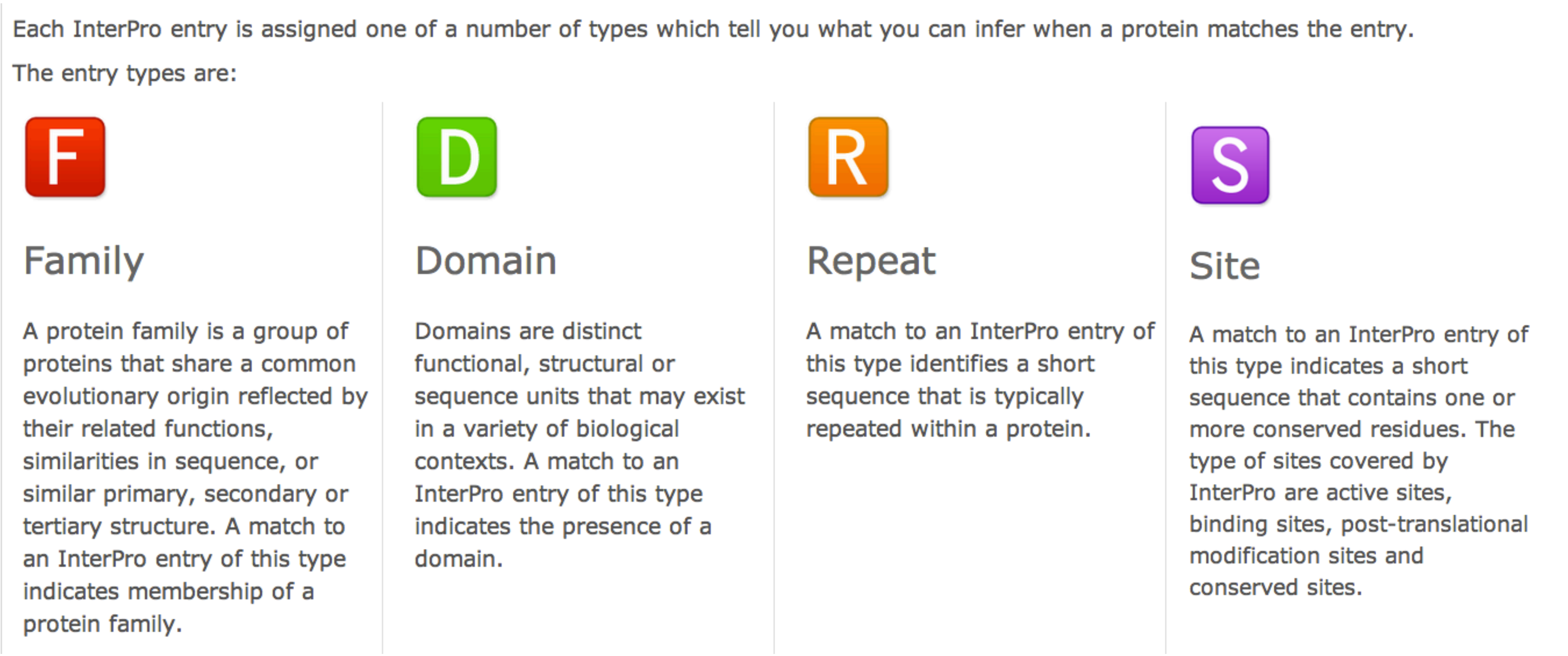
Pfam, SMART, TIGRFAM,
COG, KOG, PRK, CD, LOAD



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:

F

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.

D

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.

R

Repeat

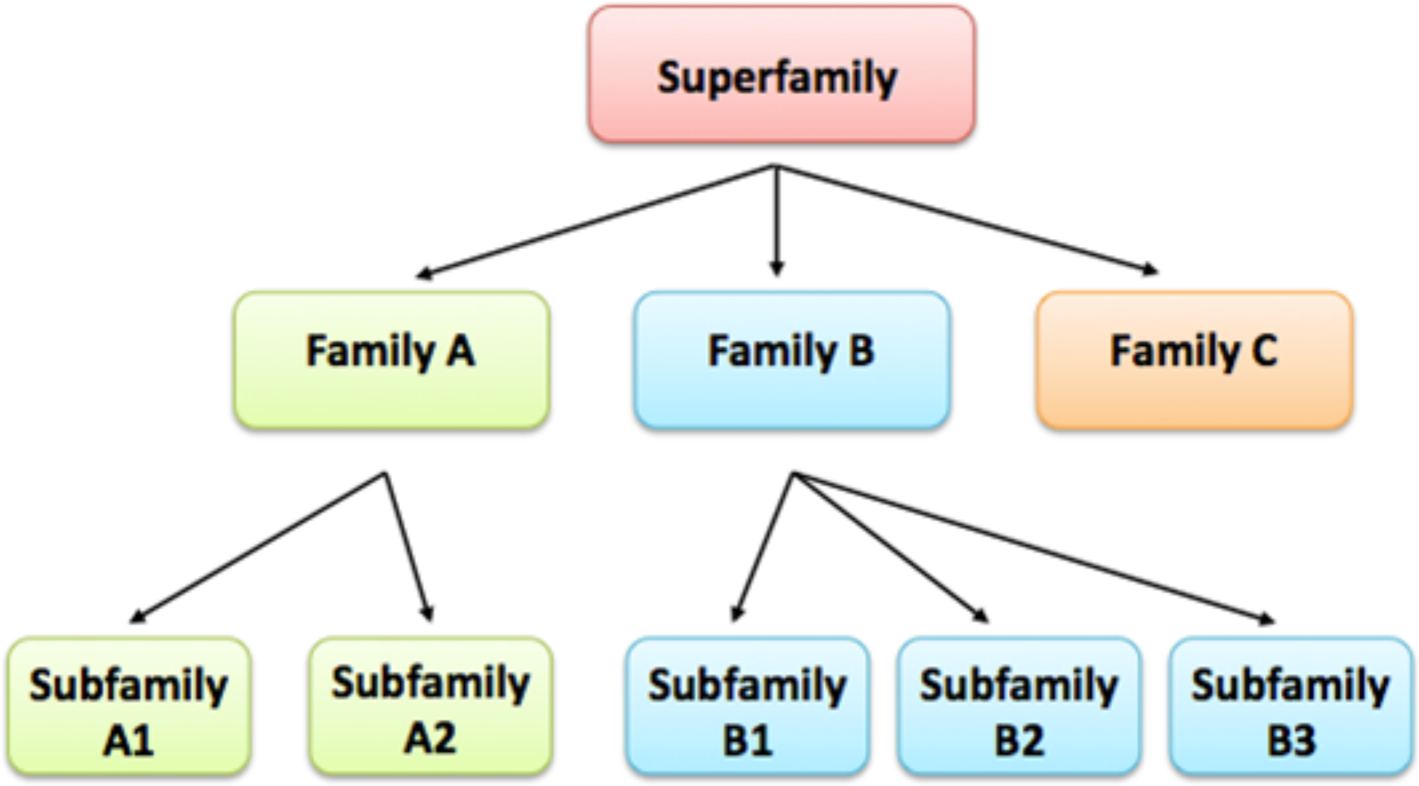
A match to an InterPro entry of this type identifies a short sequence that is typically repeated within a protein.

S

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.

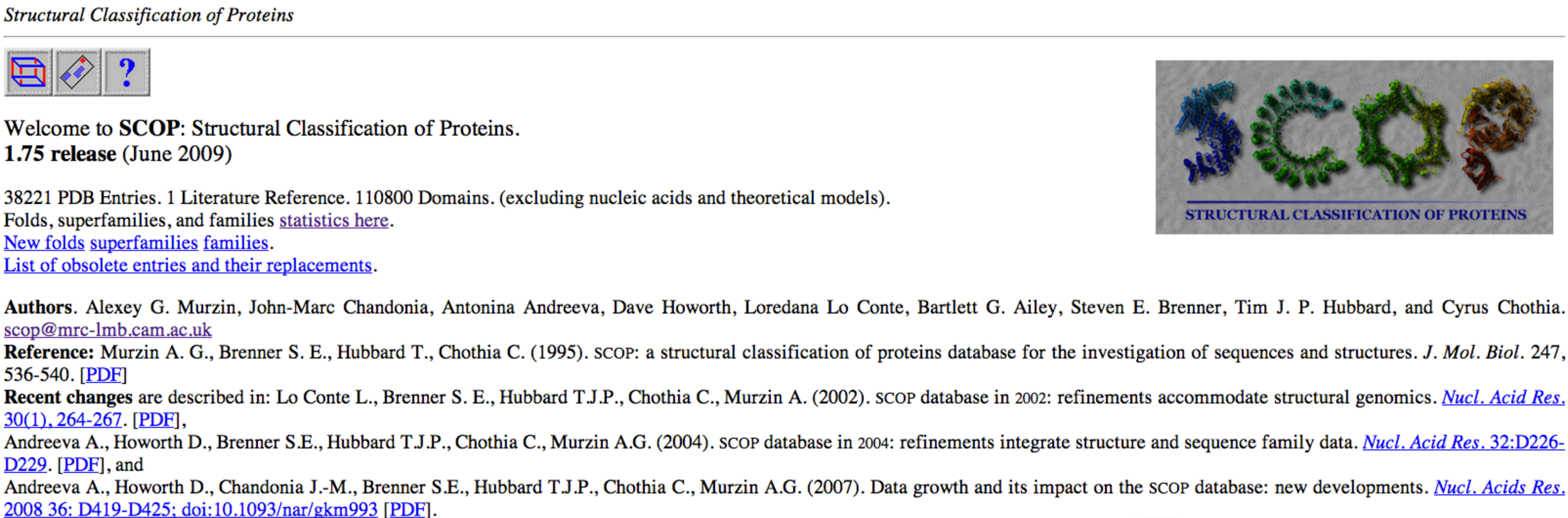
Superfamily: Probable common evolutionary origin

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



Structural Classification of Proteins

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.

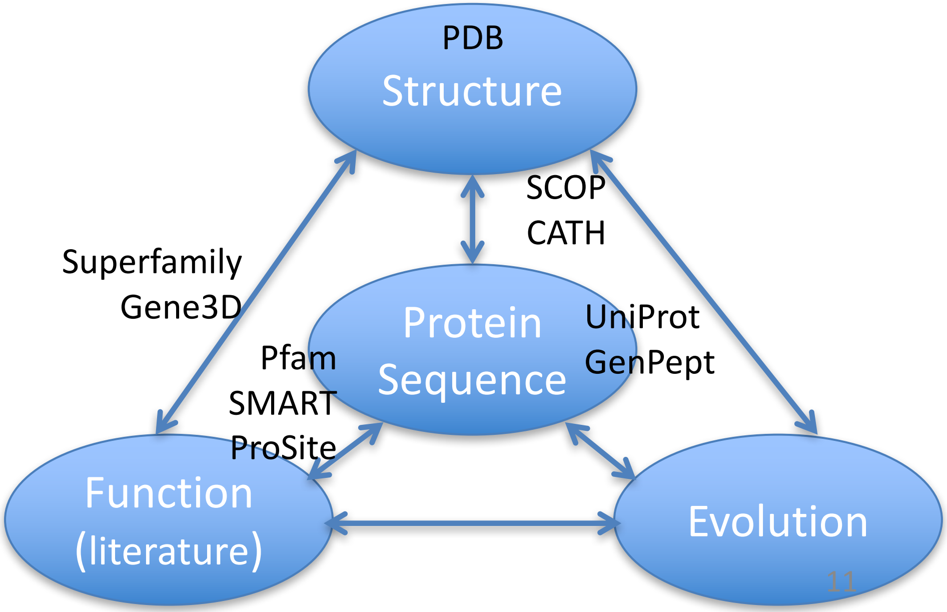
Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia.
scop@mrc-lmb.cam.ac.uk
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification 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

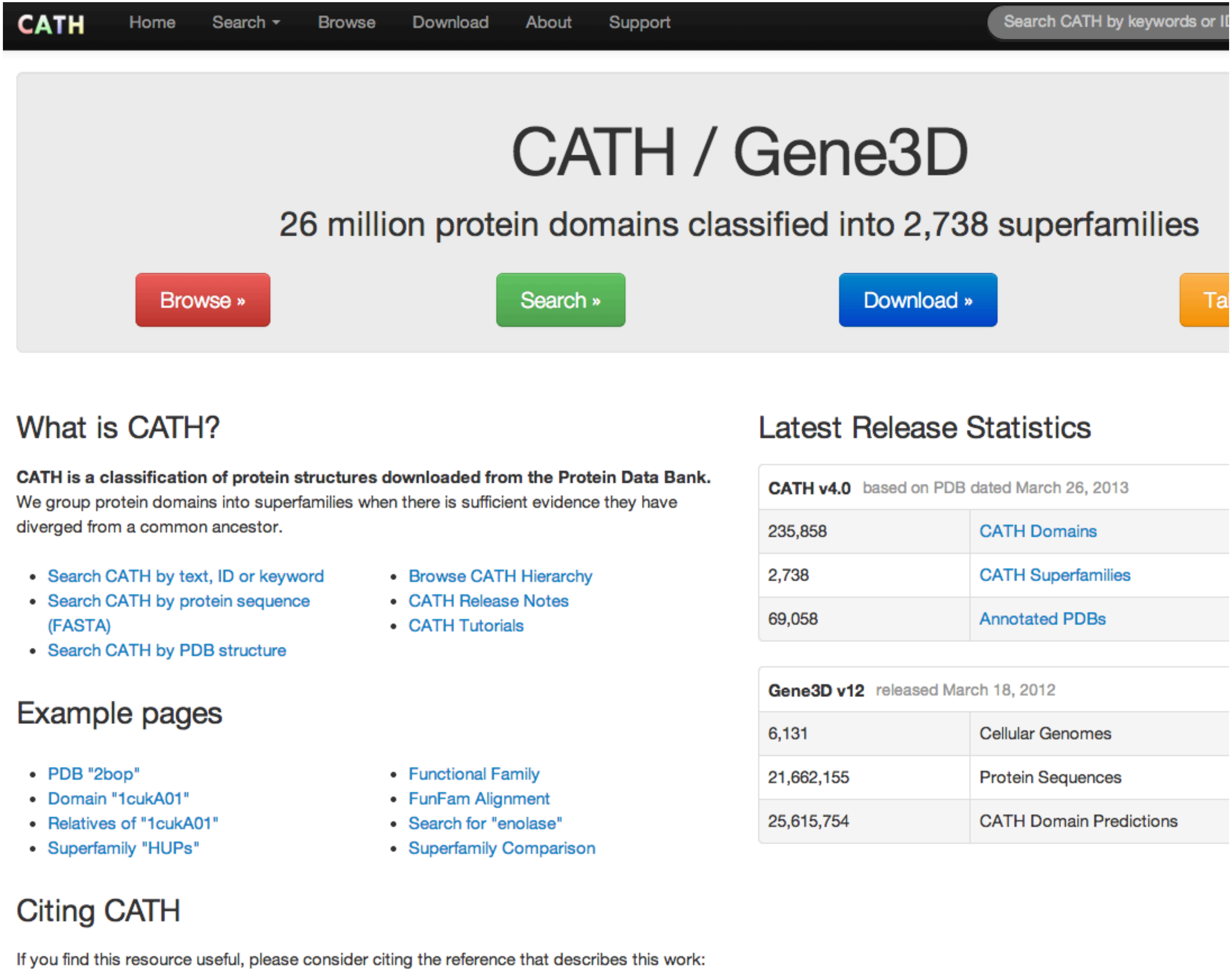
- Want to help us design and build the next generation of **SCOP** and **ASTRAL**?
  Get more details and apply here.

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)



http://www.cathdb.info/



CATH / Gene3D

26 million protein domains classified into 2,738 superfamilies

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

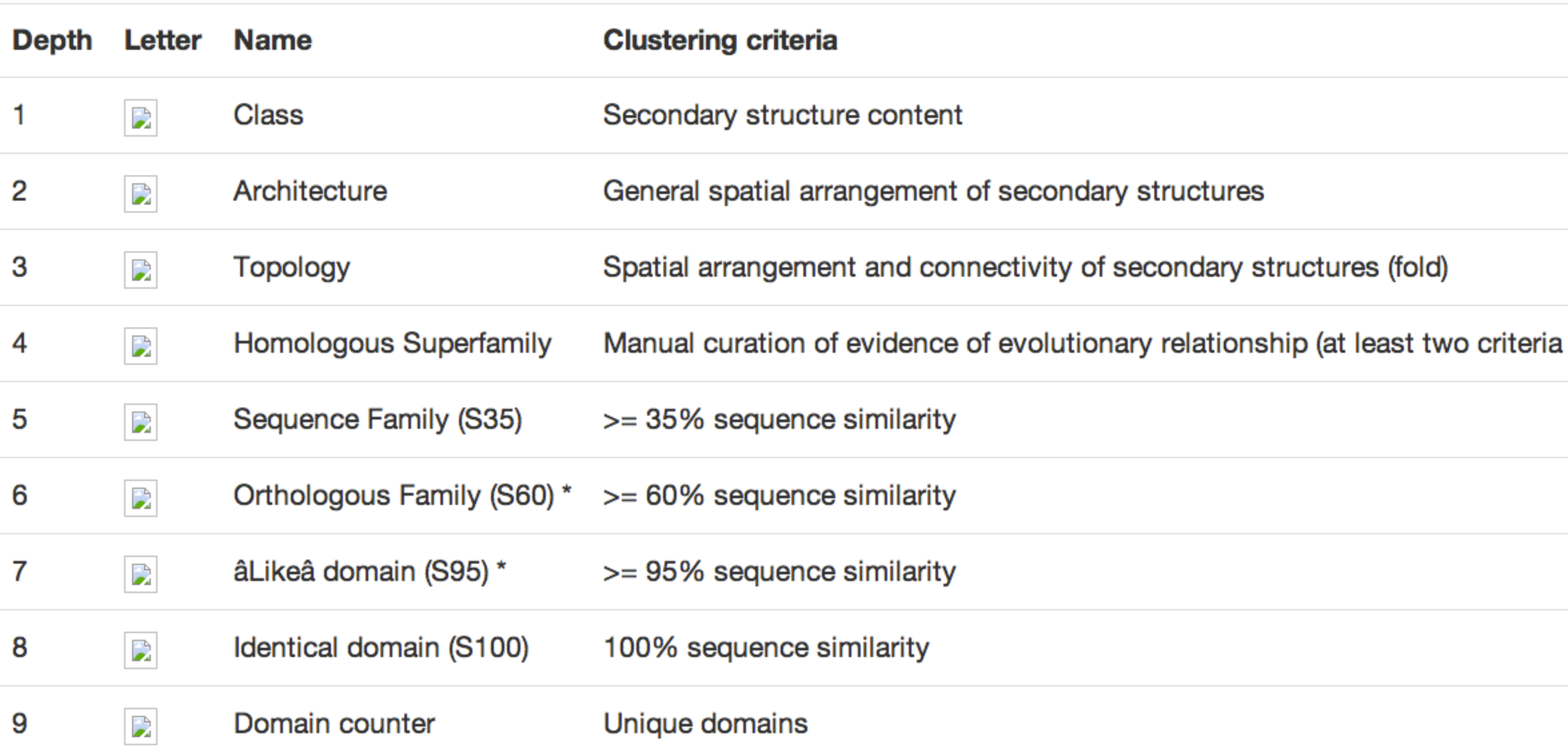
Citing CATH

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	CATH Domains
2,738	CATH Superfamilies
69,058	Annotated PDBs

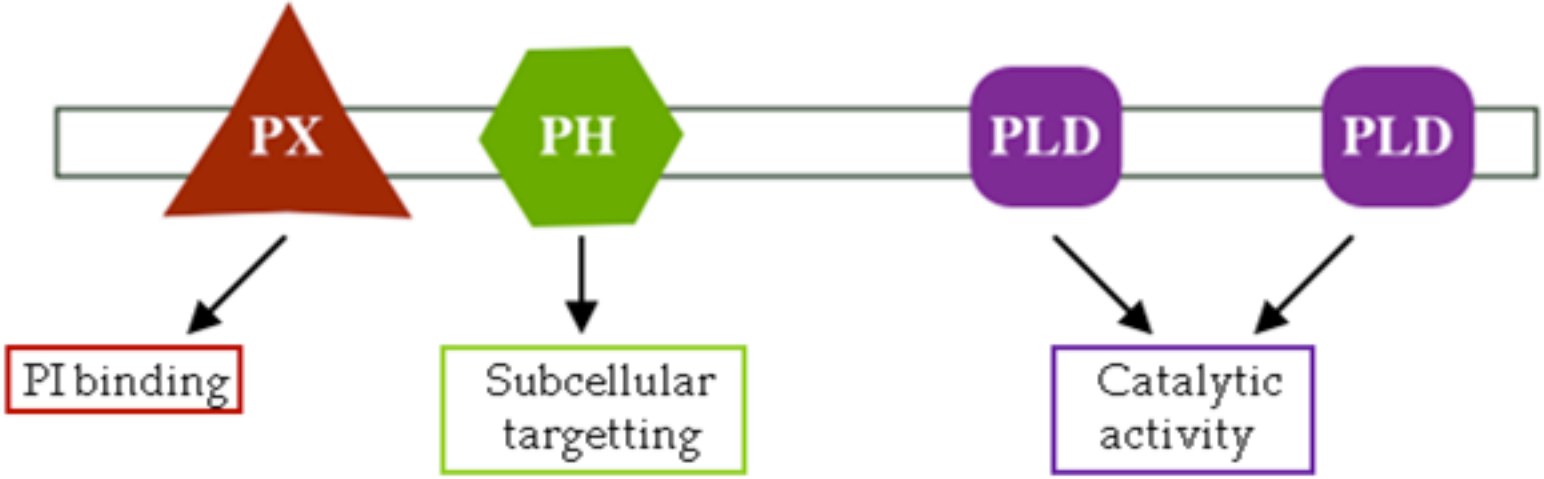
Gene3D v12	released March 18, 2012
6,131	Cellular Genomes
21,662,155	Protein Sequences
25,615,754	CATH Domain Predictions



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

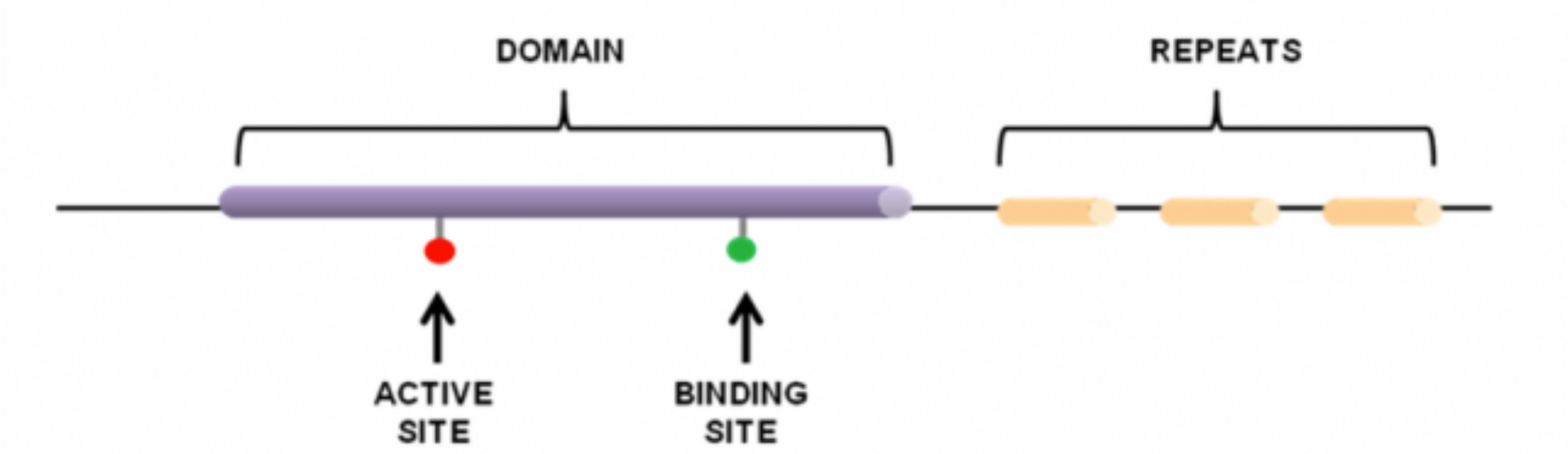
fold ~ class – superfamily ~ clan – family – subfamily – domain sequence

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



Sequences 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.

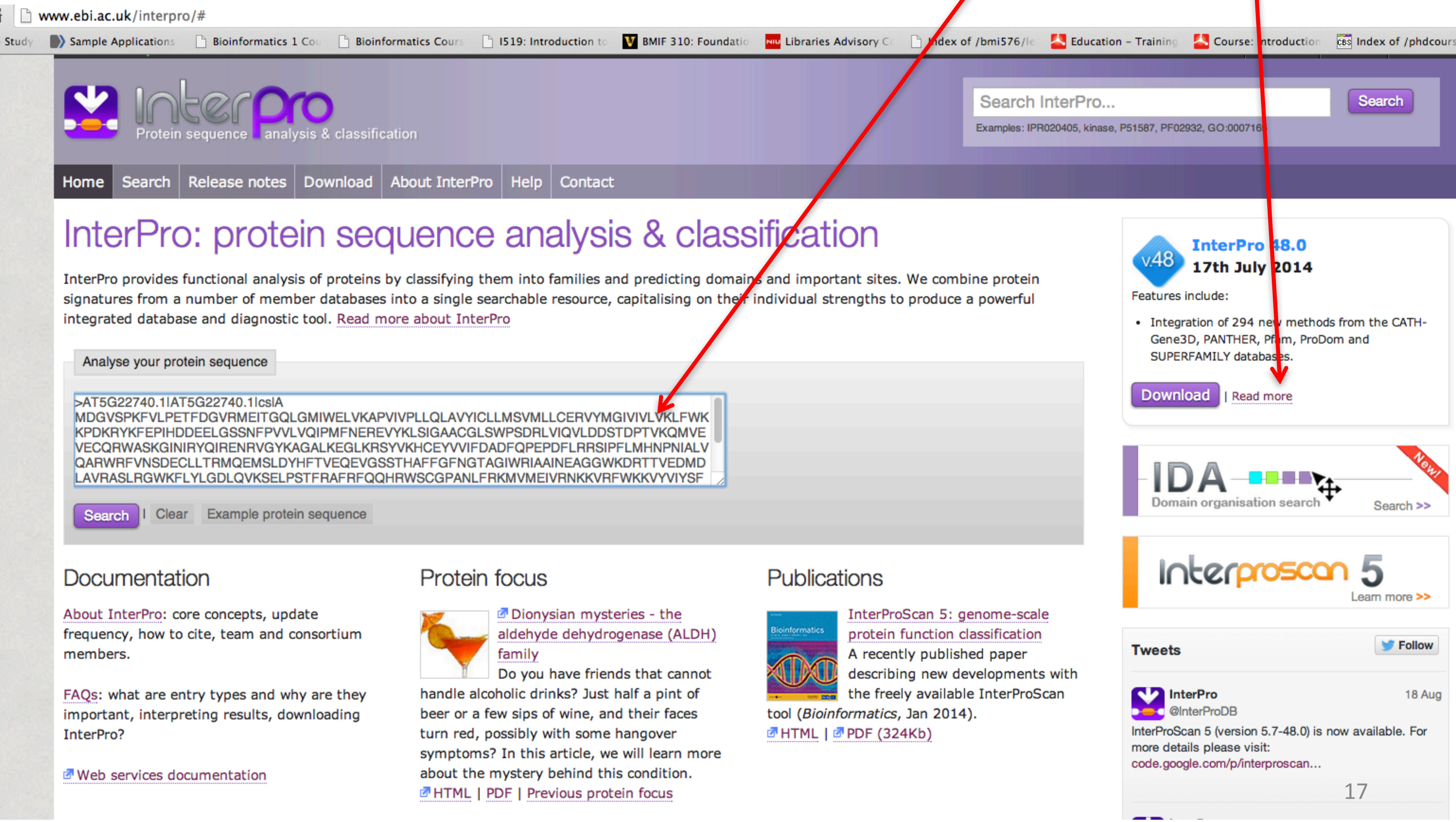
Hands on exercise 1: search against protein family databases

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



InterPro: protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. Read more about InterPro

Analyse your protein sequence

```
>AT5G22740.1|AT5G22740.1|cslA
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLVKLFWK
KPDKRYKFEPIHDDEELGSSNFPVVLVQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDSTDPTVKQMVE
VECQRWASKGINIRYQIRENRVGYKAGALKEGLKRSYVKHCEYVVIFDADFQPEPDFLRRSIPFLMHNPNIALV
QARWRFVNSDECLLTRMQEMSLDYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRTTVEDMD
LAVRASLRGWKFLYLGDLQVKSELPSTFRAFRFQQHRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSF
```

Search | Clear Example protein sequence

Documentation

About InterPro: core concepts, update frequency, how to cite, team and consortium members.

FAQs: what are entry types and why are they important, interpreting results, downloading InterPro?

Web services documentation

Protein focus

Dionysian mysteries - the aldehyde dehydrogenase (ALDH) family

Do you have friends that cannot handle alcoholic drinks? Just half a pint of beer or a few sips of wine, and their faces turn red, possibly with some hangover symptoms? In this article, we will learn more about the mystery behind this condition.

HTML | PDF | Previous protein focus

Publications

InterProScan 5: genome-scale protein function classification

A recently published paper describing new developments with the freely available InterProScan tool (*Bioinformatics*, Jan 2014).

HTML | PDF (324Kb)

v.48 InterPro 48.0
17th July 2014

Features include:

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

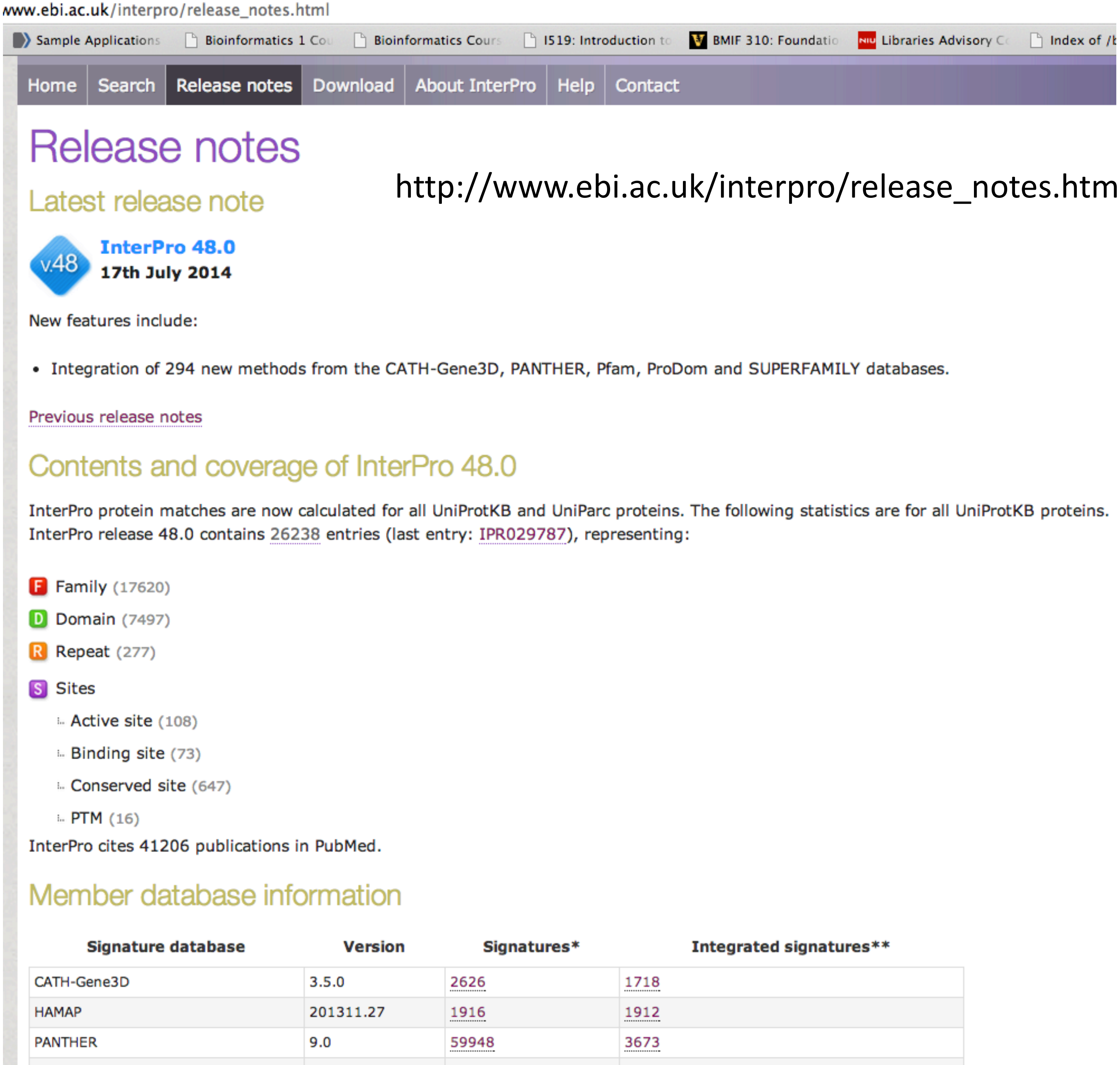
Download | Read more

Tweets Follow

InterPro @InterProDB 18 Aug

InterProScan 5 (version 5.7-48.0) is now available. For more details please visit: code.google.com/p/interproscan...

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



Release notes

Latest release note

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
  - Active site (108)
  - Binding site (73)
  - Conserved site (647)
  - PTM (16)

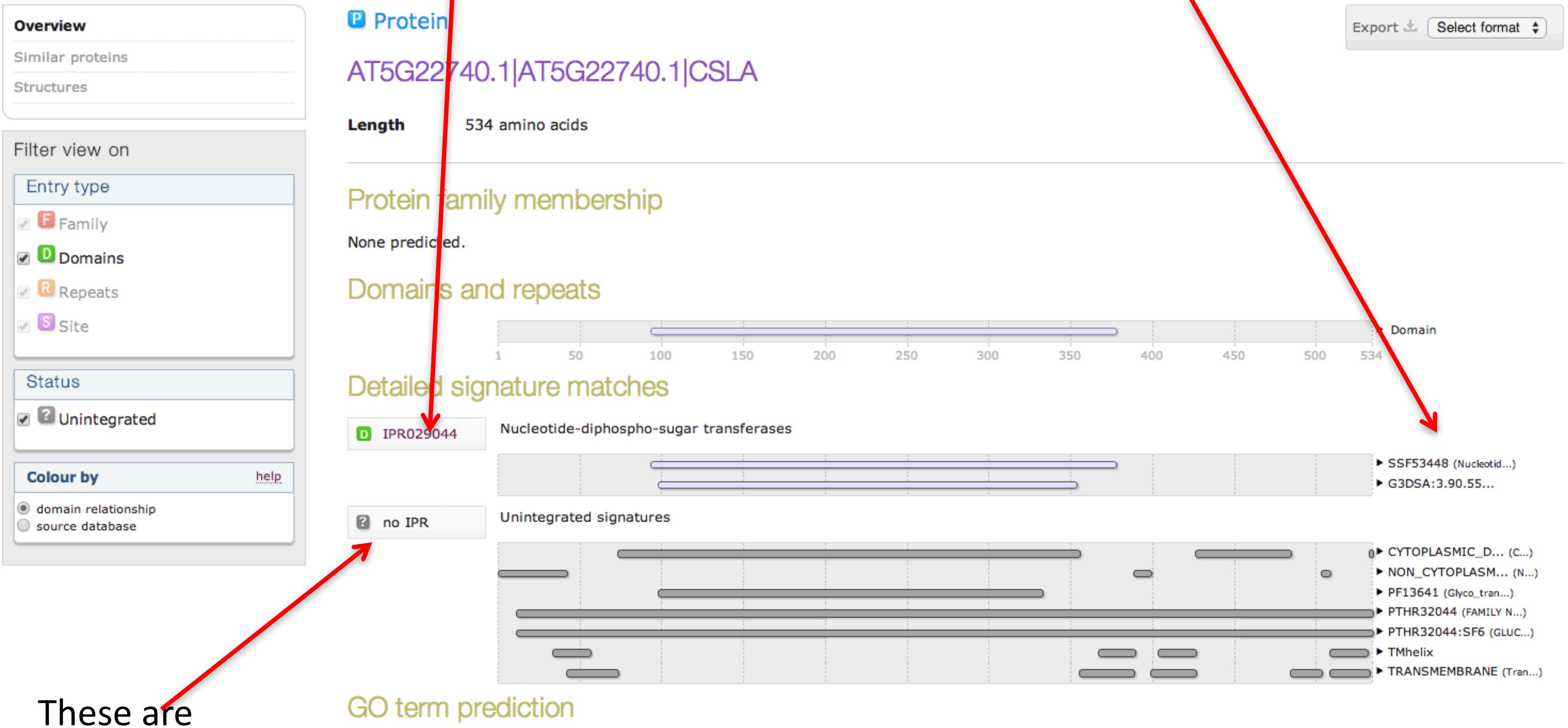
InterPro cites 41206 publications in PubMed.

Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	3.5.0	2626	1718
HAMAP	201311.27	1916	1912
PANTHER	9.0	59948	3673

Click to link to InterPro page of this domain

Click to link to individual database website



Overview
Similar proteins
Structures

Filter view on

Entry type
☑ F Family
☑ D Domains
☑ R Repeats
☑ S Site

Status
☑ ? Unintegrated

Colour by help
● domain relationship
○ source database

P Protein

AT5G22740.1|AT5G22740.1|CSLA

Export Select format

Length 534 amino acids

Protein family membership

None predicted.

Domains and repeats

Detailed signature matches

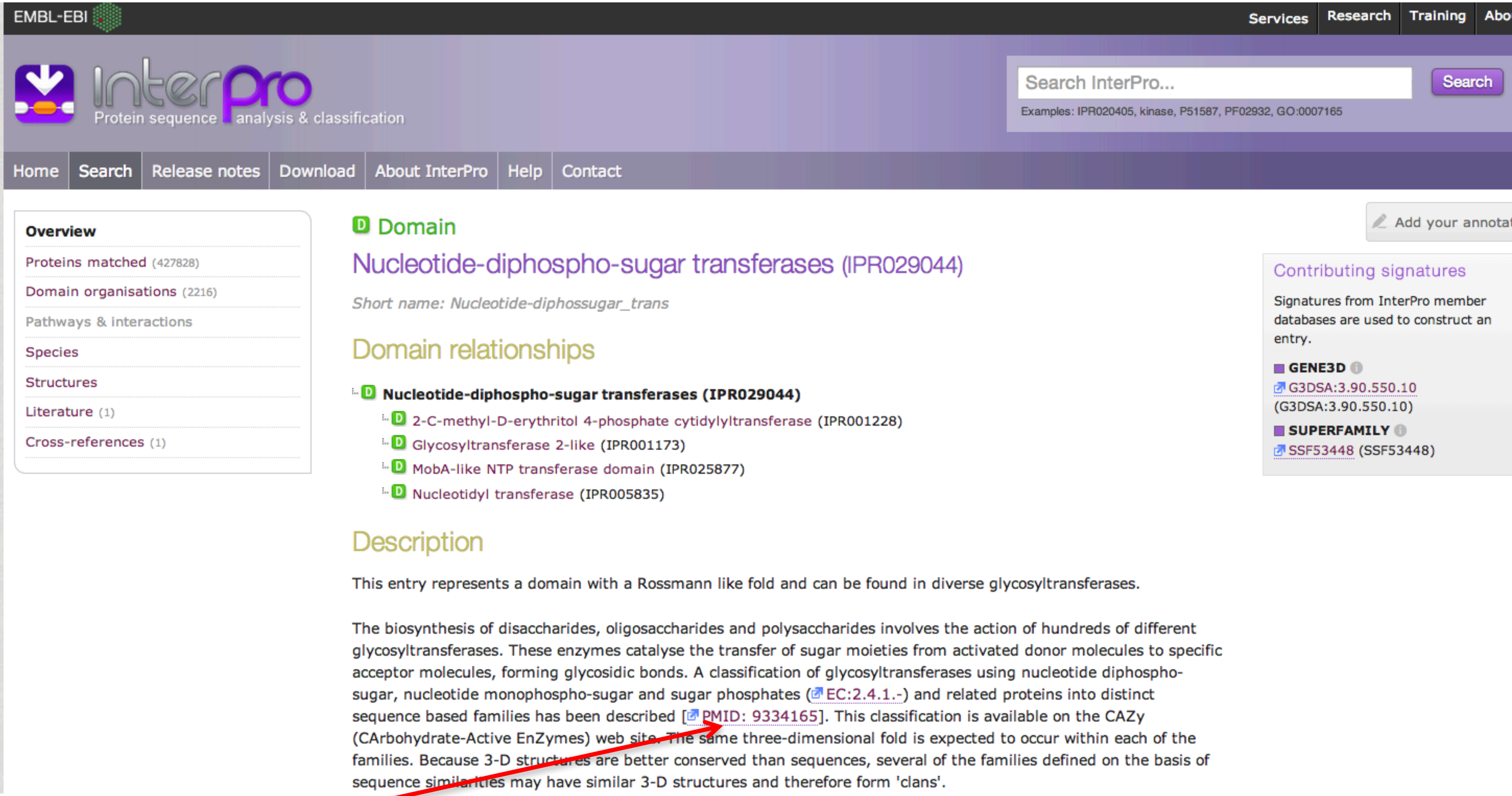
D IPR029044 Nucleotide-diphospho-sugar transferases

? no IPR Unintegrated signatures

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



D Domain

Nucleotide-diphospho-sugar transferases (IPR029044)

Short name: Nucleotide-diphossugar_trans

Domain relationships

- D **Nucleotide-diphospho-sugar transferases (IPR029044)**
  - D 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (IPR001228)
  - D Glycosyltransferase 2-like (IPR001173)
  - D MobA-like NTP transferase domain (IPR025877)
  - D 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'.

Contributing signatures

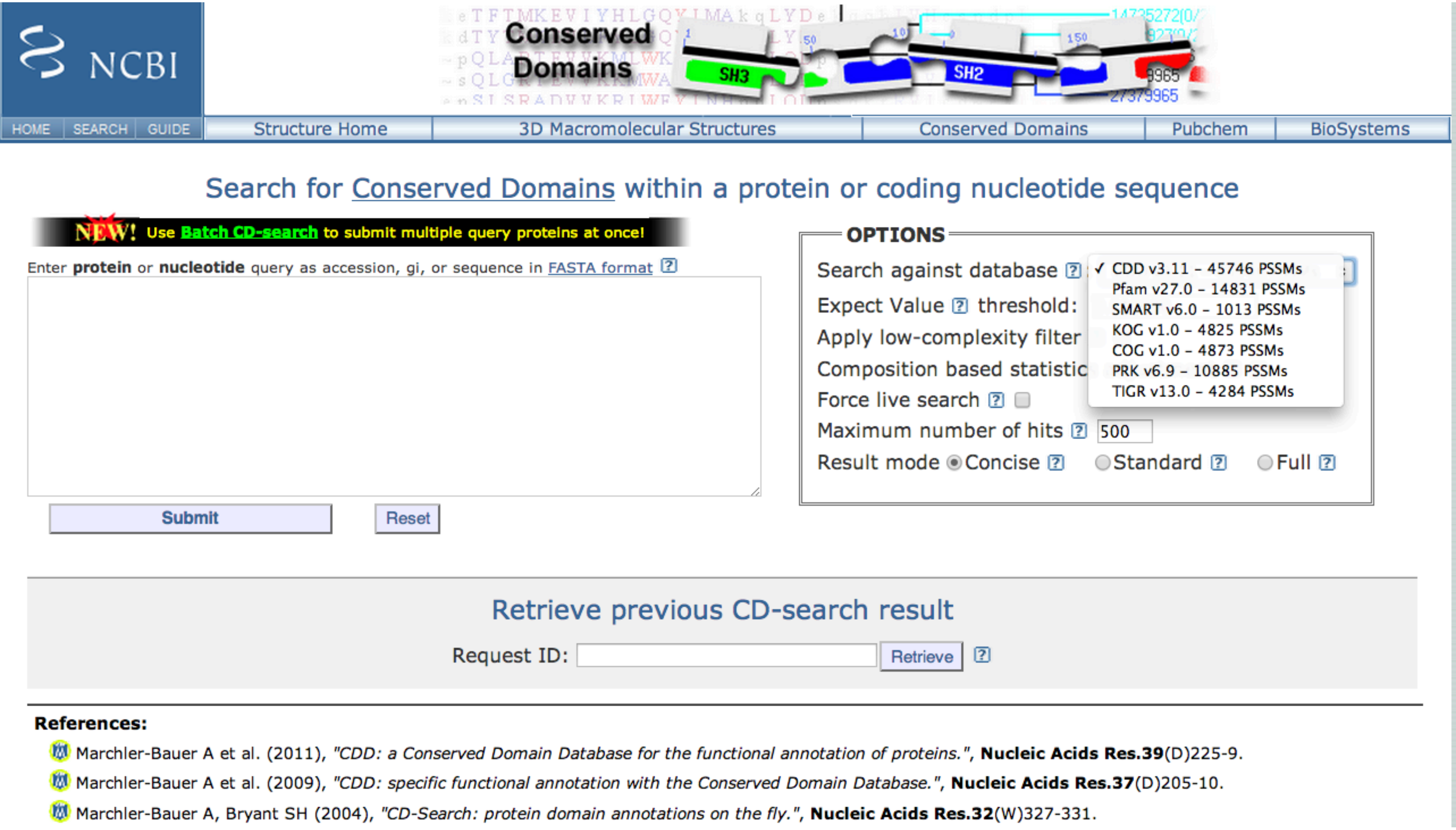
Signatures from InterPro member databases are used to construct an entry.

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



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 ? 500

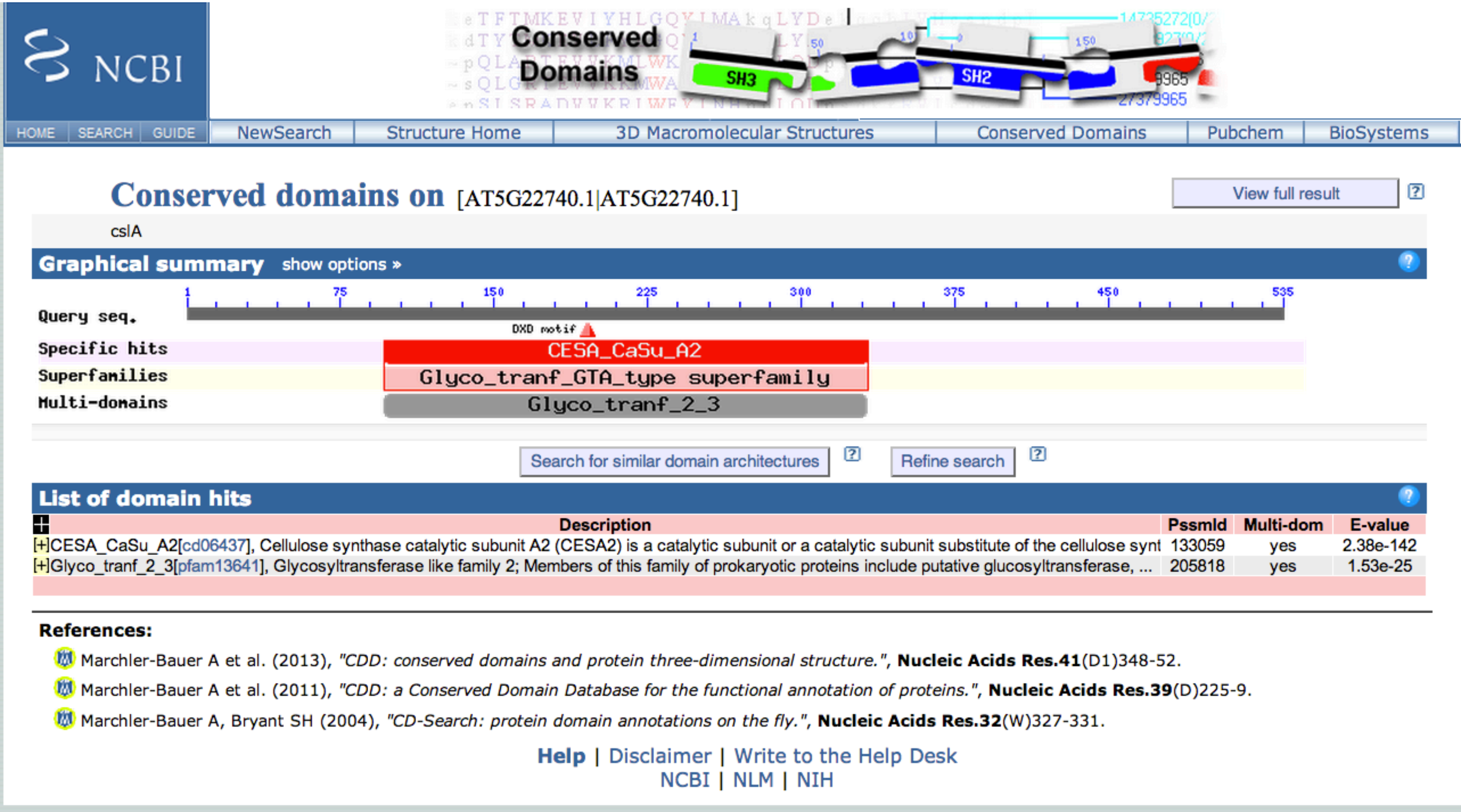
Result mode ● Concise ? ○ Standard ? ○ Full ?

Retrieve previous CD-search result

Request ID: Retrieve ?

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

cslA

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.

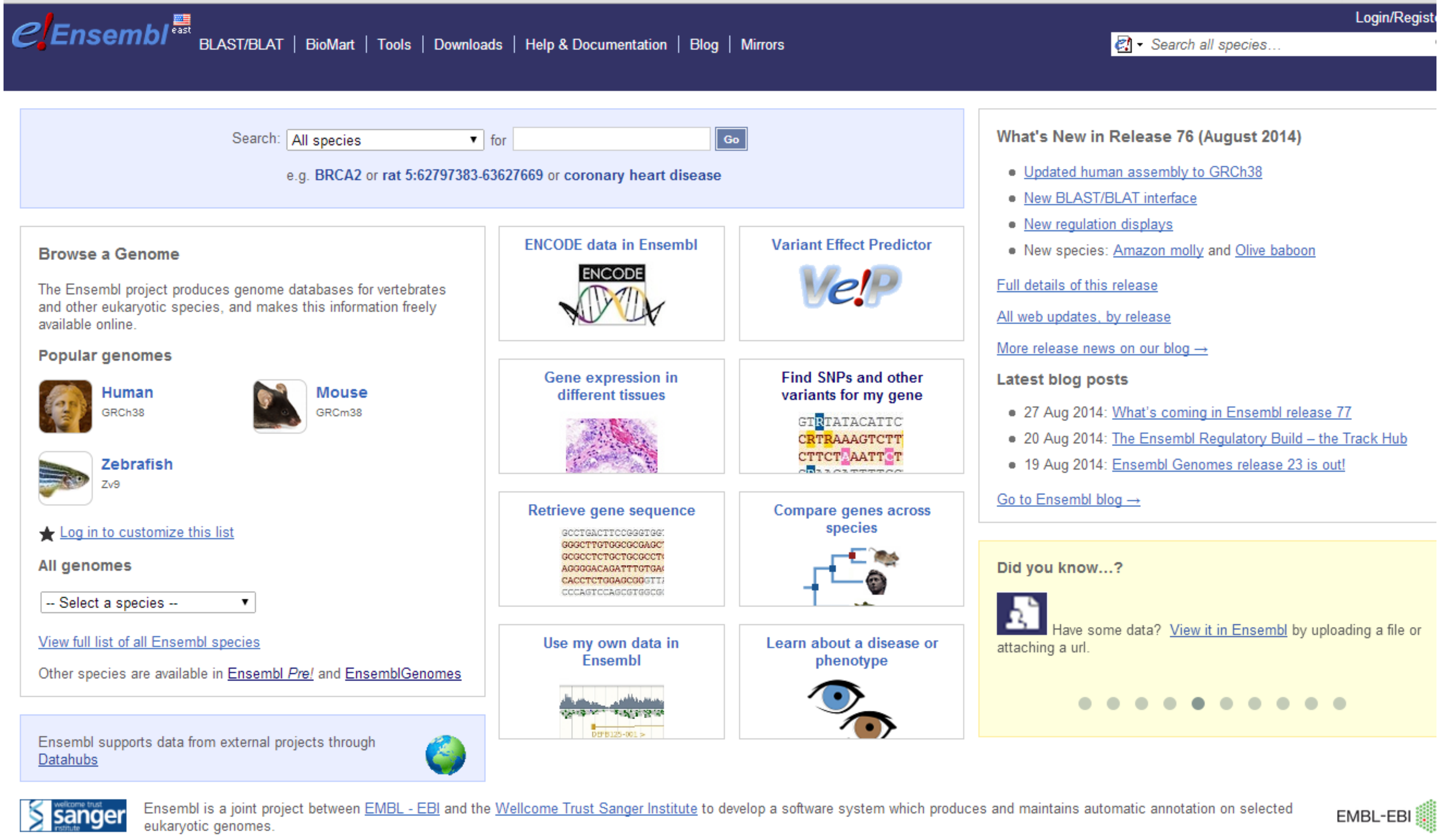
Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

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).



e!Ensembl east

BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login/Regist

Search all species...

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

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

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...?

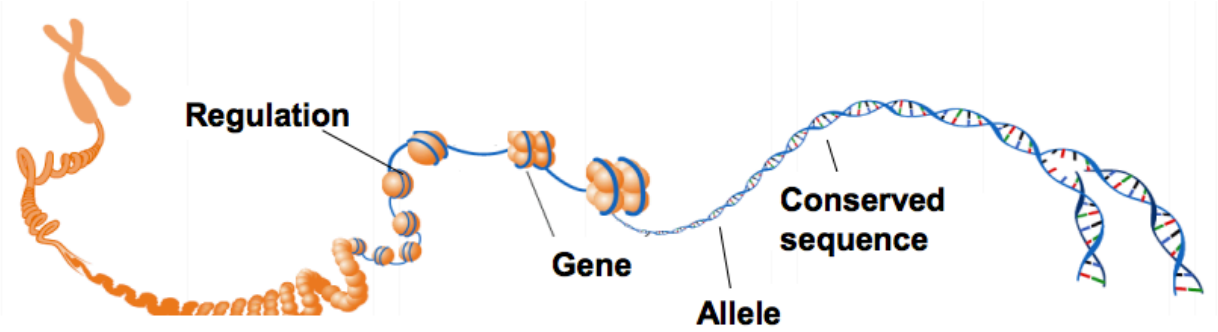
Have some data? View it in Ensembl by uploading a file or attaching a url.

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.

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.



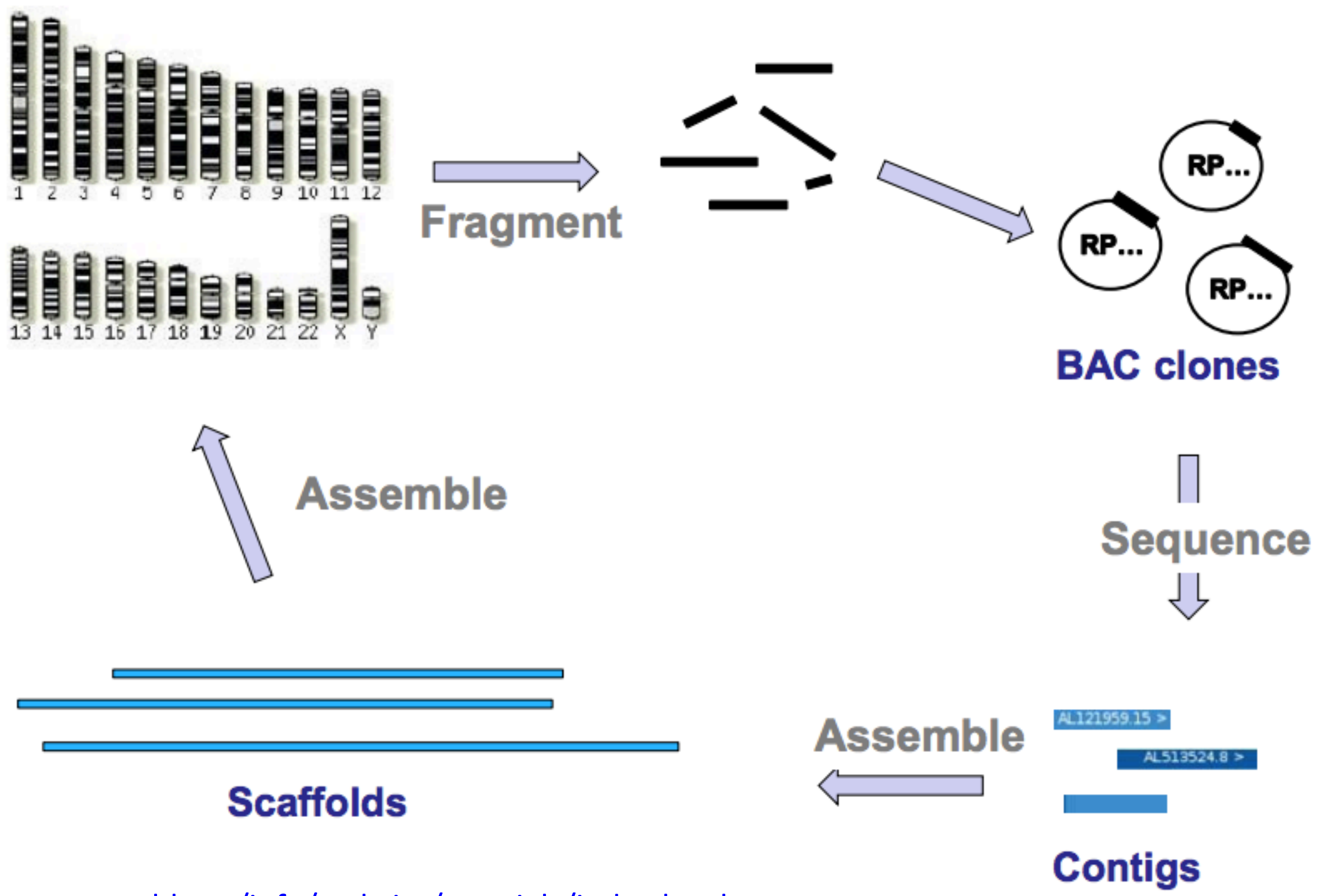
- Splice variants, proteins, non-coding RNA
- Small and large scale sequence variation, phenotype associations
- Whole genome alignments, protein trees
- Potential promoters and enhancers, DNA methylation
- User upload, custom data

Figure adapted from the ENCODE project www.nature.com/nature/focus/encode/

Short tutorial videos introducing ENSEMBL

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

Genome Sequencing



Fragment

BAC clones

Sequence

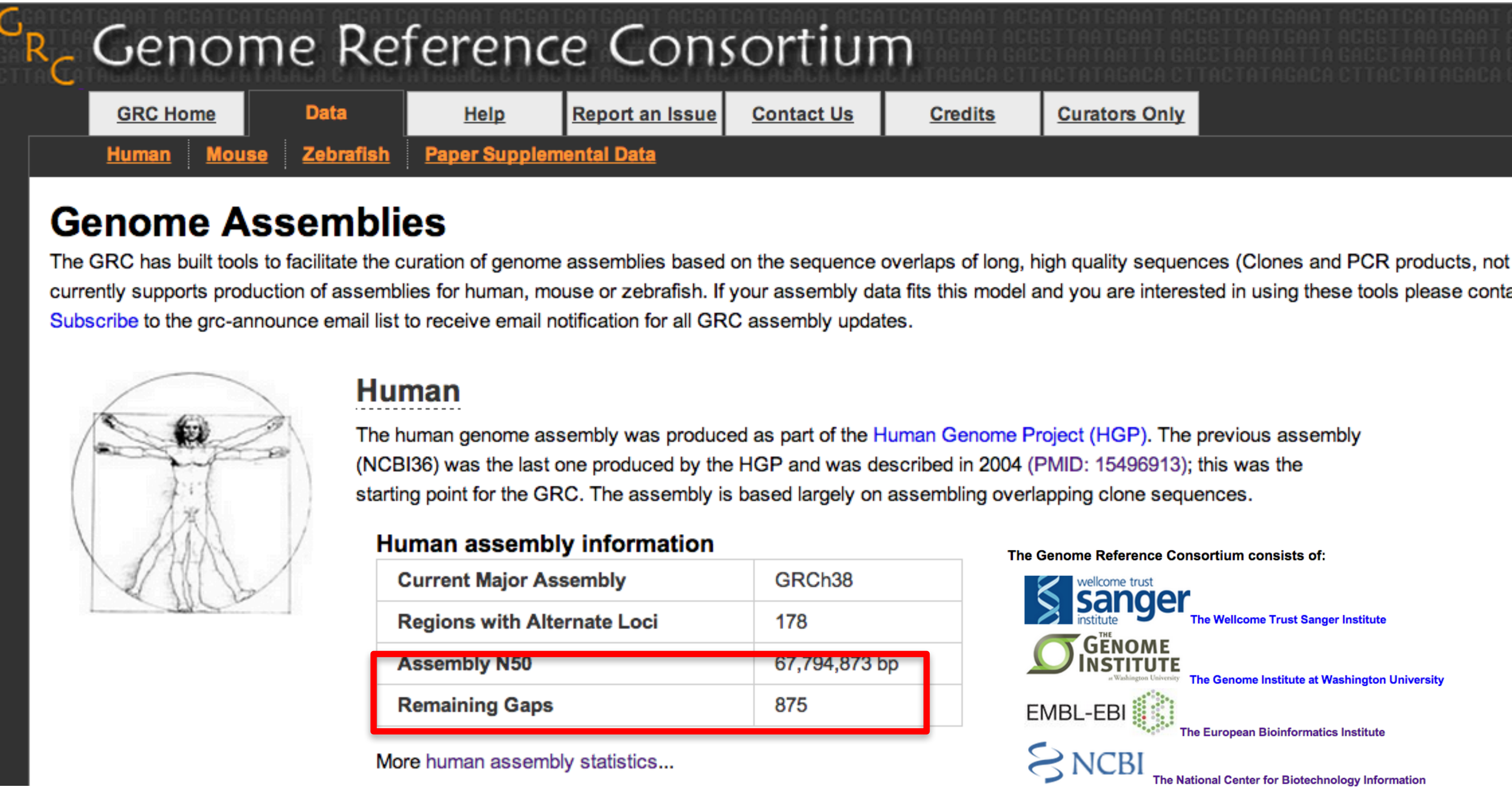
Contigs

Assemble

Scaffolds

Assemble

http://useast.ensembl.org/info/website/tutorials/index.html http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/data.shtml



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 conta

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:

The Wellcome Trust Sanger Institute

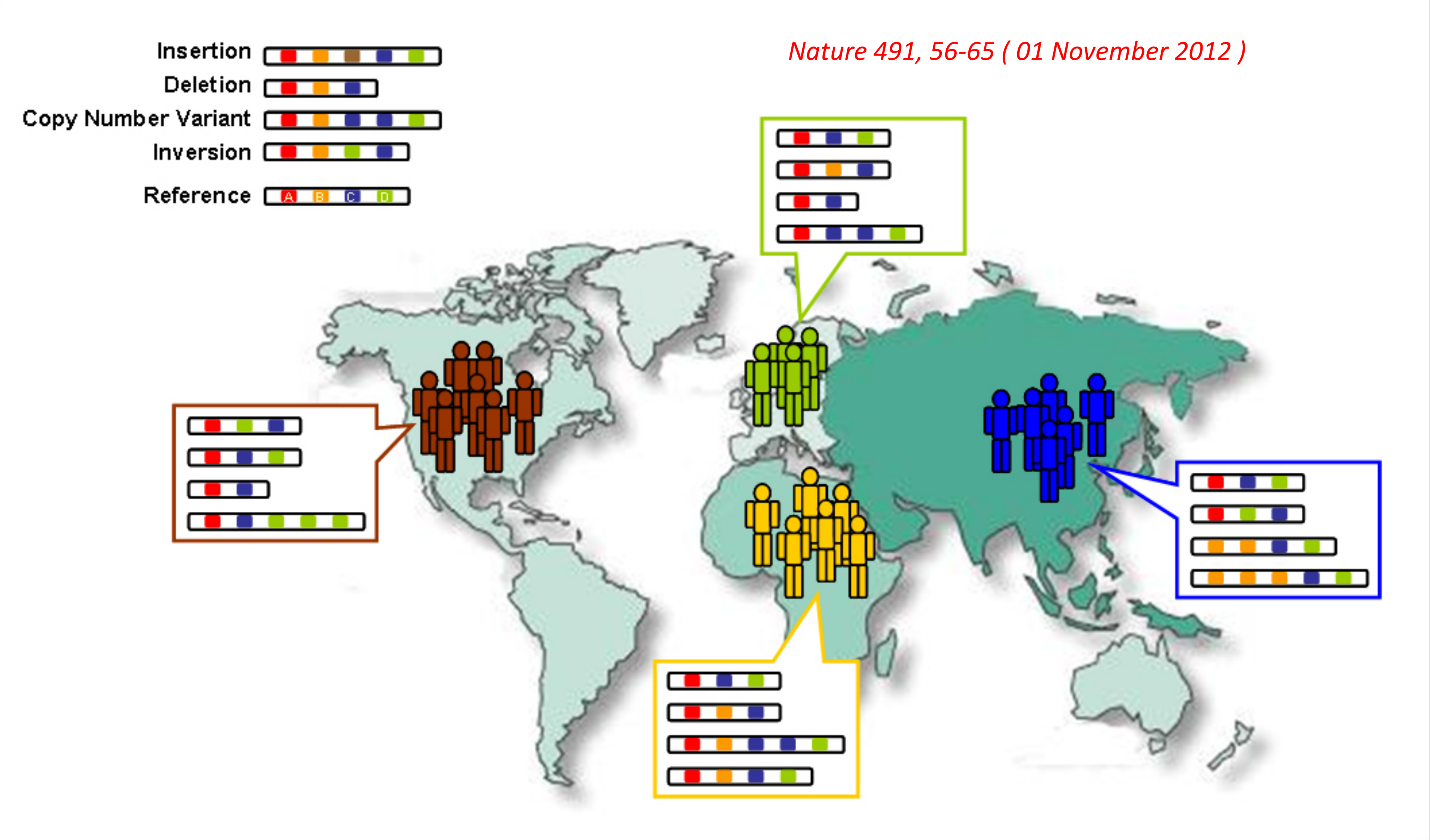
The Genome Institute at Washington University

The European Bioinformatics Institute

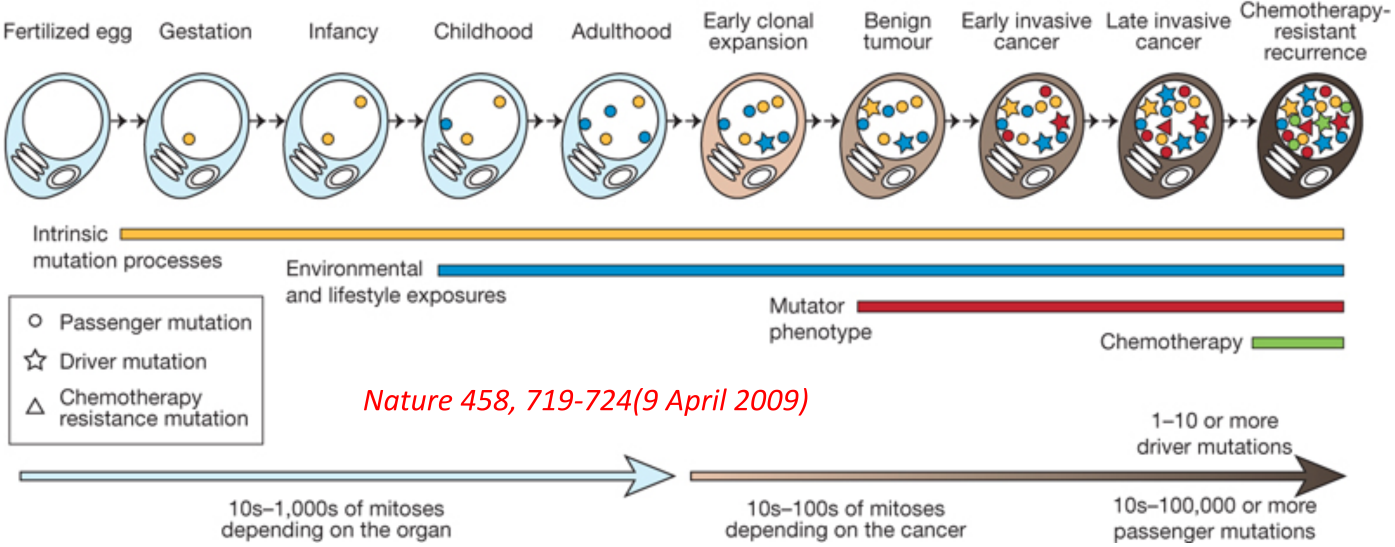
The National Center for Biotechnology Information

1000 Genomes

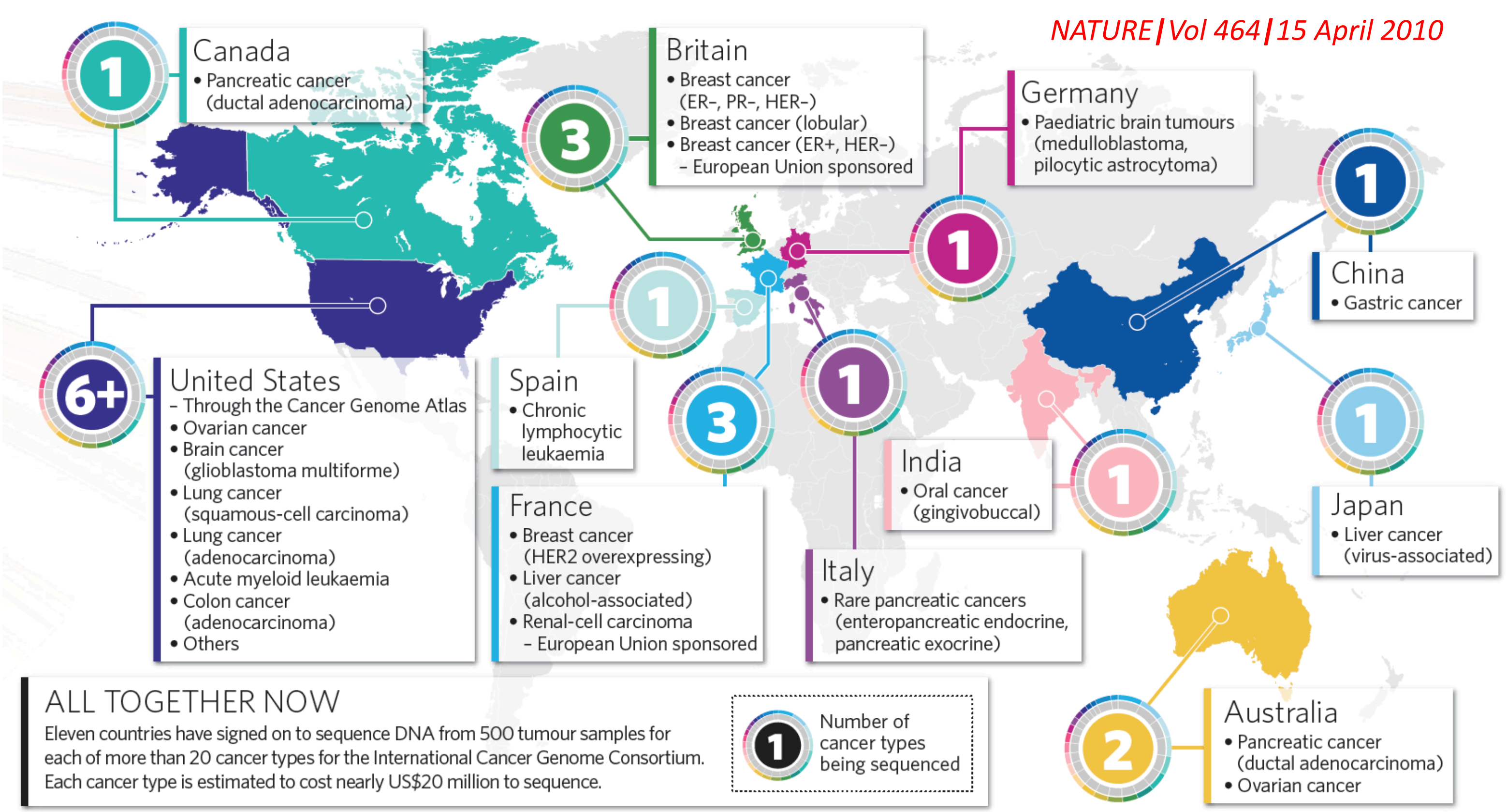
A Deep Catalog of Human Genetic Variation



Nature 491, 56-65 (01 November 2012)


Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
1–10 or more driver mutations
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
10s–100,000 or more passenger mutations

Nature 458, 719-724(9 April 2009)



NATURE | Vol 464 | 15 April 2010

Canada
- Pancreatic cancer (ductal adenocarcinoma)

Britain
- Breast cancer (ER–, PR–, HER–)
- Breast cancer (lobular)
- Breast cancer (ER+, HER–)
– European Union sponsored

Germany
- Paediatric brain tumours (medulloblastoma, pilocytic astrocytoma)

China
- Gastric cancer

United States
– Through the Cancer Genome Atlas
- Ovarian cancer
- Brain cancer (glioblastoma multiforme)
- Lung cancer (squamous-cell carcinoma)
- Lung cancer (adenocarcinoma)
- Acute myeloid leukaemia
- Colon cancer (adenocarcinoma)
- Others

Spain
- Chronic lymphocytic leukaemia

France
- Breast cancer (HER2 overexpressing)
- Liver cancer (alcohol-associated)
- Renal-cell carcinoma
– European Union sponsored

India
- Oral cancer (gingivobuccal)

Italy
- Rare pancreatic cancers (enteropancreatic endocrine, pancreatic exocrine)

Japan
- Liver cancer (virus-associated)

Australia
- Pancreatic cancer (ductal adenocarcinoma)
- Ovarian cancer

ALL TOGETHER NOW

Eleven countries have signed on to sequence DNA from 500 tumour samples for each of more than 20 cancer types for the International Cancer Genome Consortium. Each cancer type is estimated to cost nearly US$20 million to sequence.

1 Number of cancer types being sequenced

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

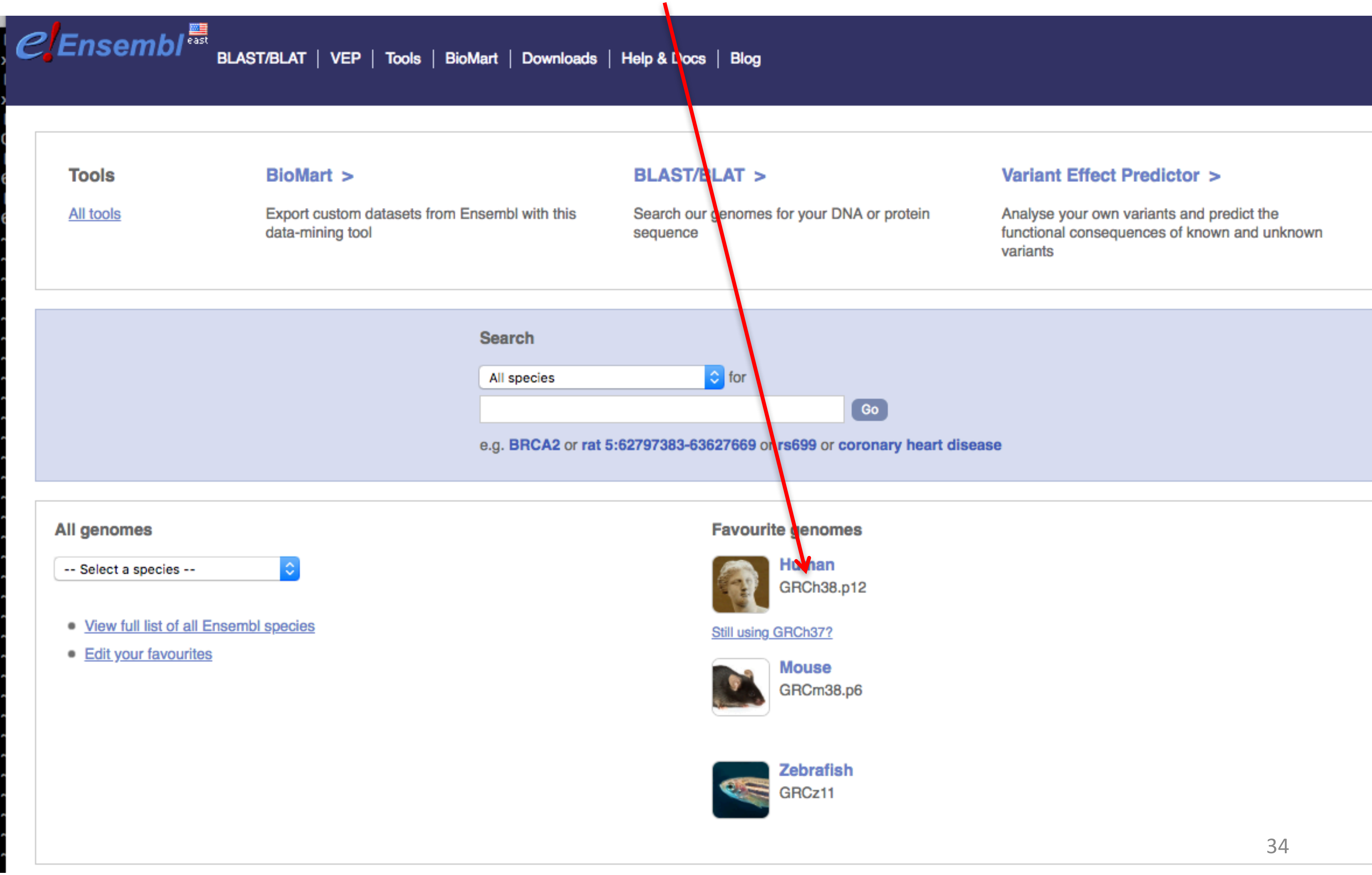
Each uses a centralized model, where the web site provides access to a large public database of genome data for many species and also integrates specialized tools, such as BLAST at NCBI and Ensembl and BLAT at UCSC.

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

http://www.ensembl.org/

Click to link to human page



e!Ensembl
east
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes

-- Select a species --

- View full list of all Ensembl species
- Edit your favourites

Favourite genomes

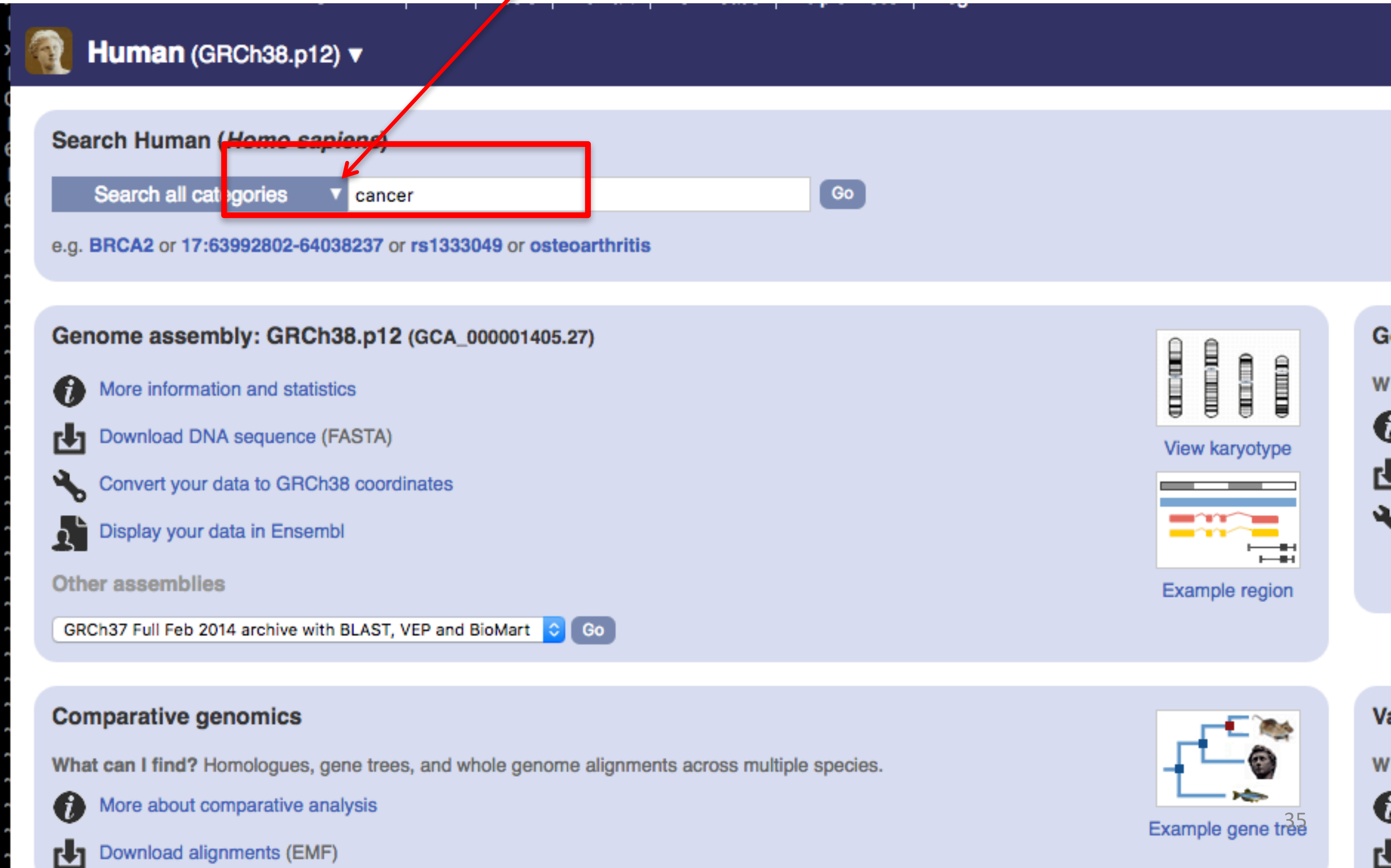
Human
GRCh38.p12

Still using GRCh37?

Mouse
GRCm38.p6

Zebrafish
GRCz11

Put “cancer” in the search box and Go



Human (GRCh38.p12) ▾

Search Human (*Homo sapiens*)

Search all categories ▾ cancer Go

e.g. BRCA2 or 17:63992802-64038237 or rs1333049 or osteoarthritis

Genome assembly: GRCh38.p12 (GCA_000001405.27)

More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh38 coordinates

Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart Go

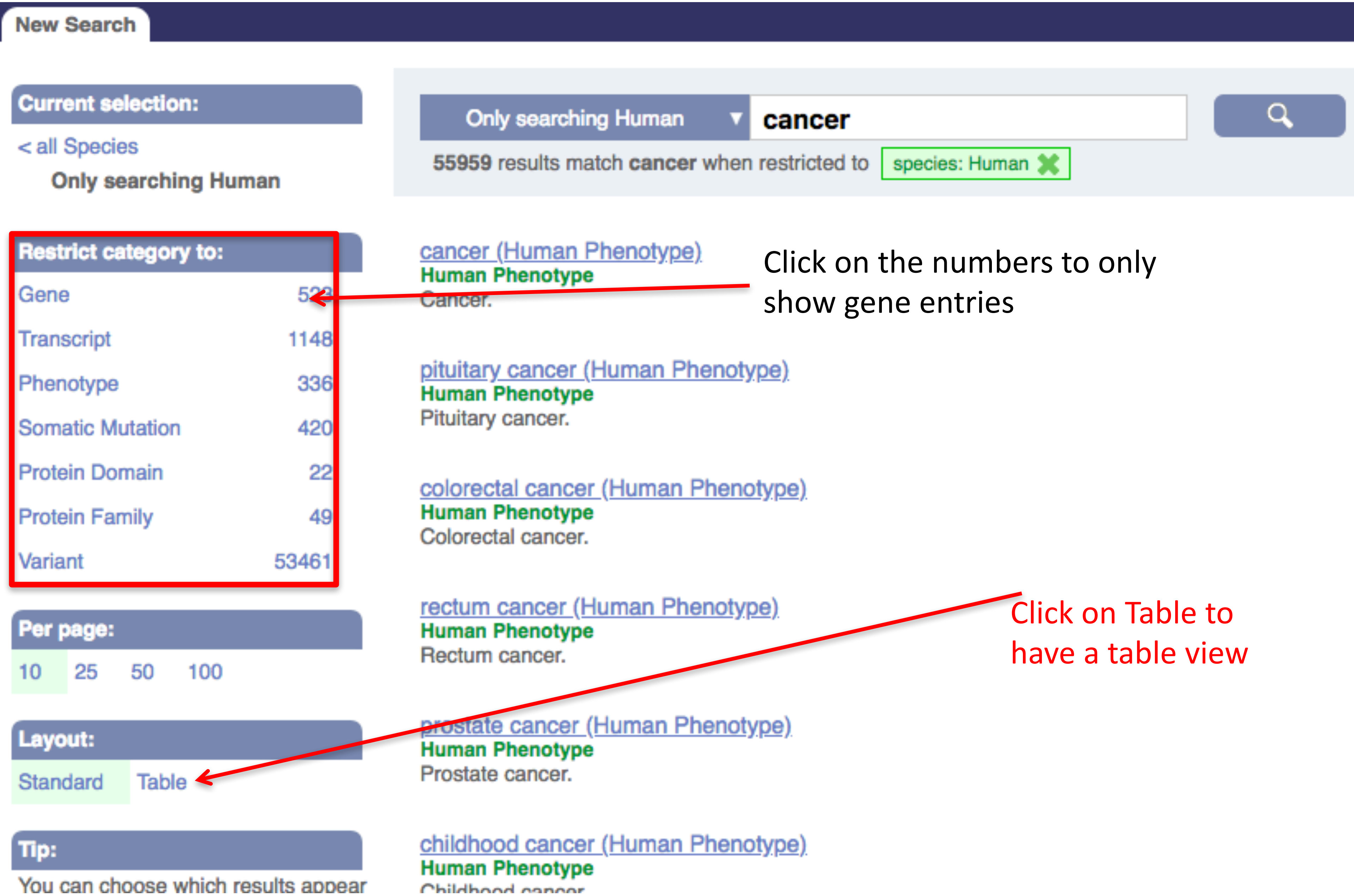
Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

This keyword search gives everything that contains “cancer”



cancer

55959 results match **cancer** when restricted to species: Human ✖

Current selection:

< all Species

Only searching Human

Restrict category to:	
Gene	523
Transcript	1148
Phenotype	336
Somatic Mutation	420
Protein Domain	22
Protein Family	49
Variant	53461

Per page:

10 25 50 100

Layout:

Standard Table

Tip:

You can choose which results appear

cancer (Human Phenotype)
Human Phenotype
Cancer.

pituitary cancer (Human Phenotype)
Human Phenotype
Pituitary cancer.

colorectal cancer (Human Phenotype)
Human Phenotype
Colorectal cancer.

rectum cancer (Human Phenotype)
Human Phenotype
Rectum cancer.

prostate cancer (Human Phenotype)
Human Phenotype
Prostate cancer.

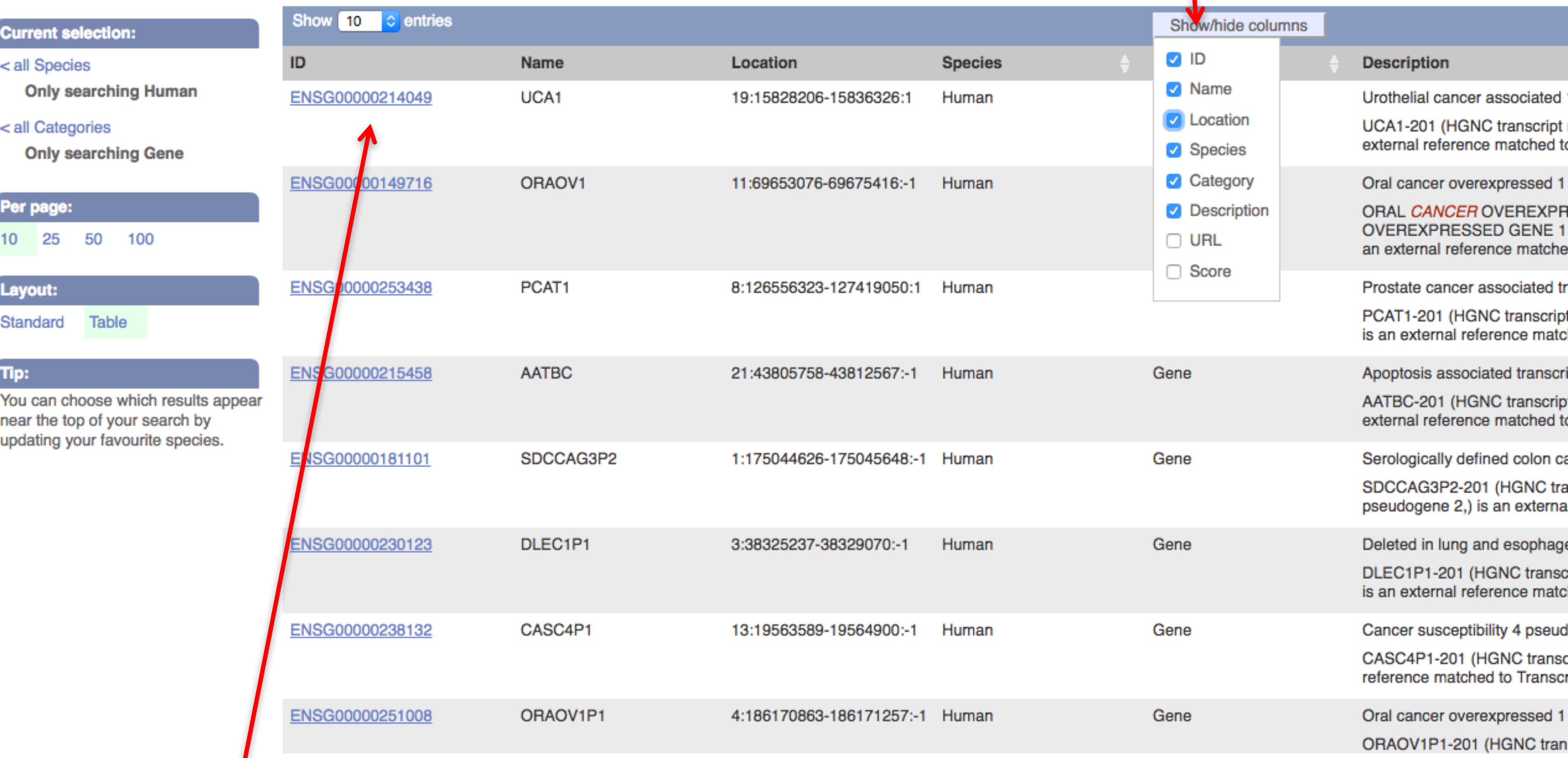
childhood cancer (Human Phenotype)
Human Phenotype
Childhood cancer

Click on the numbers to only show gene entries

Click on Table to have a table view

This is the list of genes

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



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

Show/hide columns

- ☑ ID
- ☑ Name
- ☑ Location
- ☑ Species
- ☑ Category
- ☑ Description
- □ URL
- □ Score

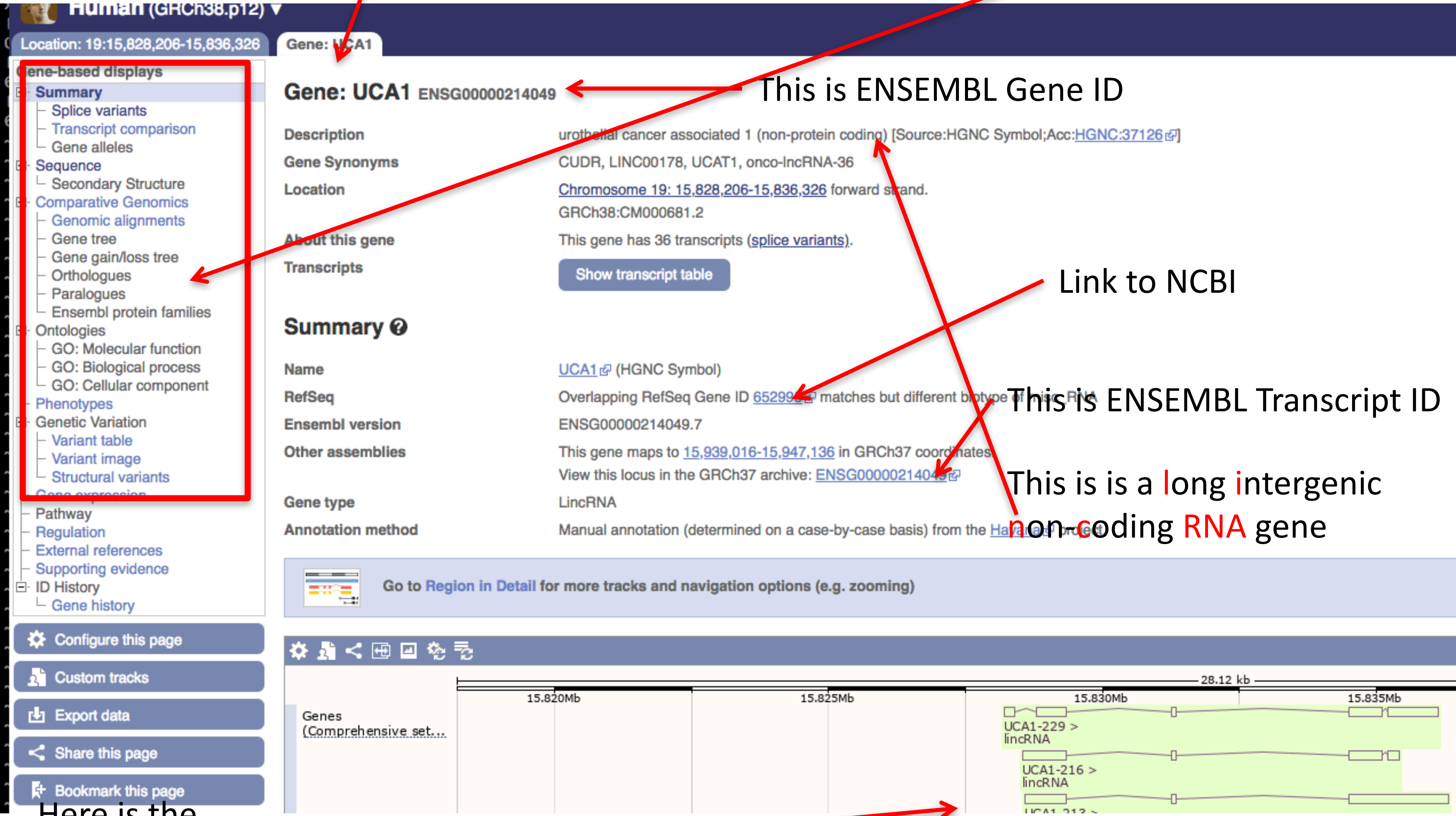
ID	Name	Location	Species		Description
ENSG00000214049	UCA1	19:15828206-15836326:1	Human		Urothelial cancer associated UCA1-201 (HGNC transcript external reference matched t
ENSG00000149716	ORAOV1	11:69653076-69675416:-1	Human		Oral cancer overexpressed 1 ORAL *CANCER* OVEREXPR OVEREXPRESSED GENE 1 an external reference matche
ENSG00000253438	PCAT1	8:126556323-127419050:1	Human		Prostate cancer associated tr PCAT1-201 (HGNC transcrip is an external reference matc
ENSG00000215458	AATBC	21:43805758-43812567:-1	Human	Gene	Apoptosis associated transcri AATBC-201 (HGNC transcrip external reference matched t
ENSG00000181101	SDCCAG3P2	1:175044626-175045648:-1	Human	Gene	Serologically defined colon ca SDCCAG3P2-201 (HGNC tra pseudogene 2,) is an externa
ENSG00000230123	DLEC1P1	3:38325237-38329070:-1	Human	Gene	Deleted in lung and esophage DLEC1P1-201 (HGNC transc is an external reference matc
ENSG00000238132	CASC4P1	13:19563589-19564900:-1	Human	Gene	Cancer susceptibility 4 pseud CASC4P1-201 (HGNC transc reference matched to Transcr
ENSG00000251008	ORAOV1P1	4:186170863-186171257:-1	Human	Gene	Oral cancer overexpressed 1 ORAOV1P1-201 (HGNC tran

The first entry in this page is a ncRNA gene.

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

Now it's showing the Gene; there is also a location tab

Many things can be explored



Human (GRCh38.p12) ▼

Location: 19:15,828,206-15,836,326 | Gene: UCA1

Gene-based displays

- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Comparative Genomics
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
  - Ensembl protein families
- Ontologies
  - GO: Molecular function
  - GO: Biological process
  - GO: Cellular component
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

Gene: UCA1 ENSG00000214049

Description	urothelial cancer associated 1 (non-protein coding) [Source:HGNC Symbol;Acc:HGNC:37126]
Gene Synonyms	CUDR, LINC00178, UCAT1, onco-lncRNA-36
Location	Chromosome 19: 15,828,206-15,836,326 forward strand. GRCh38:CM000681.2
About this gene	This gene has 36 transcripts (splice variants).
Transcripts	Show transcript table

Summary

Name	UCA1 (HGNC Symbol)
RefSeq	Overlapping RefSeq Gene ID 652995 matches but different biotype of misc_RNA
Ensembl version	ENSG00000214049.7
Other assemblies	This gene maps to 15,939,016-15,947,136 in GRCh37 coordinates. View this locus in the GRCh37 archive: ENSG00000214049
Gene type	LincRNA
Annotation method	Manual annotation (determined on a case-by-case basis) from the Havana project

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

This is ENSEMBL Gene ID

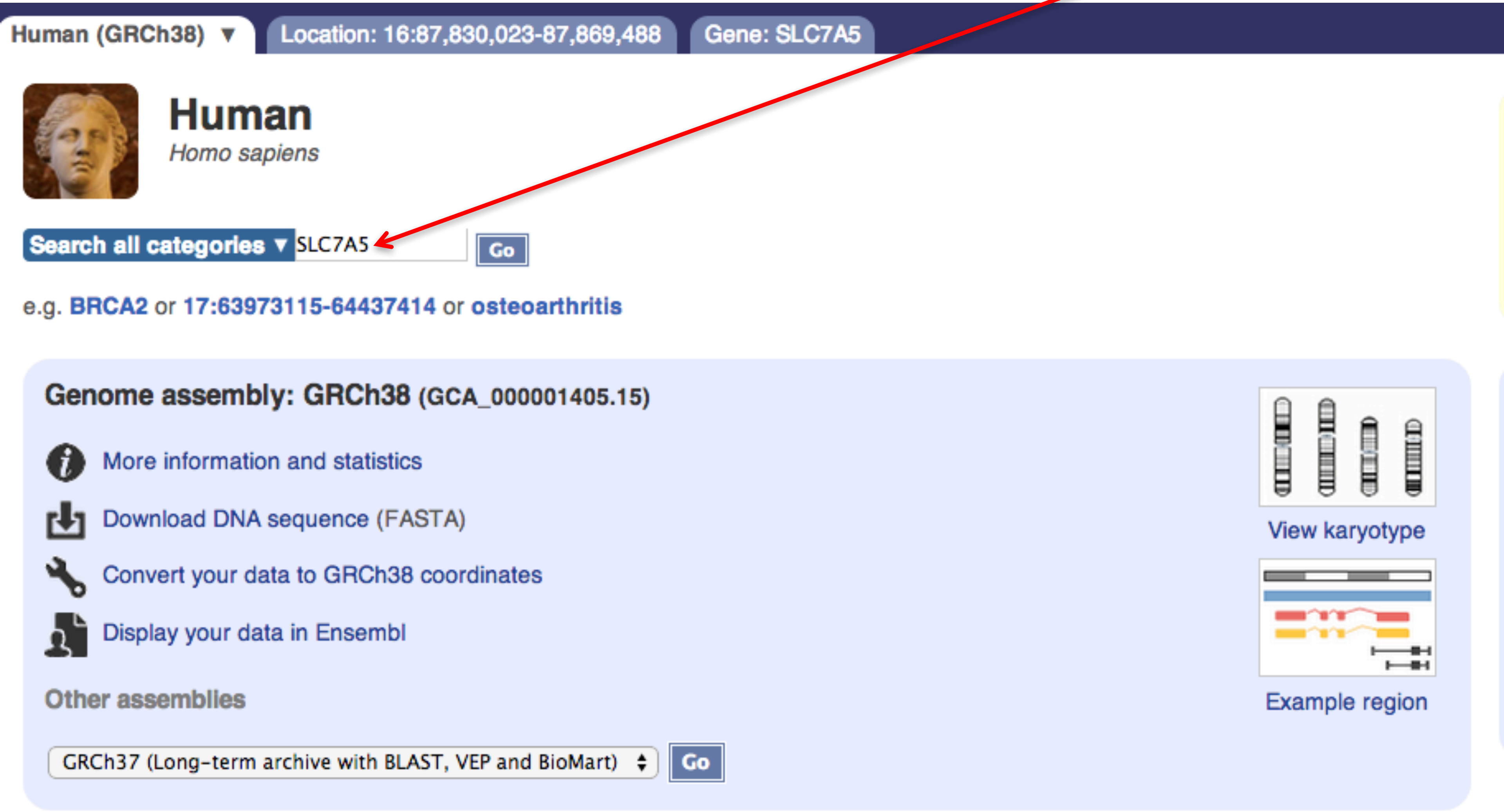
Link to NCBI

This is ENSEMBL Transcript ID

This is 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 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

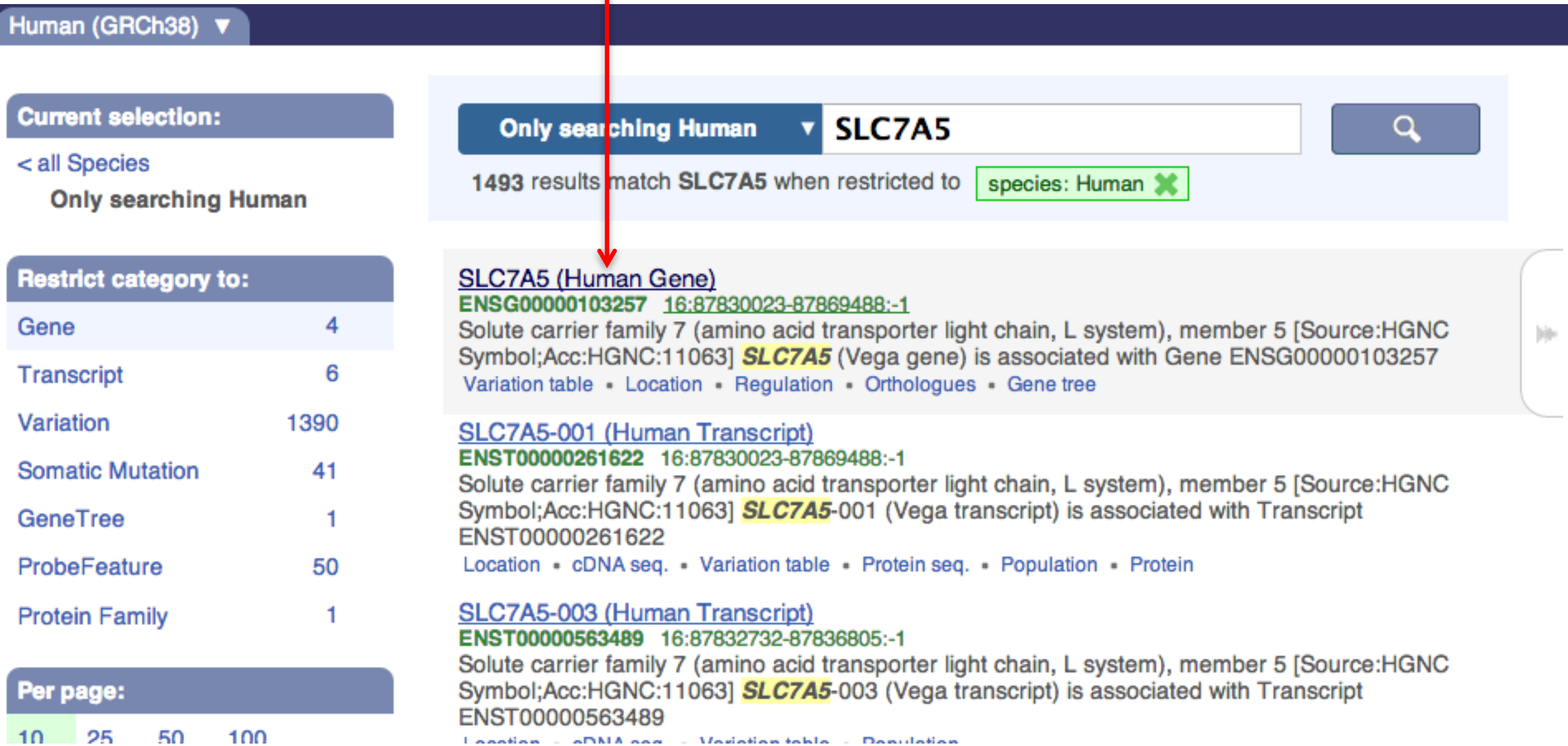
View karyotype

Example region

Other assemblies

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

Click here



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

SLC7A5

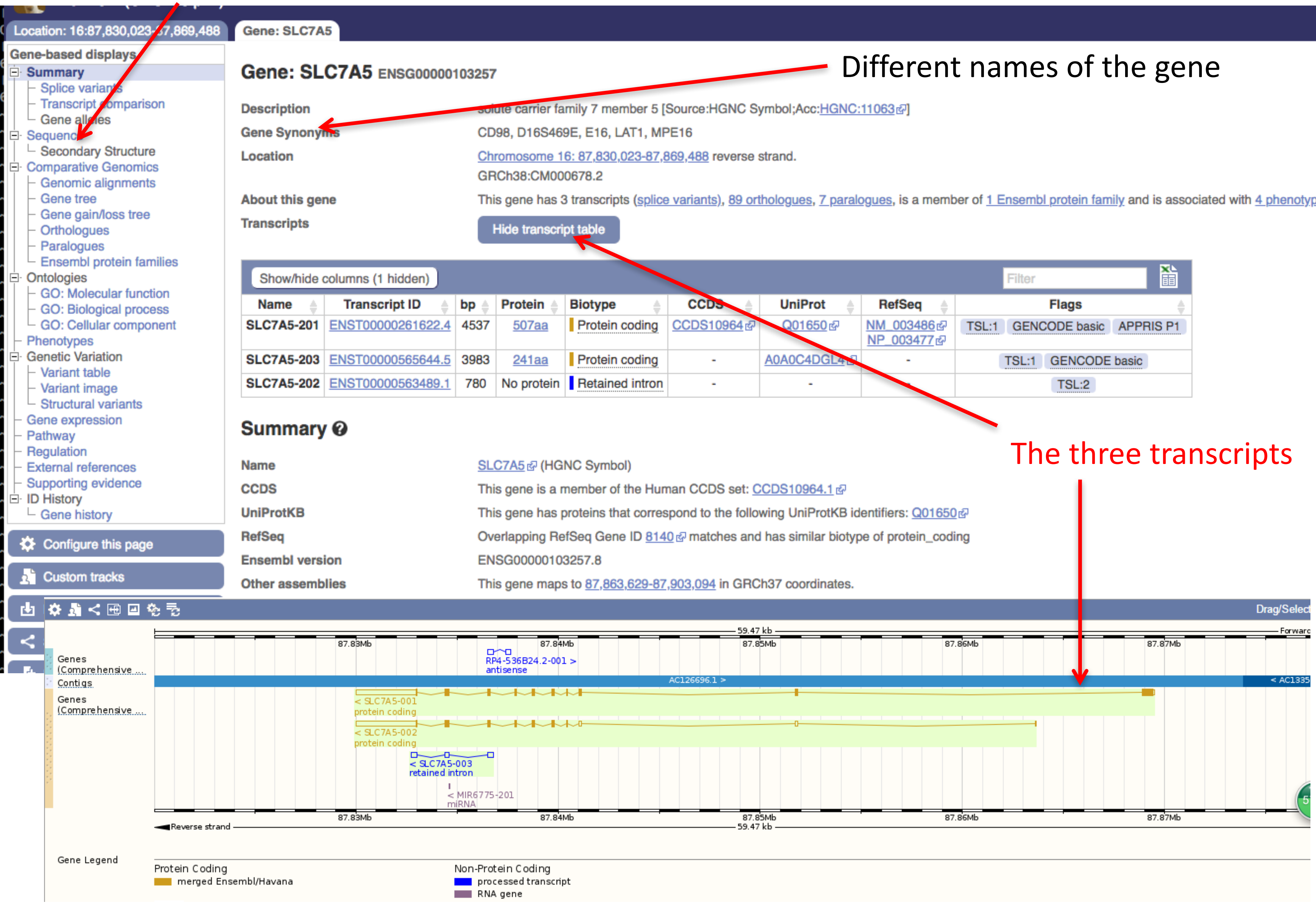
1493 results match SLC7A5 when restricted to species: Human ✖

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 • Population

Click to view the sequence page



Location: 16:87,830,023-87,869,488 | Gene: SLC7A5

Gene-based displays
- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Comparative Genomics
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
  - Ensembl protein families
- Ontologies
  - GO: Molecular function
  - GO: Biological process
  - GO: Cellular component
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history

Configure this page

Custom tracks

Gene: SLC7A5 ENSG00000103257

Description	Solute carrier family 7 member 5 [Source:HGNC Symbol;Acc:HGNC:11063]
Gene Synonyms	CD98, D16S469E, E16, LAT1, MPE16
Location	Chromosome 16: 87,830,023-87,869,488 reverse strand. GRCh38:CM000678.2
About this gene	This gene has 3 transcripts (splice variants), 89 orthologues, 7 paralogues, is a member of 1 Ensembl protein family and is associated with 4 phenotypes.
Transcripts	Hide transcript table

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
SLC7A5-201	ENST00000261622.4	4537	507aa	Protein coding	CCDS10964	Q01650	NM_003486 NP_003477	TSL:1 GENCODE basic APPRIS P1
SLC7A5-203	ENST00000565644.5	3983	241aa	Protein coding	-	A0A0C4DGL4	-	TSL:1 GENCODE basic
SLC7A5-202	ENST00000563489.1	780	No protein	Retained intron	-	-	-	TSL:2

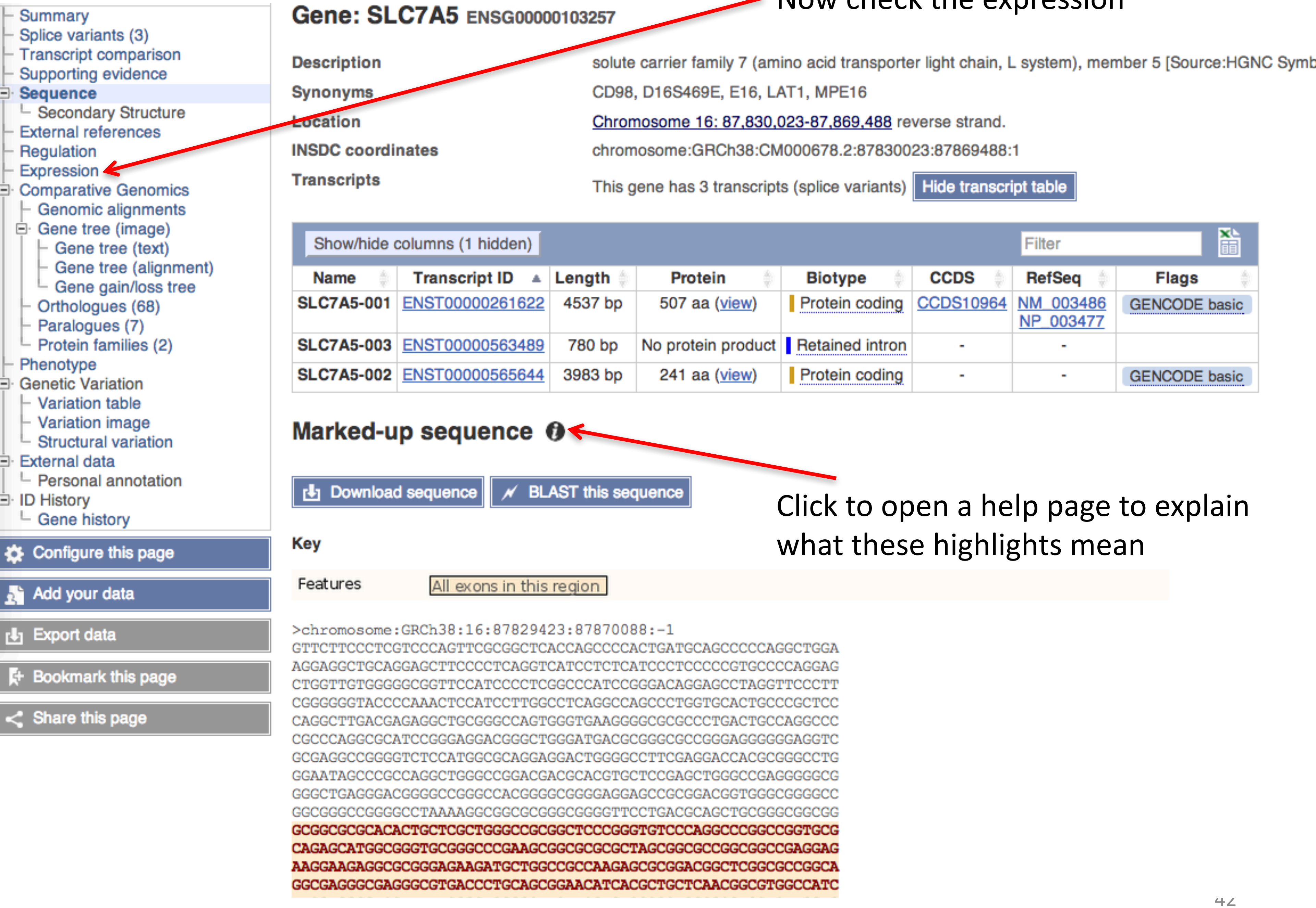
Summary

Name	SLC7A5 (HGNC Symbol)
CCDS	This gene is a member of the Human CCDS set: CCDS10964.1
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: Q01650
RefSeq	Overlapping RefSeq Gene ID 8140 matches and has similar biotype of protein_coding
Ensembl version	ENSG00000103257.8
Other assemblies	This gene maps to 87,863,629-87,903,094 in GRCh37 coordinates.

Different names of the gene

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 Symb
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	ENST00000261622	4537 bp	507 aa (view)	Protein coding	CCDS10964	NM_003486 NP_003477	GENCODE basic
SLC7A5-003	ENST00000563489	780 bp	No protein product	Retained intron	-	-	
SLC7A5-002	ENST00000565644	3983 bp	241 aa (view)	Protein coding	-	-	GENCODE basic

Marked-up sequence

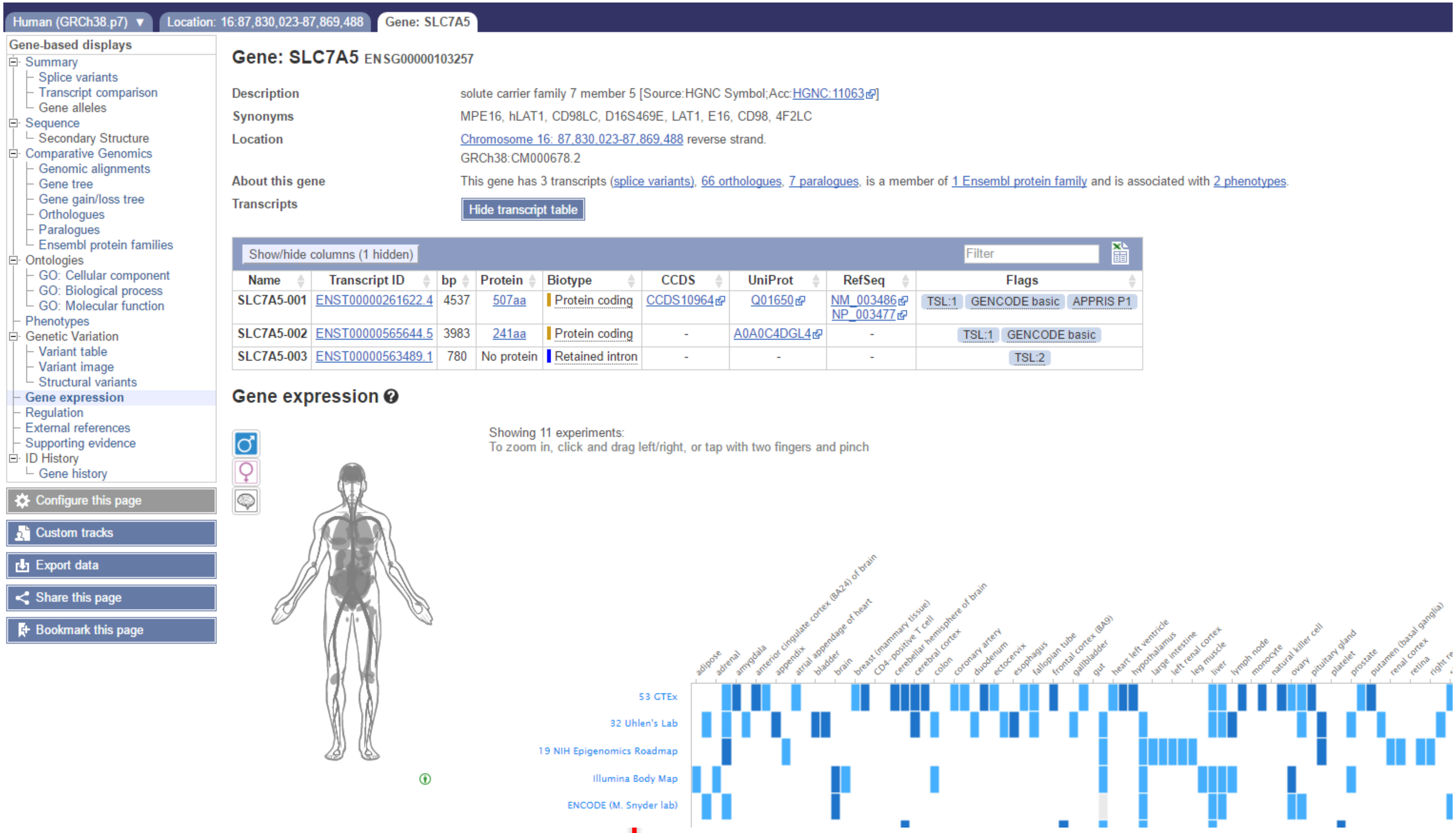
Download sequence | BLAST this sequence

Key

Features: All exons in this region

```
>chromosome:GRCh38:16:87829423:87870088:-1
GTTCTTCCCTCGTCCCAGTTCGCGGCTCACCAGCCCCACTGATGCAGCCCCCAGGCTGGA
AGGAGGCTGCAGGAGCTTCCCCTCAGGTCATCCTCTCATCCCTCCCCCGTGCCCCAGGAG
CTGGTTGTGGGGGCGGTTCCATCCCCTCGGCCCATCCGGGACAGGAGCCTAGGTTCCCTT
CGGGGGGTACCCCAAACTCCATCCTTGGCCTCAGGCCAGCCCTGGTGCACTGCCCGCTCC
CAGGCTTGACGAGAGGCTGCGGGCCAGTGGGTGAAGGGGCGCGCCCTGACTGCCAGGCCC
CGCCCAGGCGCATCCGGGAGGACGGGCTGGGATGACGCGGGCGCCGGGAGGGGGAGGTC
GCGAGGCCGGGGTCTCCATGGCGCAGGAGGACTGGGGCCTTCGAGGACCACGCGGGCCTG
GGAATAGCCCGCCAGGCTGGGCCGGACGACGCACGTGCTCCGAGCTGGGCCGAGGGGGCG
GGGCTGAGGGACGGGGCCGGGCCACGGGGCGGGGAGGAGCCGCGGACGGTGGGCGGGGCC
GGCGGGCCGGGGCCTAAAAGGCGGCGCGGGCGGGGTTCCTGACGCAGCTGCGGGCGGCGG
GCGGCGCGCACACTGCTCGCTGGGCCGCGGCTCCCGGGTGTCCCAGGCCCGGCCGGTGCG
CAGAGCATGGCGGGTGCGGGCCCGAAGCGGCGCGCTAGCGGCGCCGGCGGCCGAGGAG
AAGGAAGAGGCGCGGGAGAAGATGCTGGCCGCCAAGAGCGCGGACGGCTCGGCGCCGGCA
GGCGAGGGCGAGGGCGTGACCCTGCAGCGGAACATCACGCTGCTCAACGGCGTGGCCATC
```

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



Gene: SLC7A5 ENSG00000103257

Description: solute carrier family 7 member 5 [Source:HGNC Symbol;Acc:HGNC:11063]

Synonyms: MPE16, hLAT1, CD98LC, D16S469E, LAT1, E16, CD98, 4F2LC

Location: Chromosome 16: 87,830,023-87,869,488 reverse strand.
GRCh38:CM000678.2

About this gene: This gene has 3 transcripts (splice variants), 66 orthologues, 7 paralogues, is a member of 1 Ensembl protein family and is associated with 2 phenotypes.

Transcripts: Hide transcript table

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
SLC7A5-001	ENST00000261622.4	4537	507aa	Protein coding	CCDS10964	Q01650	NM_003486 NP_003477	TSL:1 GENCODE basic APPRIS P1
SLC7A5-002	ENST00000565644.5	3983	241aa	Protein coding	-	A0A0C4DGL4	-	TSL:1 GENCODE basic
SLC7A5-003	ENST00000563489.1	780	No protein	Retained intron	-	-	-	TSL:2

Gene expression

Showing 11 experiments:
To zoom in, click and drag left/right, or tap with two fingers and pinch

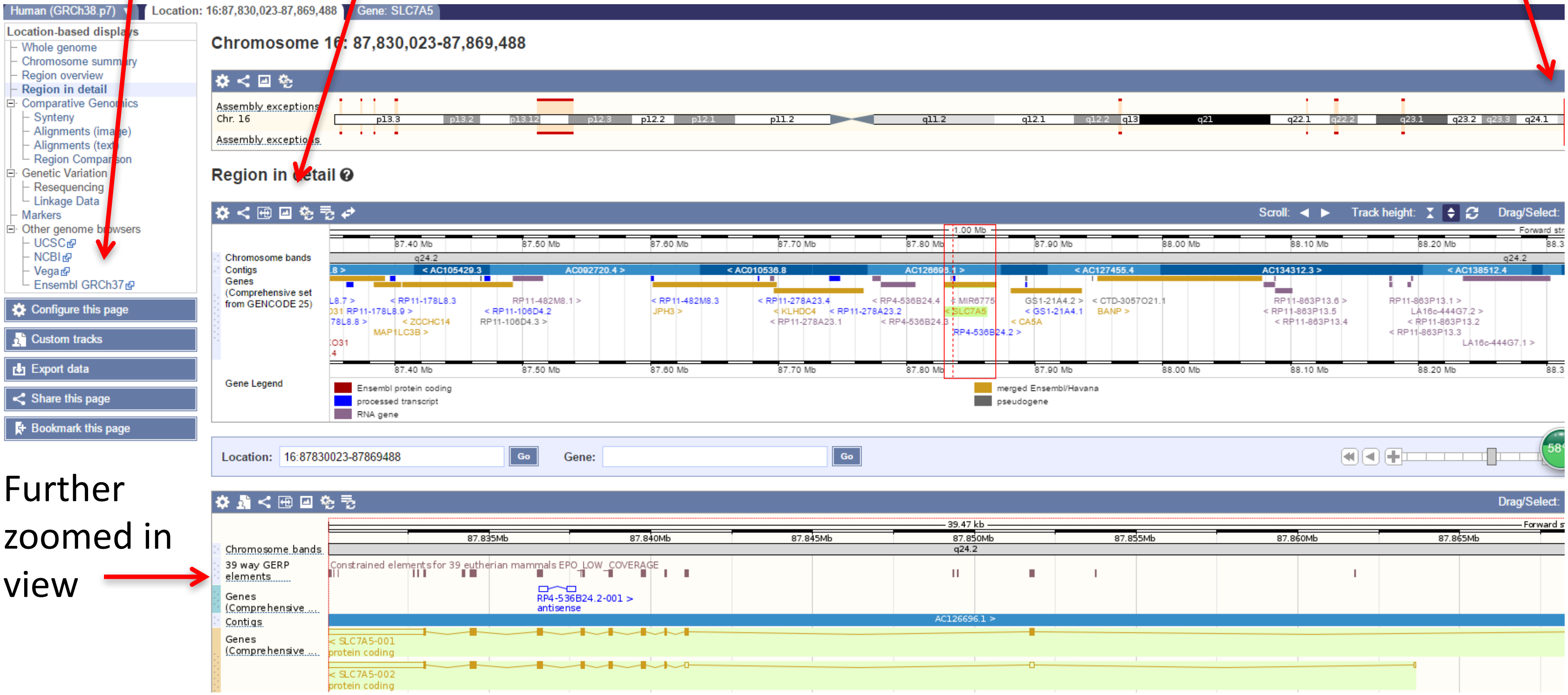
Different genome-wide expression studies

Links to other genome browsers

Zoomed in view

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

Further zoomed in view



Human (GRCh38.p7) Location: 16:87,830,023-87,869,488 Gene: SLC7A5

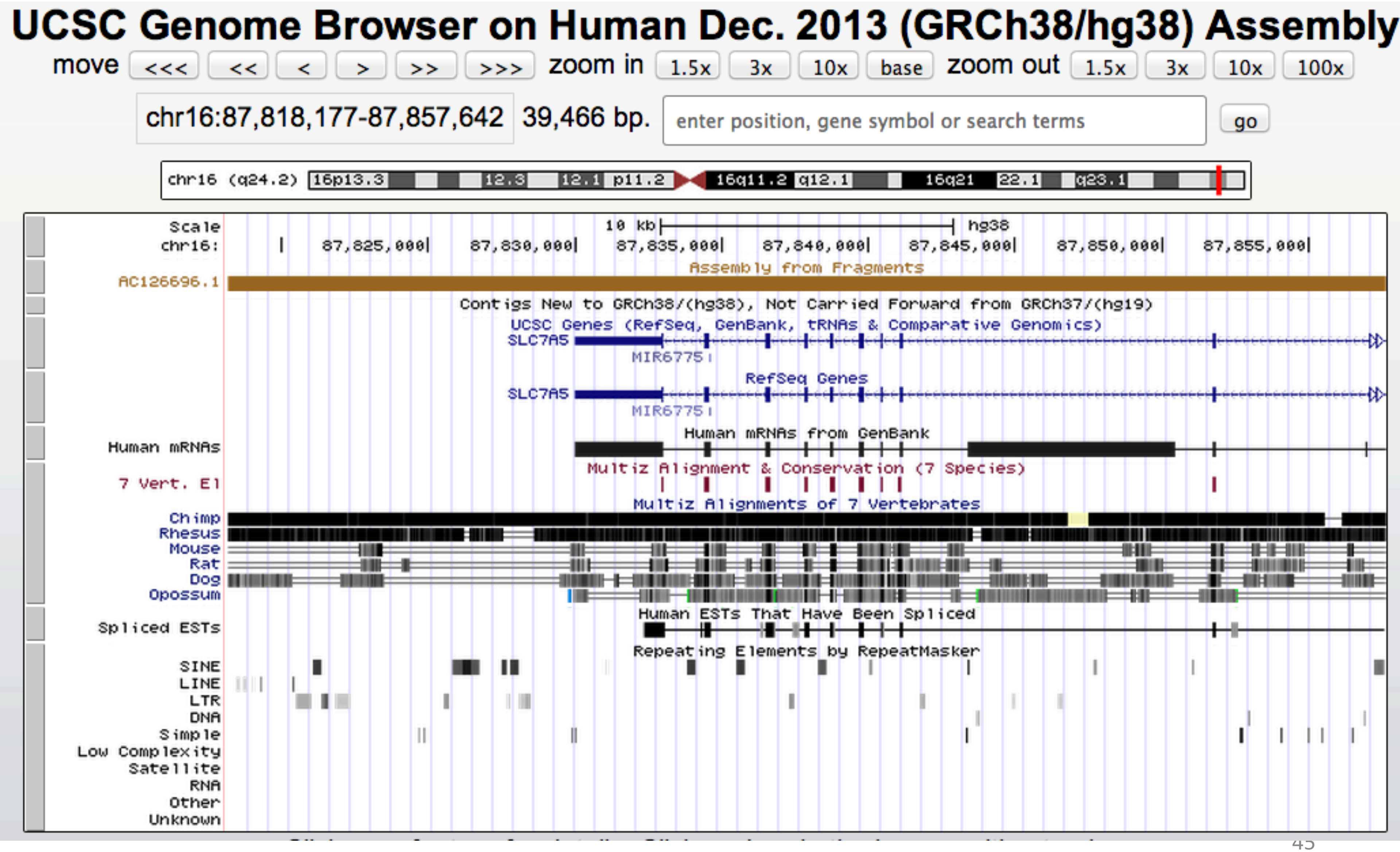
Chromosome 16: 87,830,023-87,869,488

Region in detail

Location: 16:87830023-87869488 Go Gene: Go

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

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

Next lecture: ExPASy and DTU tools