EBI web resources II: Ensembl and InterPro

Yanbin Yin

Spring 2013

Outline

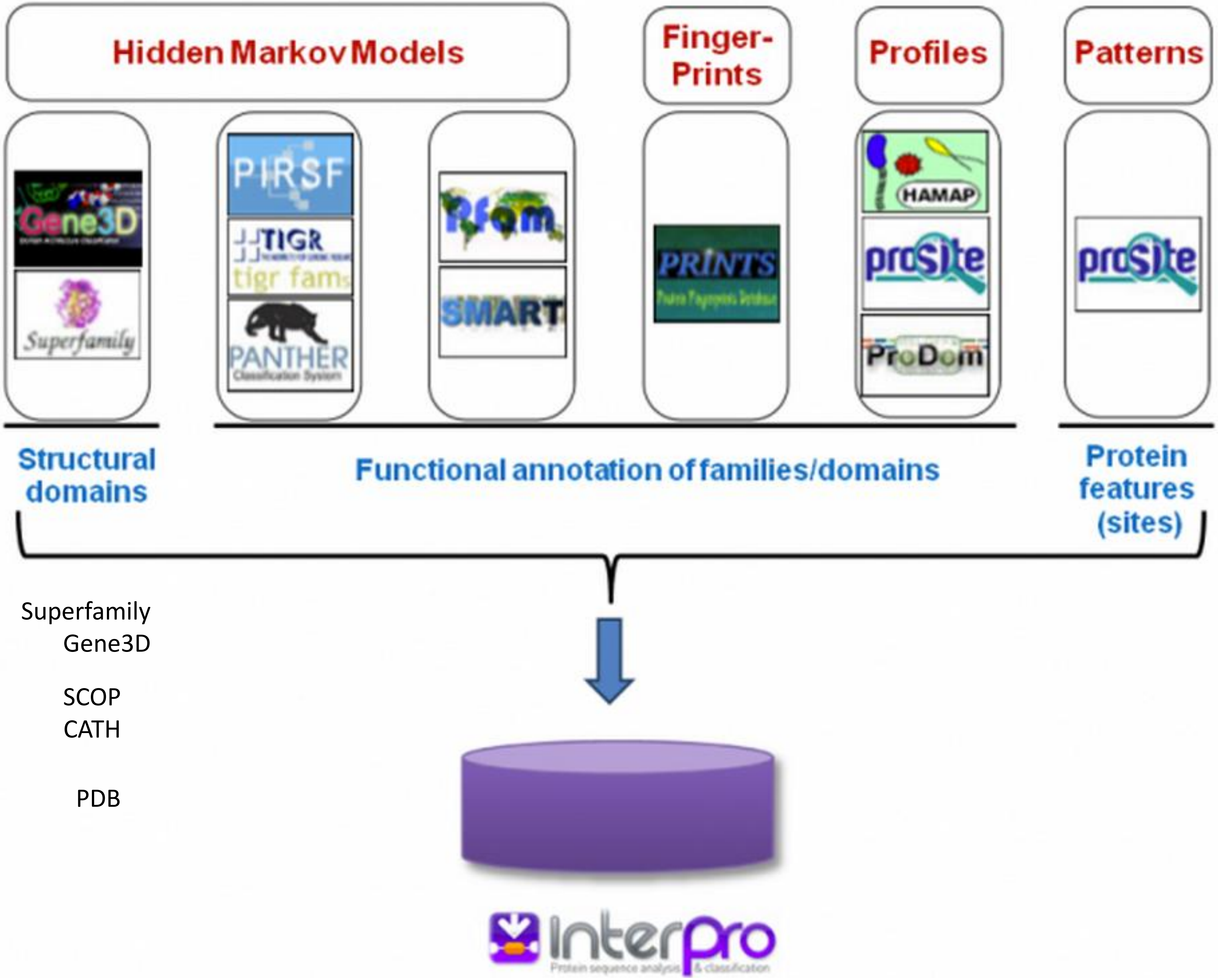
- Intro to genome annotation
- Protein family/domain databases
  - InterPro, Pfam, Superfamily etc.
- Genome browser
  - Ensembl
- Hands on Practice

Genome annotation

- Predict genes (where are the genes?)
  - protein coding
  - RNA coding

- Function annotation (What are the genes?)
  - Search against UniProt or NCBI-nr (GenPept)
  - Search against protein family/domain databases
  - Search against Pathway databases

Function vocabularies defined in Gene Ontology



Superfamily
Gene3D

SCOP
CATH

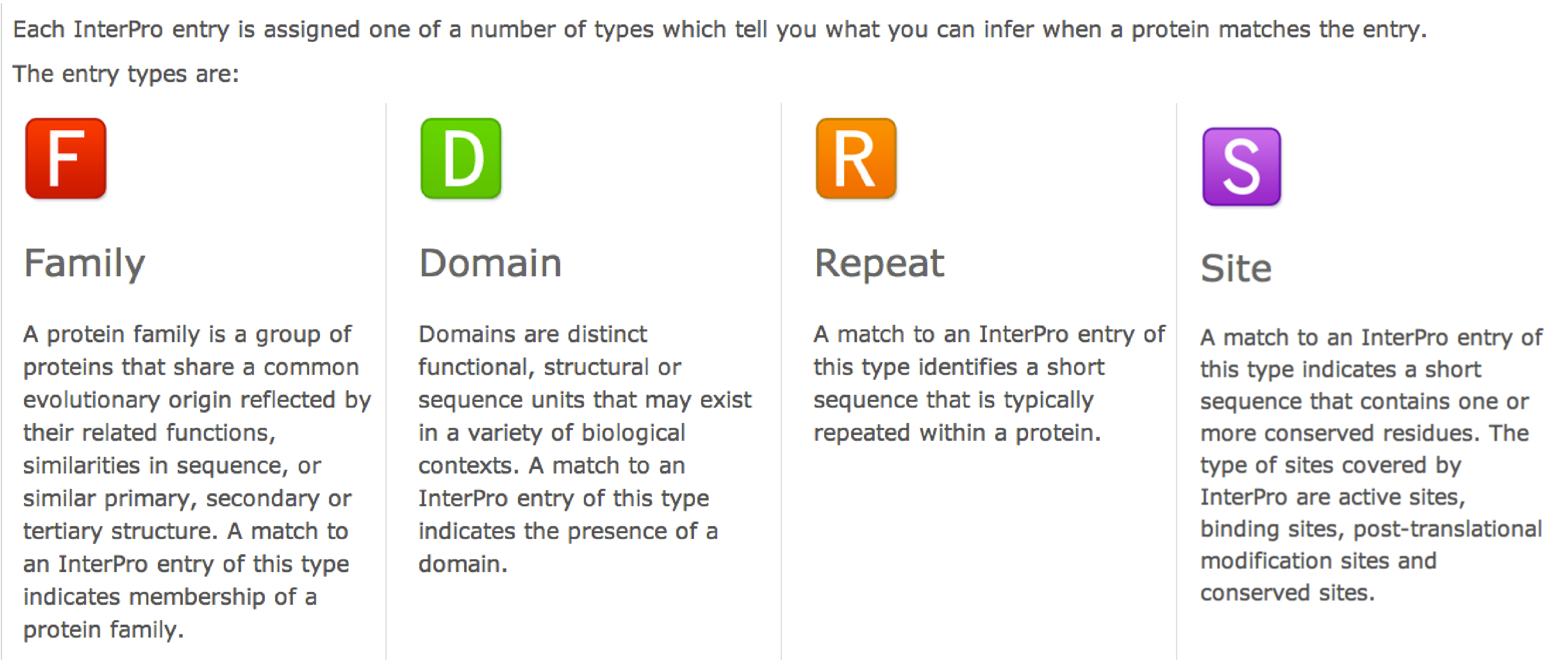
PDB

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



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 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.

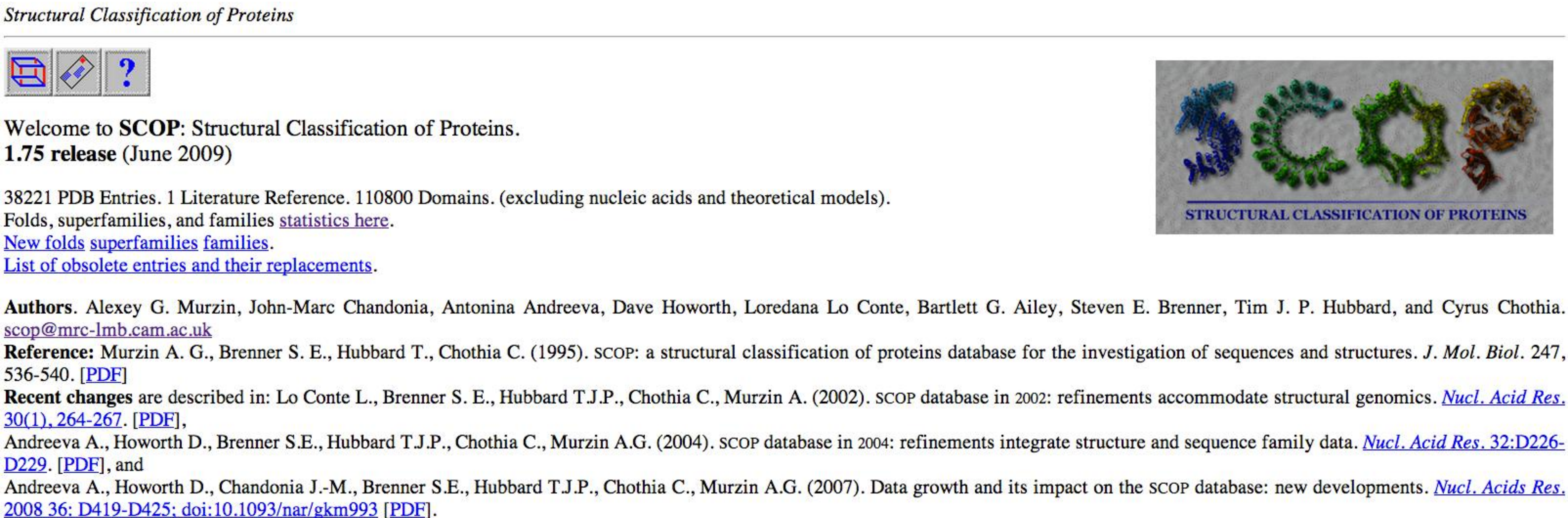
Superfamily: Probable common evolutionary origin

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.

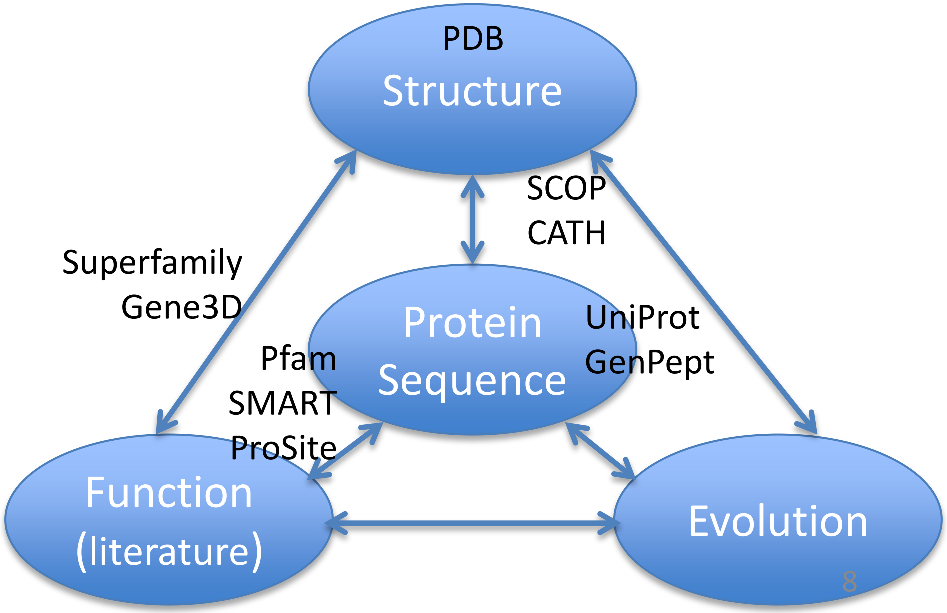
Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia.
scop@mrc-lmb.cam.ac.uk
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification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



PDB
Structure

SCOP
CATH

Superfamily
Gene3D

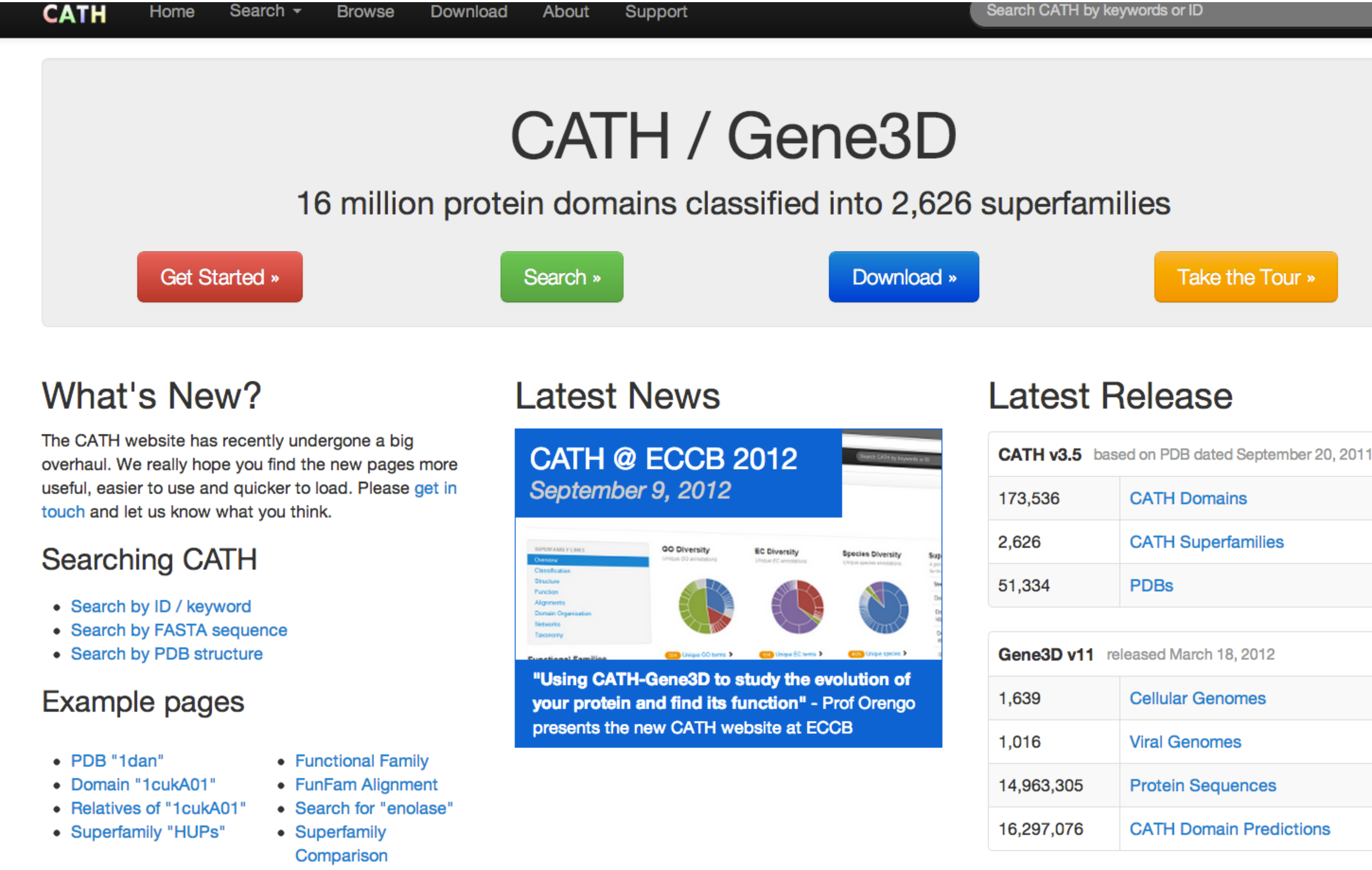
Protein
Sequence

UniProt
GenPept

Pfam
SMART
ProSite

Function
(literature)

Evolution



CATH / Gene3D

16 million protein domains classified into 2,626 superfamilies

What's New?

The CATH website has recently undergone a big overhaul. We really hope you find the new pages more useful, easier to use and quicker to load. Please get in touch and let us know what you think.

Searching CATH

- Search by ID / keyword
- Search by FASTA sequence
- Search by PDB structure

Example pages

- PDB "1dan"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HUPs"
- Functional Family
- FunFam Alignment
- Search for "enolase"
- Superfamily Comparison

Latest News

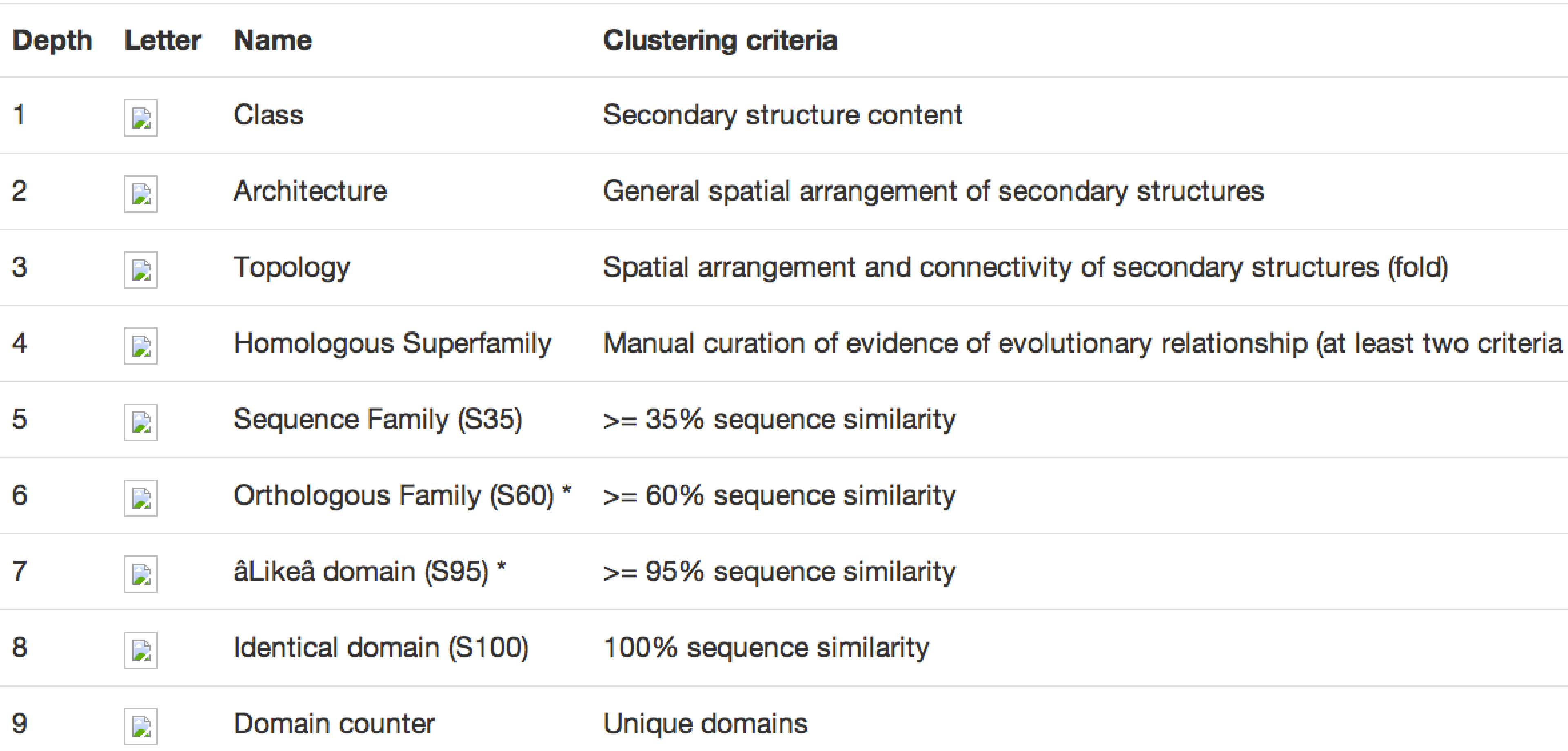
CATH @ ECCB 2012
September 9, 2012

"Using CATH-Gene3D to study the evolution of your protein and find its function" - Prof Orengo presents the new CATH website at ECCB

Latest Release

CATH v3.5	based on PDB dated September 20, 2011
173,536	CATH Domains
2,626	CATH Superfamilies
51,334	PDBs

Gene3D v11	released March 18, 2012
1,639	Cellular Genomes
1,016	Viral Genomes
14,963,305	Protein Sequences
16,297,076	CATH Domain Predictions

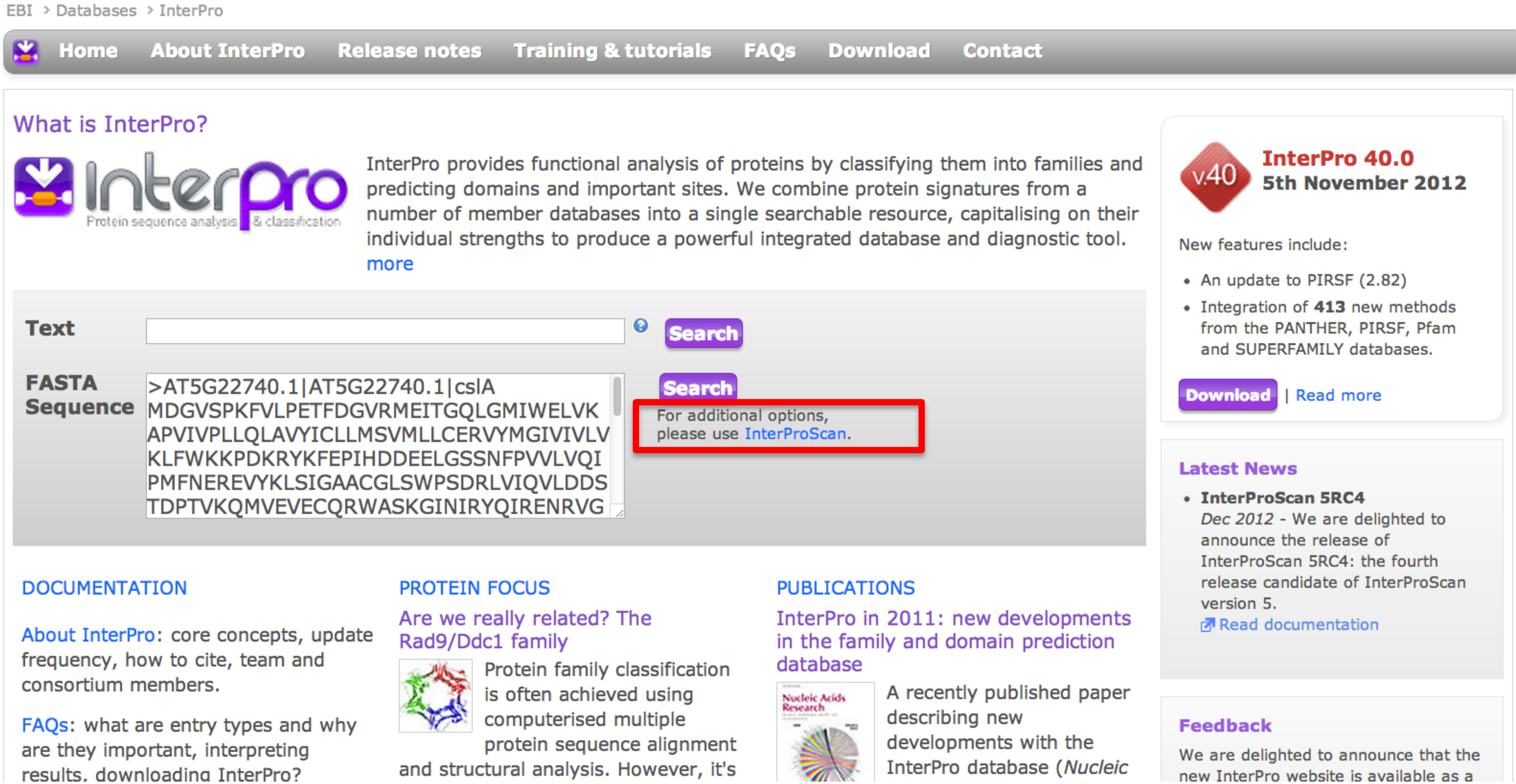


Depth	Letter	Name	Clustering criteria
1		Class	Secondary structure content
2		Architecture	General spatial arrangement of secondary structures
3		Topology	Spatial arrangement and connectivity of secondary structures (fold)
4		Homologous Superfamily	Manual curation of evidence of evolutionary relationship (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

Hands on exercise 1: search against protein family databases

Google “interpro”



What is 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. more

Text

FASTA Sequence

```
>AT5G22740.1|AT5G22740.1|cslA
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVK
APVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLV
KLFWKKPDKRYKFEPIHDDEELGSSNFPVVLVQI
PMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDS
TDPTVKQMVEVECQRWASKGINIRYQIRENRVG
```

For additional options, please use InterProScan.

InterPro 40.0
5th November 2012

v.40

New features include:

- An update to PIRSF (2.82)
- Integration of **413** new methods from the PANTHER, PIRSF, Pfam and SUPERFAMILY databases.

Download | Read more

Latest News

- **InterProScan 5RC4**
  Dec 2012 - We are delighted to announce the release of InterProScan 5RC4: the fourth release candidate of InterProScan version 5.
  Read documentation

Feedback

We are delighted to announce that the new InterPro website is available as a

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?

PROTEIN FOCUS

Are we really related? The Rad9/Ddc1 family

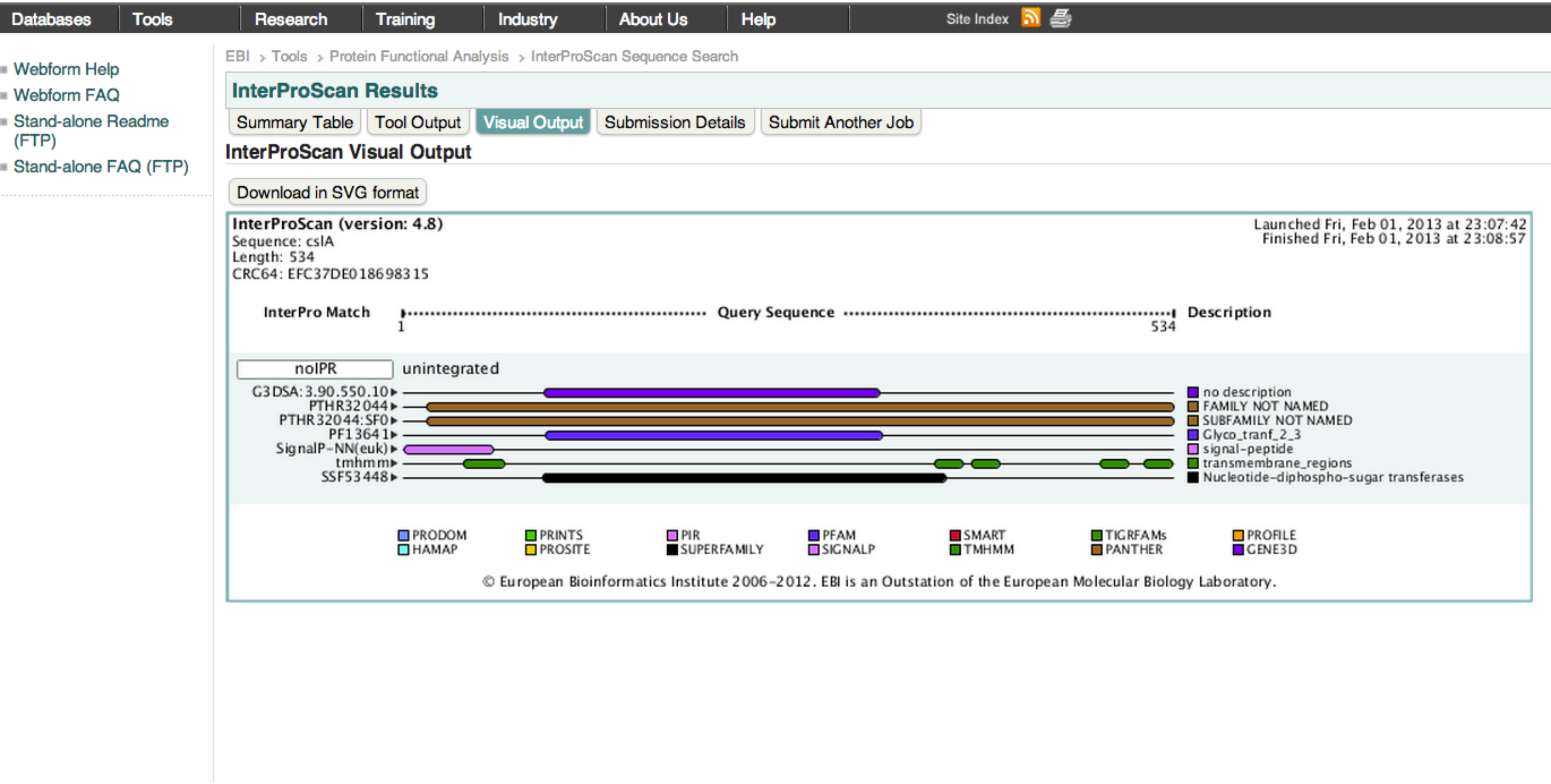
Protein family classification is often achieved using computerised multiple protein sequence alignment and structural analysis. However, it's

PUBLICATIONS

InterPro in 2011: new developments in the family and domain prediction database

Nucleic Acids Research

A recently published paper describing new developments with the InterPro database (*Nucleic*



InterProScan Results

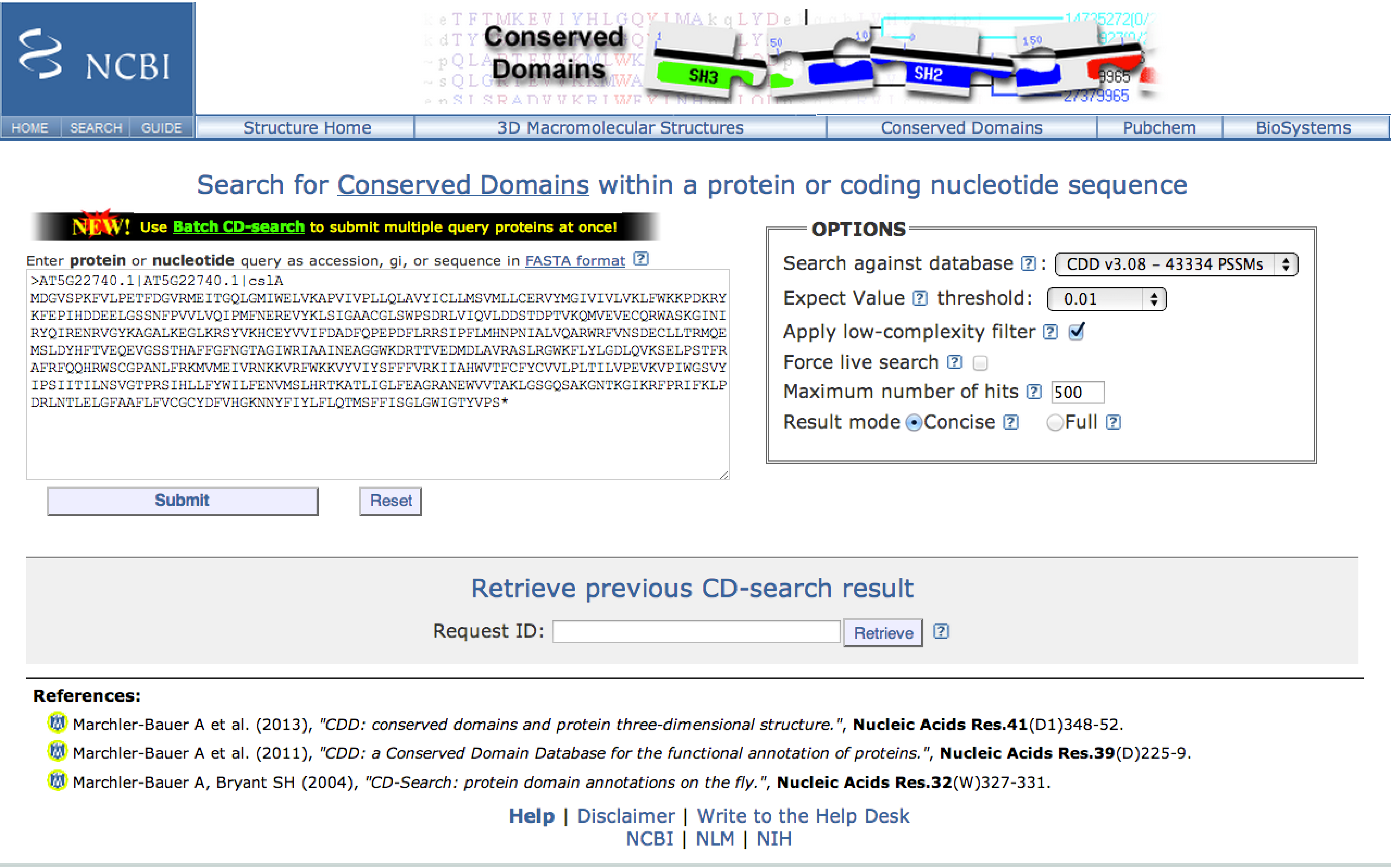
InterProScan Visual Output

InterProScan (version: 4.8)
Sequence: cslA
Length: 534
CRC64: EFC37DE018698315

Launched Fri, Feb 01, 2013 at 23:07:42
Finished Fri, Feb 01, 2013 at 23:08:57

© European Bioinformatics Institute 2006-2012. EBI is an Outstation of the European Molecular Biology Laboratory.

Google “NCBI CDD search”



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 ?

```
>AT5G22740.1|AT5G22740.1|cslA
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLVKLFWKKPDKRY
KFEPIHDDEELGSSNFPVVLVQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDSTDPTVKQMVEVECQRWASKGINI
RYQIRENRVGYKAGALKEGLKRSYVKHCEYVVIFDADFQPEPDFLRRSIPFLMHNPNIALVQARWRFVNSDECLLTRMQE
MSLDYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRTTVEDMDLAVRASLRGWKFLYLGDLQVKSELPSTFR
AFRFQQHRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFFFVRKIIAHWVTFCFYCVVLPLTILVPEVKVPIWGSVY
IPSIITILNSVGTPRSIHLLFYWILFENVMSLHRTKATLIGLFEAGRANEWVVTAKLGSGQSAKGNTKGIKRFPRIFKLP
DRLNTLELGFAAFLFVCGCYDFVHGKNNYFIYLFLQTMSFFISGLGWIGTYVPS*
```

OPTIONS

Search against database ?: CDD v3.08 – 43334 PSSMs

Expect Value ? threshold: 0.01

Apply low-complexity filter ? ☑

Force live search ? □

Maximum number of hits ? 500

Result mode ● Concise ? ○ Full ?

Retrieve previous CD-search result

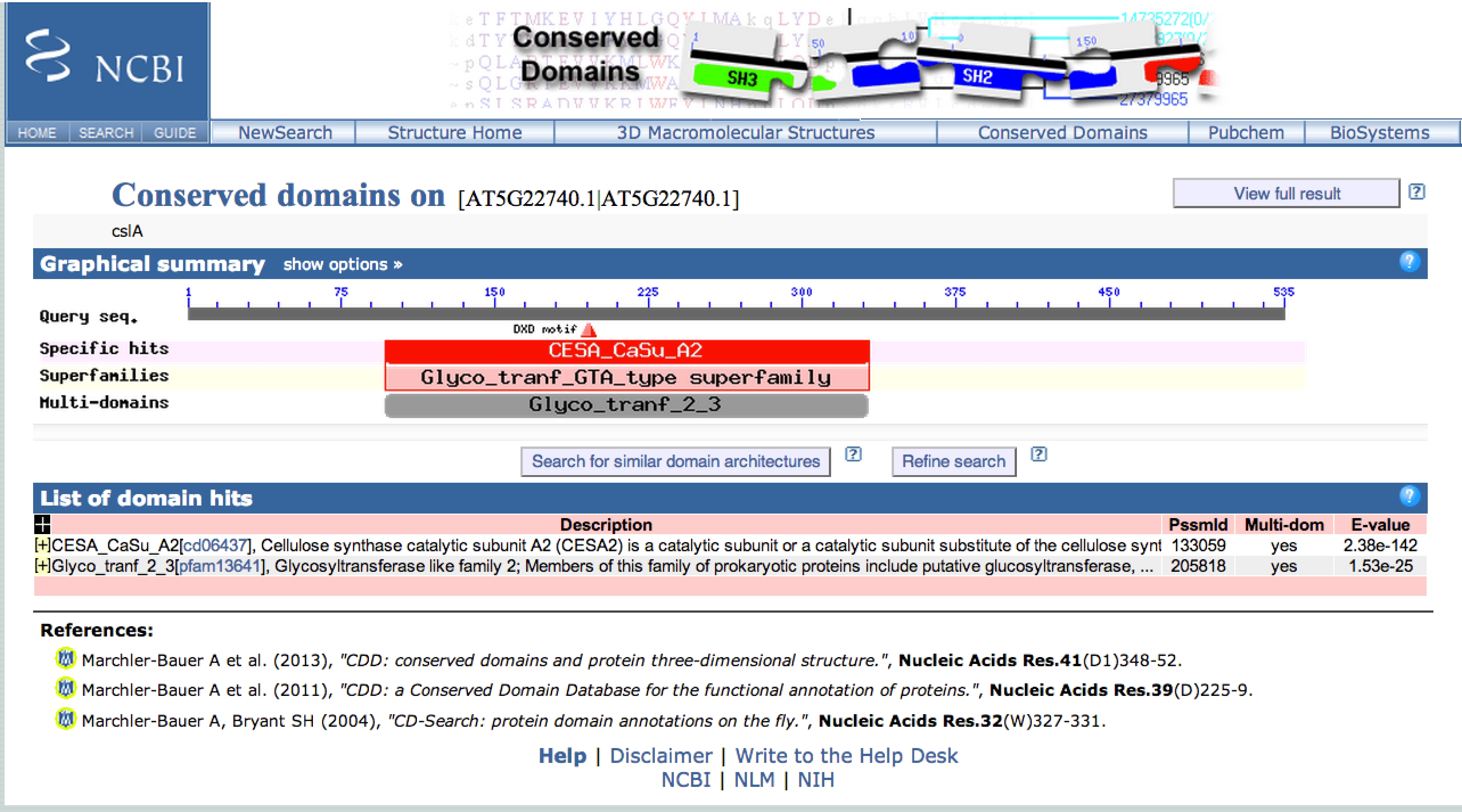
Request ID:

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.



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

cslA

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

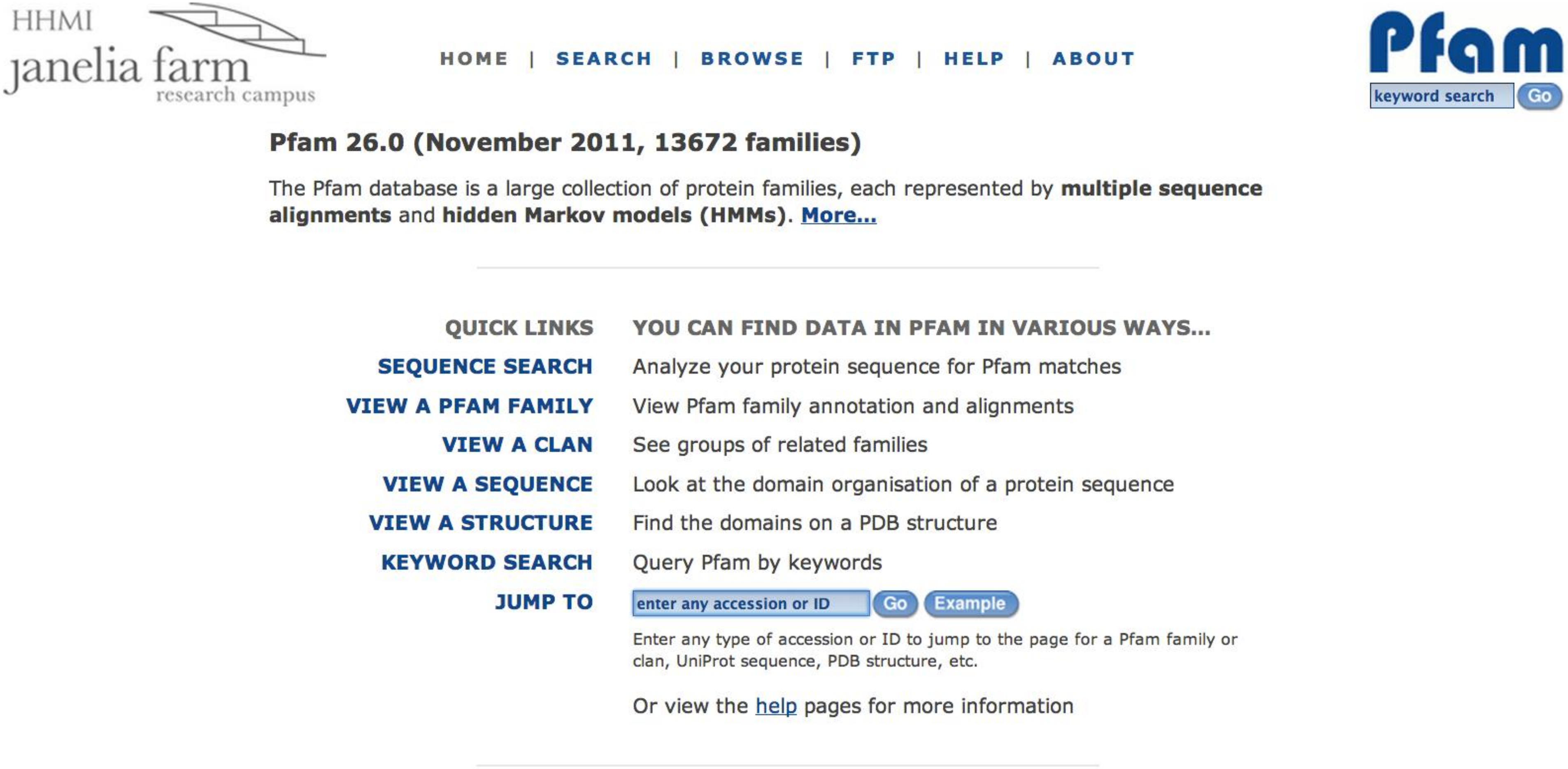
References:

- 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

Google “Pfam”

You will see two pfam sites:
Sanger pfam and Janellia pfam



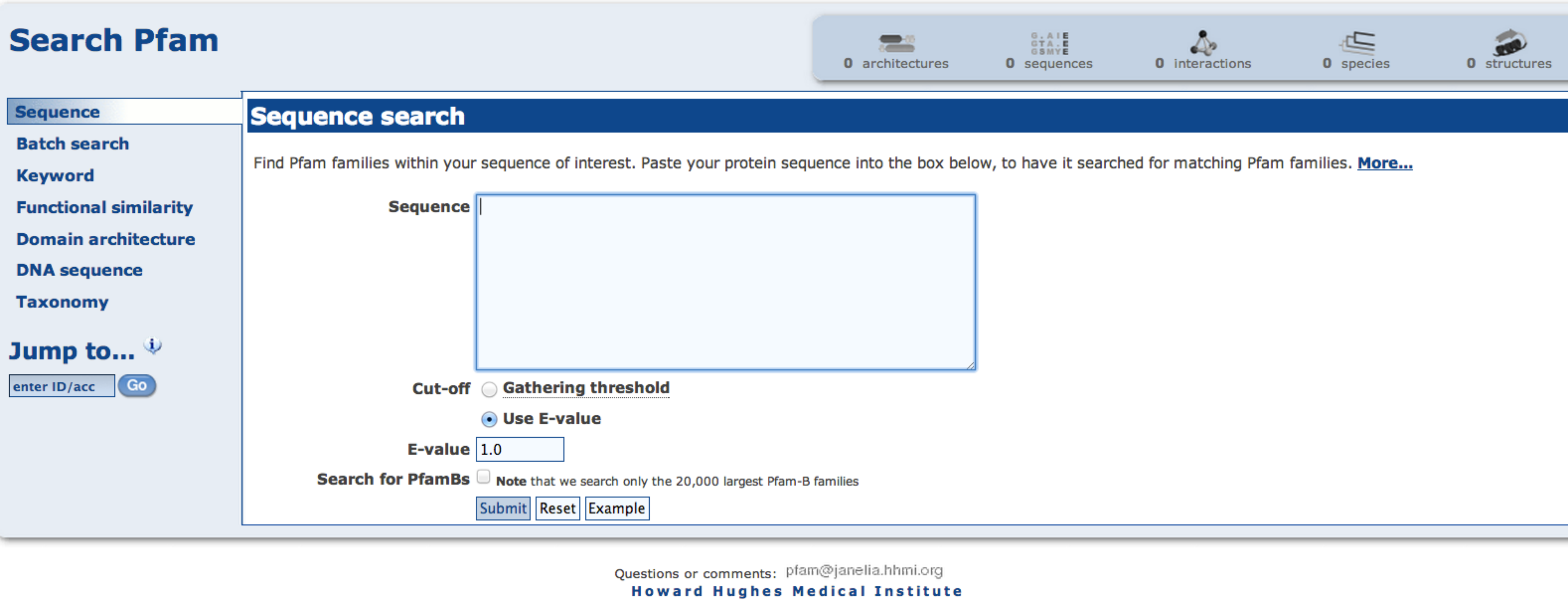
Pfam 26.0 (November 2011, 13672 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information



Search Pfam

Sequence search

Find Pfam families within your sequence of interest. Paste your protein sequence into the box below, to have it searched for matching Pfam families. More...

Sequence

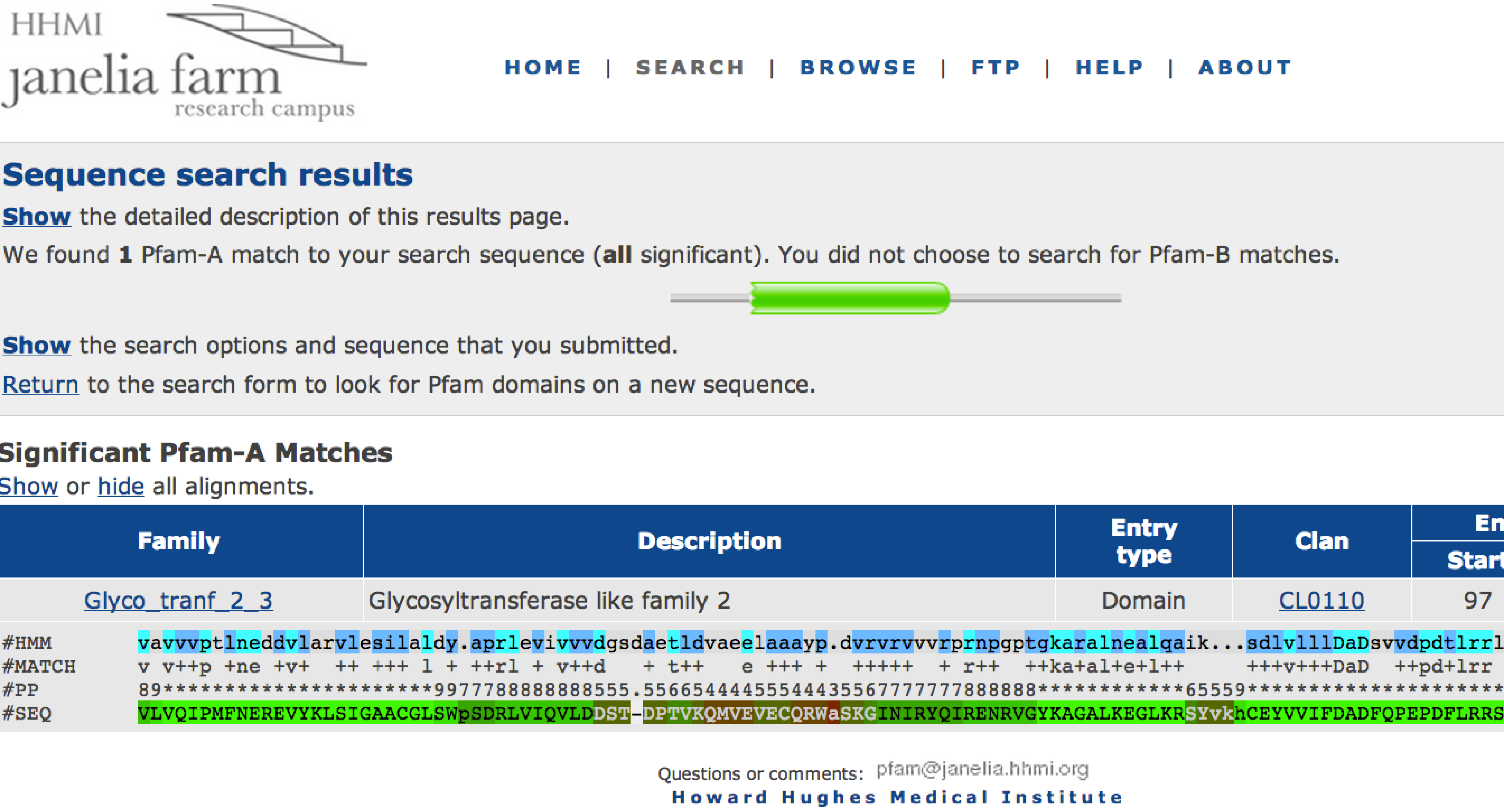
Cut-off ○ Gathering threshold

● Use E-value

E-value 1.0

Search for PfamBs □ **Note** that we search only the 20,000 largest Pfam-B families

Questions or comments: pfam@janelia.hhmi.org
Howard Hughes Medical Institute



HHMI janelia farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Sequence search results

Show the detailed description of this results page.

We found **1** Pfam-A match to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

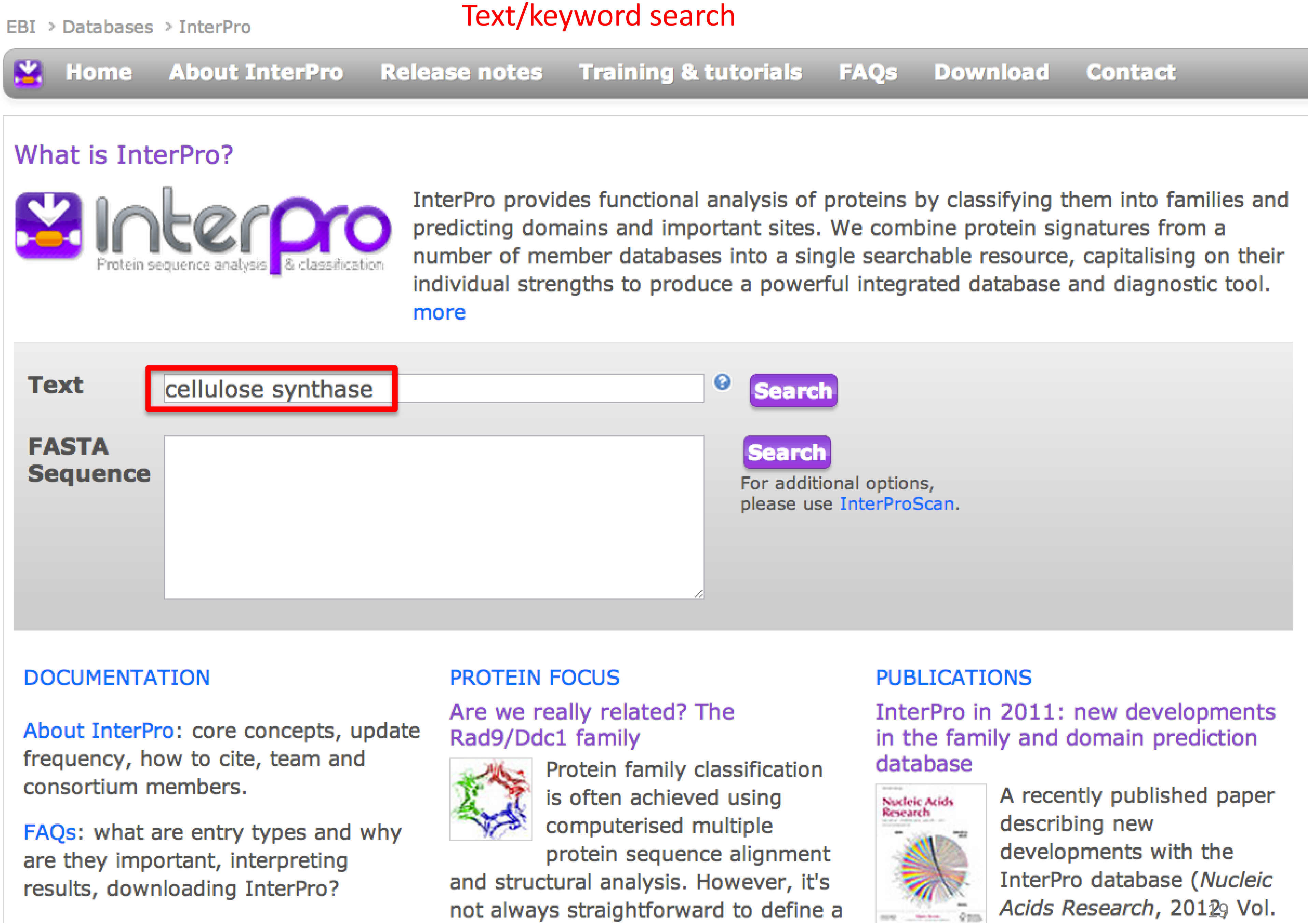
Family	Description	Entry type	Clan	En
				Start
Glyco_tranf_2_3	Glycosyltransferase like family 2	Domain	CL0110	97

```
#HMM    vavvvptlneddvlarvlesilaldy.aprlevivvvdgsdaetldvaeelaaayp.dvrvrvvvrprnpgptgkaralnealqaik...sdlvlllDaDsvvdpdtlrrl
#MATCH  v v++p +ne +v+  ++ +++ l + ++rl + v++d   + t++   e +++ +  +++++  + r++  ++ka+al+e+l++     +++v+++DaD  ++pd+lrr
#PP     89*********************9977788888888555.556654444555444355677777777888888************65559*********************
#SEQ    VLVQIPMFNEREVYKLSIGAACGLSWpSDRLVIQVLDDST-DPTVKQMVEVECQRWaSKGINIRYQIRENRVGYKAGALKEGLKRSYvkhCEYVVIFDADFQPEPDFLRRS
```

Questions or comments: pfam@janelia.hhmi.org

Howard Hughes Medical Institute

Text/keyword search



What is InterPro?

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. more

Text cellulose synthase

FASTA Sequence

For additional options, please use InterProScan.

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?

PROTEIN FOCUS

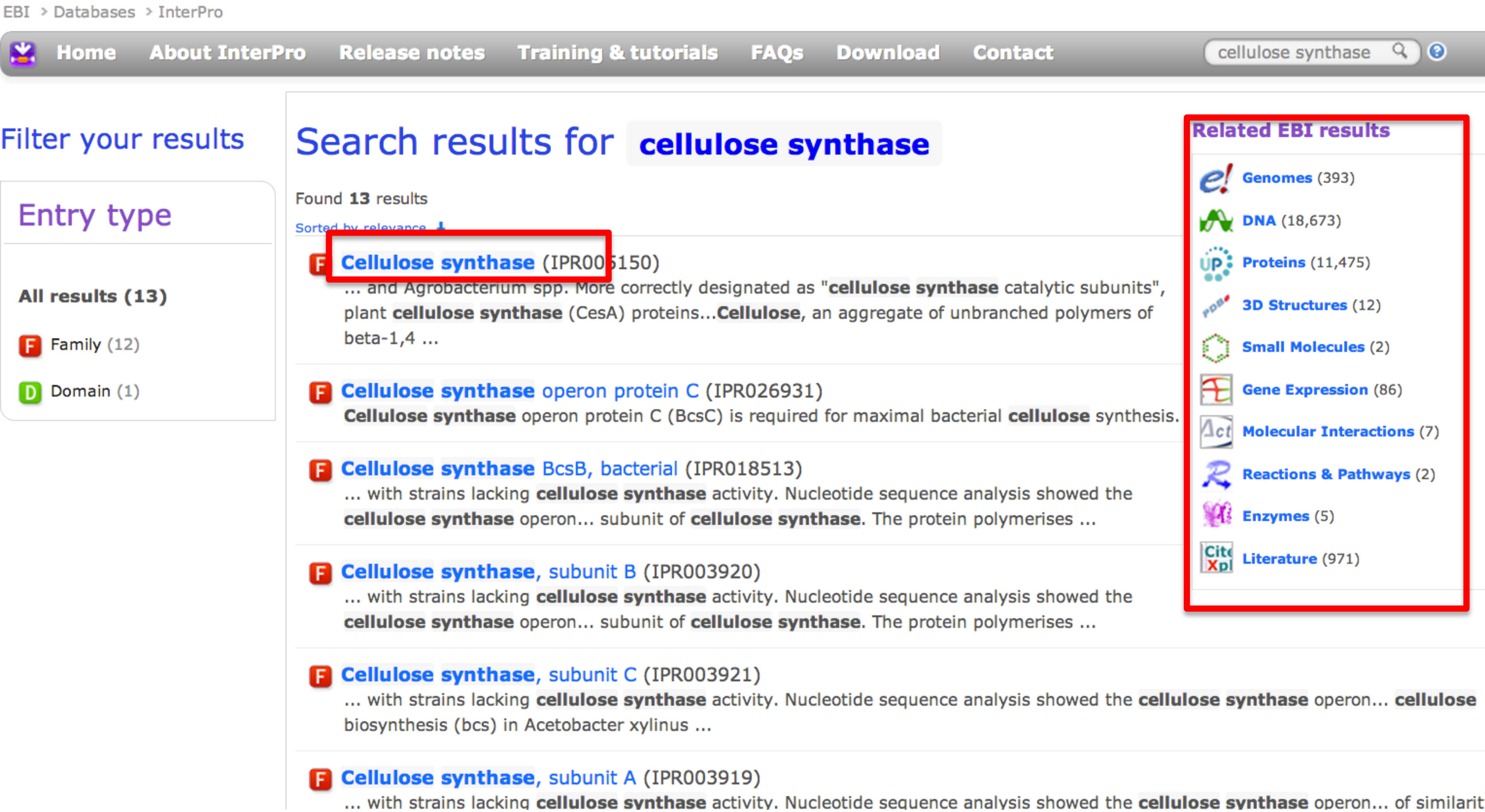
Are we really related? The Rad9/Ddc1 family

Protein family classification is often achieved using computerised multiple protein sequence alignment and structural analysis. However, it's not always straightforward to define a

PUBLICATIONS

InterPro in 2011: new developments in the family and domain prediction database

A recently published paper describing new developments with the InterPro database (*Nucleic Acids Research*, 2012, Vol.



Filter your results

Entry type

All results (13)

F Family (12)

D Domain (1)

Search results for **cellulose synthase**

Found **13** results

Sorted by relevance

F **Cellulose synthase** (IPR005150)
... and Agrobacterium spp. More correctly designated as "**cellulose synthase** catalytic subunits", plant **cellulose synthase** (CesA) proteins...**Cellulose**, an aggregate of unbranched polymers of beta-1,4 ...

F **Cellulose synthase** operon protein C (IPR026931)
Cellulose synthase operon protein C (BcsC) is required for maximal bacterial **cellulose** synthesis.

F **Cellulose synthase** BcsB, bacterial (IPR018513)
... with strains lacking **cellulose synthase** activity. Nucleotide sequence analysis showed the **cellulose synthase** operon... subunit of **cellulose synthase**. The protein polymerises ...

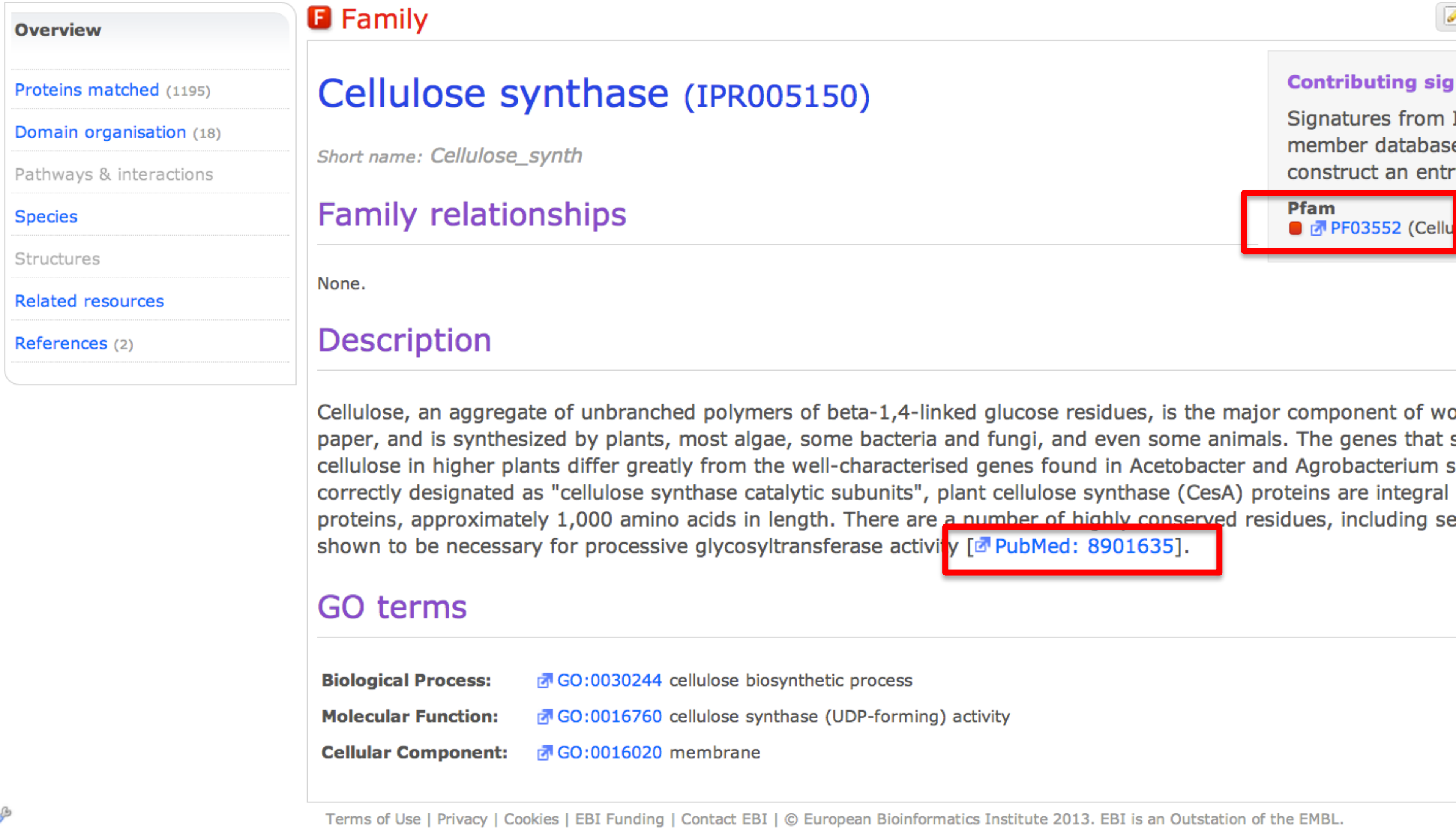
F **Cellulose synthase**, subunit B (IPR003920)
... with strains lacking **cellulose synthase** activity. Nucleotide sequence analysis showed the **cellulose synthase** operon... subunit of **cellulose synthase**. The protein polymerises ...

F **Cellulose synthase**, subunit C (IPR003921)
... with strains lacking **cellulose synthase** activity. Nucleotide sequence analysis showed the **cellulose synthase** operon... **cellulose** biosynthesis (bcs) in Acetobacter xylinus ...

F **Cellulose synthase**, subunit A (IPR003919)
... with strains lacking **cellulose synthase** activity. Nucleotide sequence analysis showed the **cellulose synthase** operon... of similarit

Related EBI results

Genomes	(393)
DNA	(18,673)
Proteins	(11,475)
3D Structures	(12)
Small Molecules	(2)
Gene Expression	(86)
Molecular Interactions	(7)
Reactions & Pathways	(2)
Enzymes	(5)
Literature	(971)



F Family

Cellulose synthase (IPR005150)

Short name: Cellulose_synth

Family relationships

None.

Description

Cellulose, an aggregate of unbranched polymers of beta-1,4-linked glucose residues, is the major component of wo paper, and is synthesized by plants, most algae, some bacteria and fungi, and even some animals. The genes that s cellulose in higher plants differ greatly from the well-characterised genes found in Acetobacter and Agrobacterium s correctly designated as "cellulose synthase catalytic subunits", plant cellulose synthase (CesA) proteins are integral proteins, approximately 1,000 amino acids in length. There are a number of highly conserved residues, including se shown to be necessary for processive glycosyltransferase activity [PubMed: 8901635].

GO terms

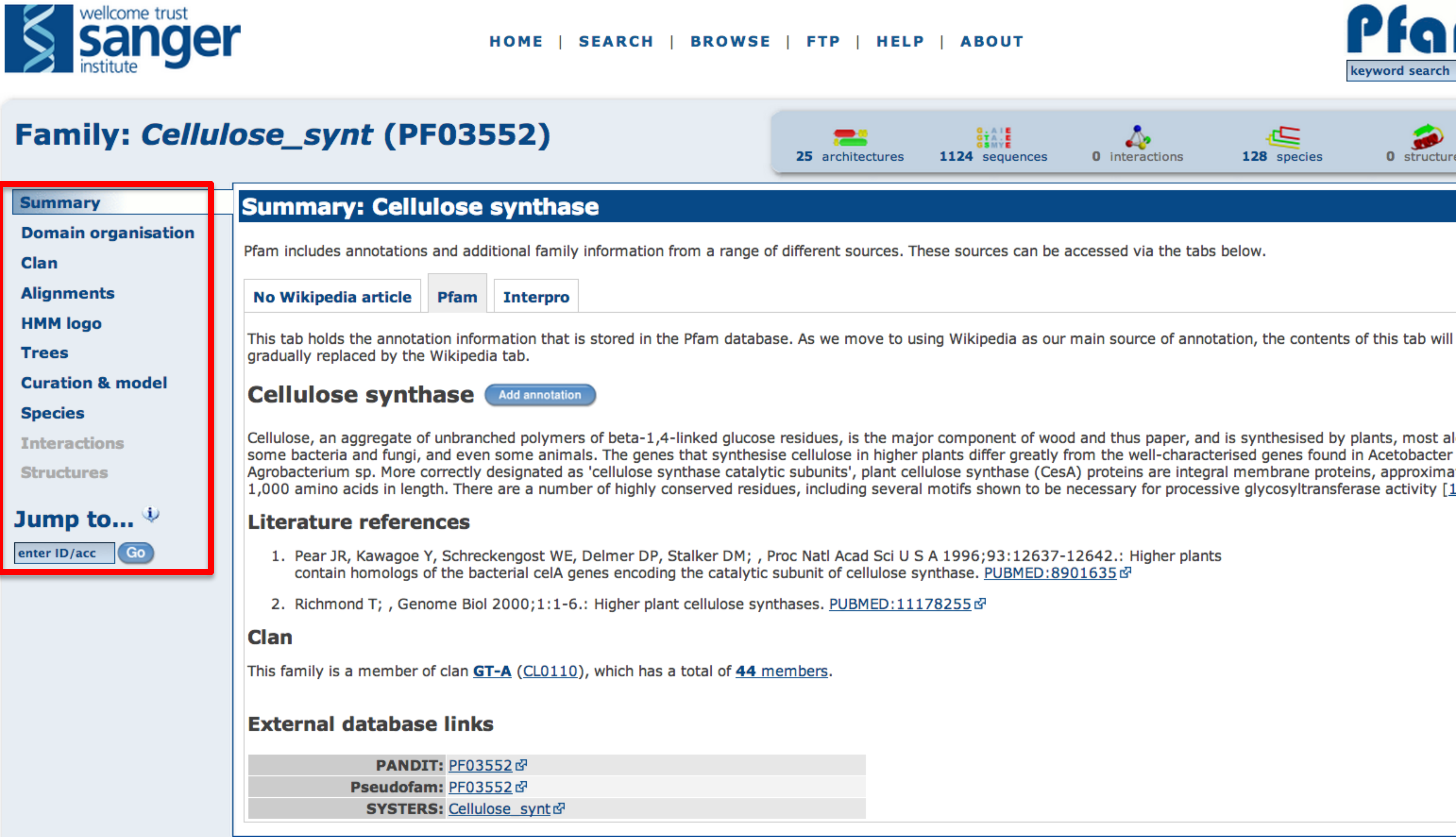
Biological Process:	GO:0030244 cellulose biosynthetic process
Molecular Function:	GO:0016760 cellulose synthase (UDP-forming) activity
Cellular Component:	GO:0016020 membrane

Contributing sig

Signatures from I member database construct an entr

Pfam
🟥 PF03552 (Cellu

Terms of Use | Privacy | Cookies | EBI Funding | Contact EBI | © European Bioinformatics Institute 2013. EBI is an Outstation of the EMBL.



Family: *Cellulose_synt* (PF03552)

Summary: Cellulose synthase

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will gradually replaced by the Wikipedia tab.

Cellulose synthase

Cellulose, an aggregate of unbranched polymers of beta-1,4-linked glucose residues, is the major component of wood and thus paper, and is synthesised by plants, most al some bacteria and fungi, and even some animals. The genes that synthesise cellulose in higher plants differ greatly from the well-characterised genes found in Acetobacter Agrobacterium sp. More correctly designated as 'cellulose synthase catalytic subunits', plant cellulose synthase (CesA) proteins are integral membrane proteins, approxima 1,000 amino acids in length. There are a number of highly conserved residues, including several motifs shown to be necessary for processive glycosyltransferase activity [1

Literature references

1. Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM; , Proc Natl Acad Sci U S A 1996;93:12637-12642.: Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. PUBMED:8901635
2. Richmond T; , Genome Biol 2000;1:1-6.: Higher plant cellulose synthases. PUBMED:11178255

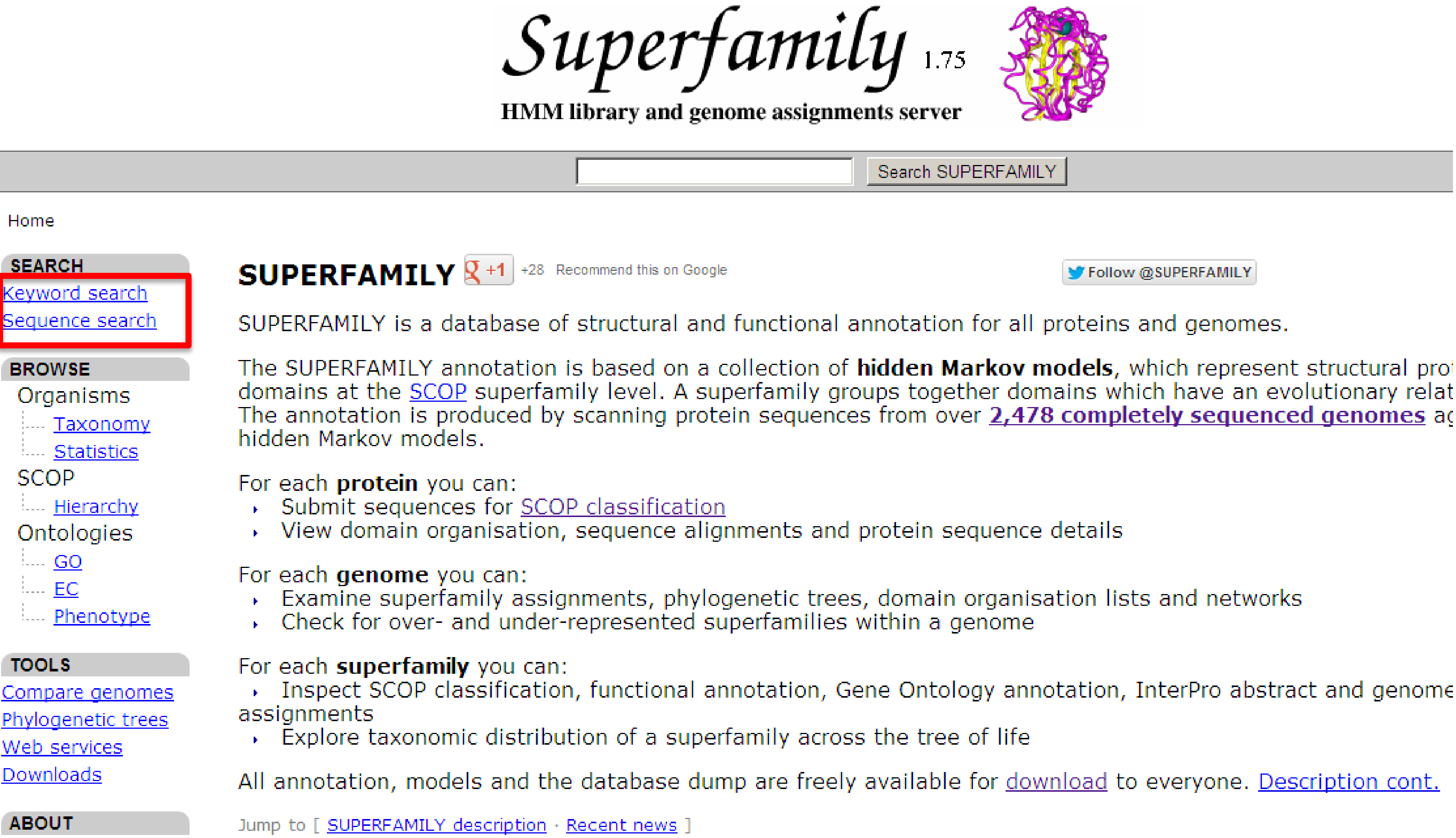
Clan

This family is a member of clan **GT-A** (CL0110), which has a total of **44** members.

External database links

PANDIT:	PF03552
Pseudofam:	PF03552
SYSTERS:	Cellulose_synt

http://supfam.cs.bris.ac.uk/SUPERFAMILY/



SUPERFAMILY

SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes.

The SUPERFAMILY annotation is based on a collection of **hidden Markov models**, which represent structural pro domains at the SCOP superfamily level. A superfamily groups together domains which have an evolutionary relat The annotation is produced by scanning protein sequences from over **2,478 completely sequenced genomes** ag hidden Markov models.

For each **protein** you can:

- Submit sequences for SCOP classification
- View domain organisation, sequence alignments and protein sequence details

For each **genome** you can:

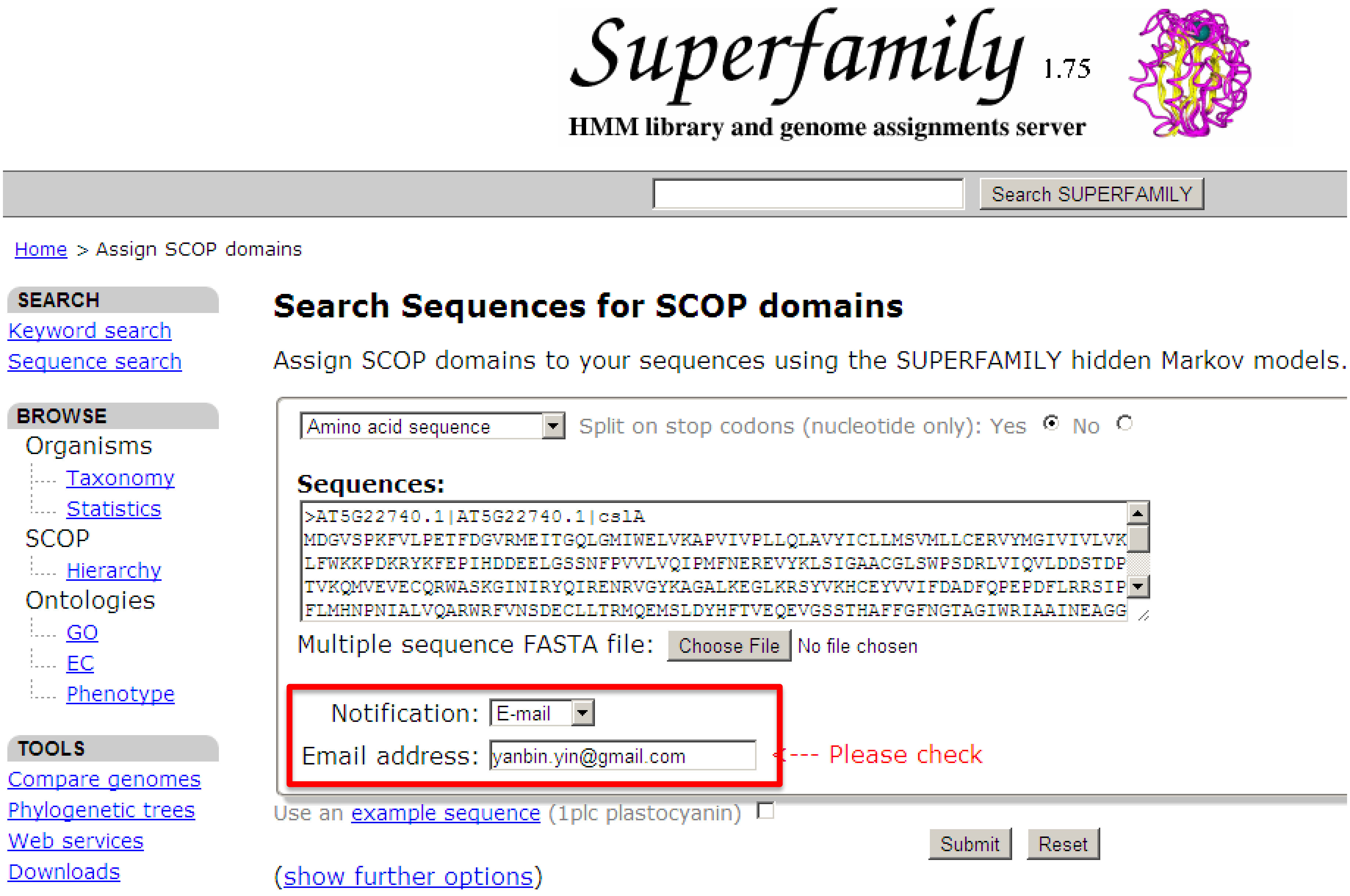
- Examine superfamily assignments, phylogenetic trees, domain organisation lists and networks
- Check for over- and under-represented superfamilies within a genome

For each **superfamily** you can:

- Inspect SCOP classification, functional annotation, Gene Ontology annotation, InterPro abstract and genome assignments
- Explore taxonomic distribution of a superfamily across the tree of life

All annotation, models and the database dump are freely available for download to everyone. Description cont.

Jump to [SUPERFAMILY description · Recent news]



Superfamily 1.75

HMM library and genome assignments server

Home > Assign SCOP domains

Search Sequences for SCOP domains

Assign SCOP domains to your sequences using the SUPERFAMILY hidden Markov models.

Amino acid sequence ▾ Split on stop codons (nucleotide only): Yes ● No ○

Sequences:

```
>AT5G22740.1|AT5G22740.1|cslA
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVLVK
LFWKKPDKRYKFEPIHDDEELGSSNFPVVLVQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDSTDP
TVKQMVEVECQRWASKGINIRYQIRENRVGYKAGALKEGLKRSYVKHCEYVVIFDADFQPEPDFLRRSIP
FLMHNPNIALVQARWRFVNSDECLLTRMQEMSLDYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGG
```

Multiple sequence FASTA file: Choose File No file chosen

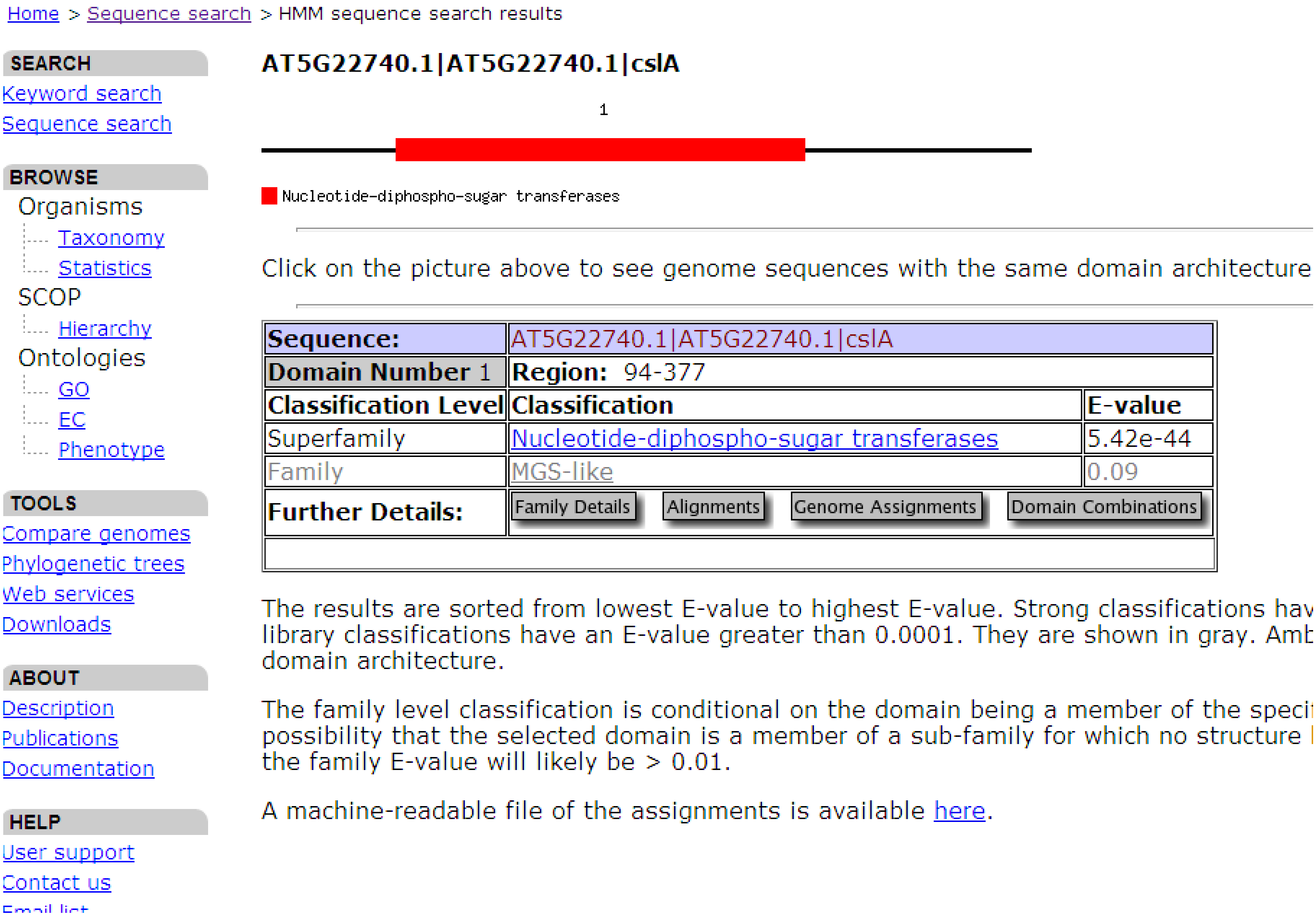
Notification: E-mail ▾

Email address: yanbin.yin@gmail.com

<--- Please check

Use an example sequence (1plc plastocyanin) □

(show further options)



Home > Sequence search > HMM sequence search results

SEARCH

Keyword search

Sequence search

BROWSE

Organisms

- Taxonomy
- Statistics

SCOP

- Hierarchy

Ontologies

- GO
- EC
- Phenotype

TOOLS

Compare genomes

Phylogenetic trees

Web services

Downloads

ABOUT

Description

Publications

Documentation

HELP

User support

Contact us

Email list

AT5G22740.1|AT5G22740.1|cslA

🟥 Nucleotide-diphospho-sugar transferases

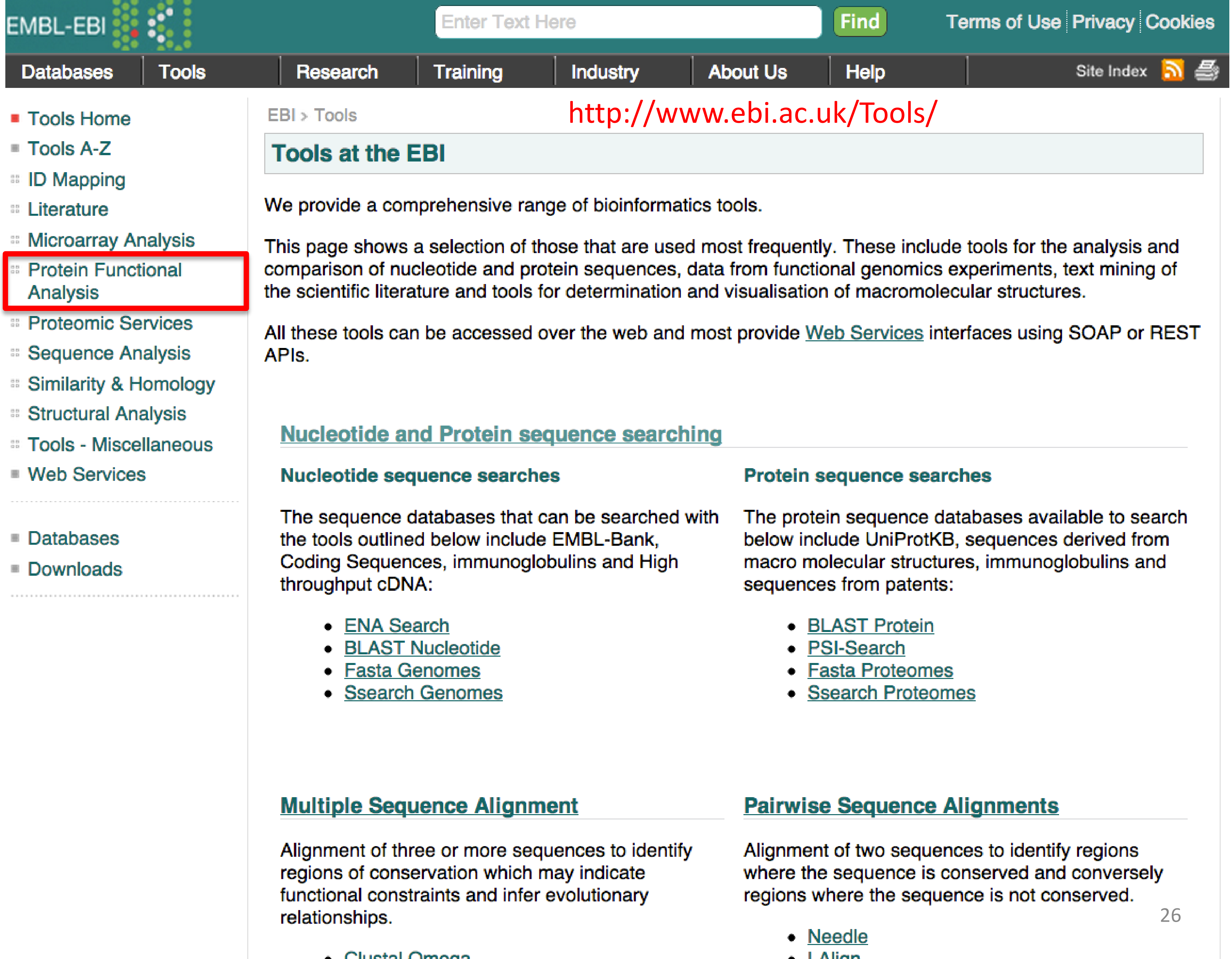
Click on the picture above to see genome sequences with the same domain architecture

Sequence:	AT5G22740.1|AT5G22740.1|cslA	
Domain Number 1	**Region:** 94-377	
Classification Level	**Classification**	**E-value**
Superfamily	Nucleotide-diphospho-sugar transferases	5.42e-44
Family	MGS-like	0.09
Further Details:	Family Details Alignments Genome Assignments Domain Combinations	

The results are sorted from lowest E-value to highest E-value. Strong classifications hav library classifications have an E-value greater than 0.0001. They are shown in gray. Amb domain architecture.

The family level classification is conditional on the domain being a member of the speci possibility that the selected domain is a member of a sub-family for which no structure the family E-value will likely be > 0.01.

A machine-readable file of the assignments is available here.



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

Tools at the EBI

We provide a comprehensive range of bioinformatics tools.

This page shows a selection of those that are used most frequently. These include tools for the analysis and comparison of nucleotide and protein sequences, data from functional genomics experiments, text mining of the scientific literature and tools for determination and visualisation of macromolecular structures.

All these tools can be accessed over the web and most provide Web Services interfaces using SOAP or REST APIs.

Nucleotide and Protein sequence searching

Nucleotide sequence searches

The sequence databases that can be searched with the tools outlined below include EMBL-Bank, Coding Sequences, immunoglobulins and High throughput cDNA:

- ENA Search
- BLAST Nucleotide
- Fasta Genomes
- Ssearch Genomes

Protein sequence searches

The protein sequence databases available to search below include UniProtKB, sequences derived from macro molecular structures, immunoglobulins and sequences from patents:

- BLAST Protein
- PSI-Search
- Fasta Proteomes
- Ssearch Proteomes

Multiple Sequence Alignment

Alignment of three or more sequences to identify regions of conservation which may indicate functional constraints and infer evolutionary relationships.

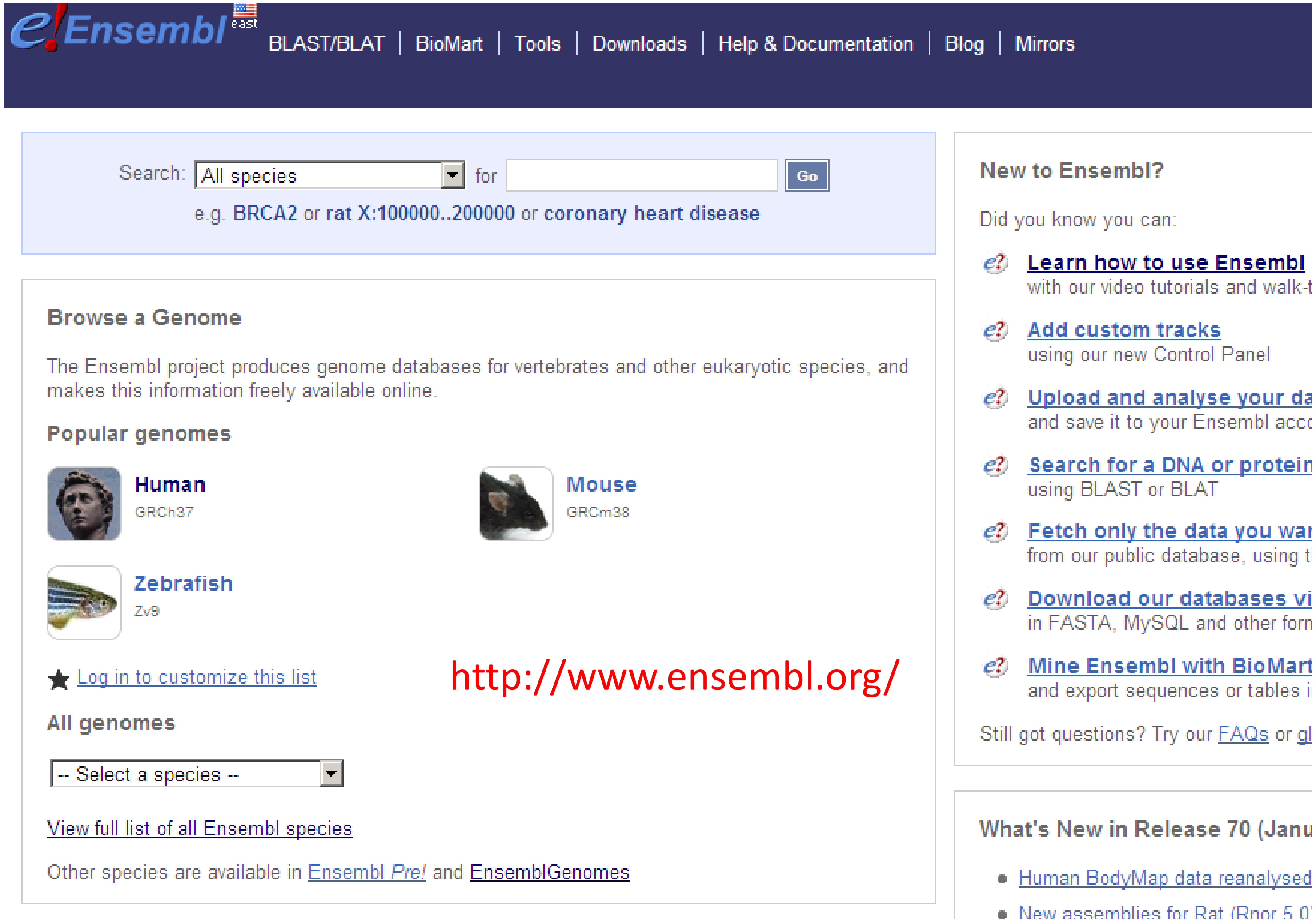
- Clustal Omega

Pairwise Sequence Alignments

Alignment of two sequences to identify regions where the sequence is conserved and conversely regions where the sequence is not conserved.

- Needle
- LAlign

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
Go

Search: All species for

e.g. BRCA2 or rat X:100000..200000 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 GRCh37

Mouse GRCm38

Zebrafish Zv9

★ Log in to customize this list

http://www.ensembl.org/

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

New to Ensembl?

Did you know you can:

- Learn how to use Ensembl with our video tutorials and walk-t
- Add custom tracks using our new Control Panel
- Upload and analyse your da and save it to your Ensembl acco
- Search for a DNA or protein using BLAST or BLAT
- Fetch only the data you wa from our public database, using t
- Download our databases vi in FASTA, MySQL and other forn
- Mine Ensembl with BioMart and export sequences or tables i

Still got questions? Try our FAQs or gl

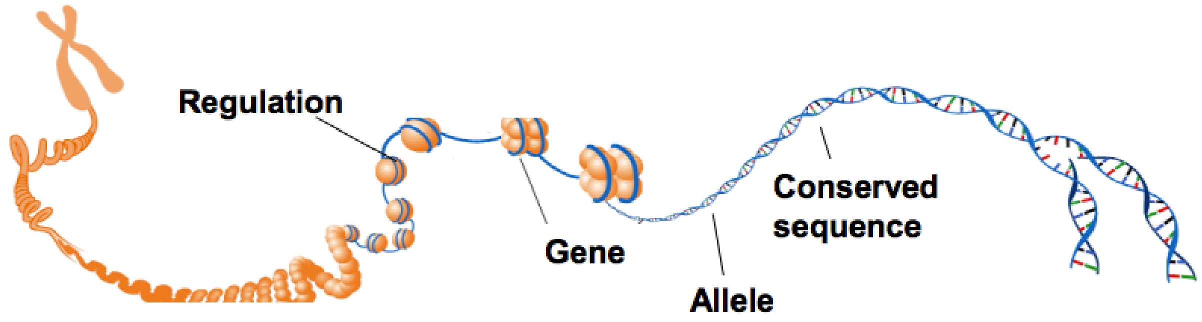
What's New in Release 70 (Janu

- Human BodyMap data reanalysed
- New assemblies for Rat (Rnor 5.0

What do we need genome browsers?

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

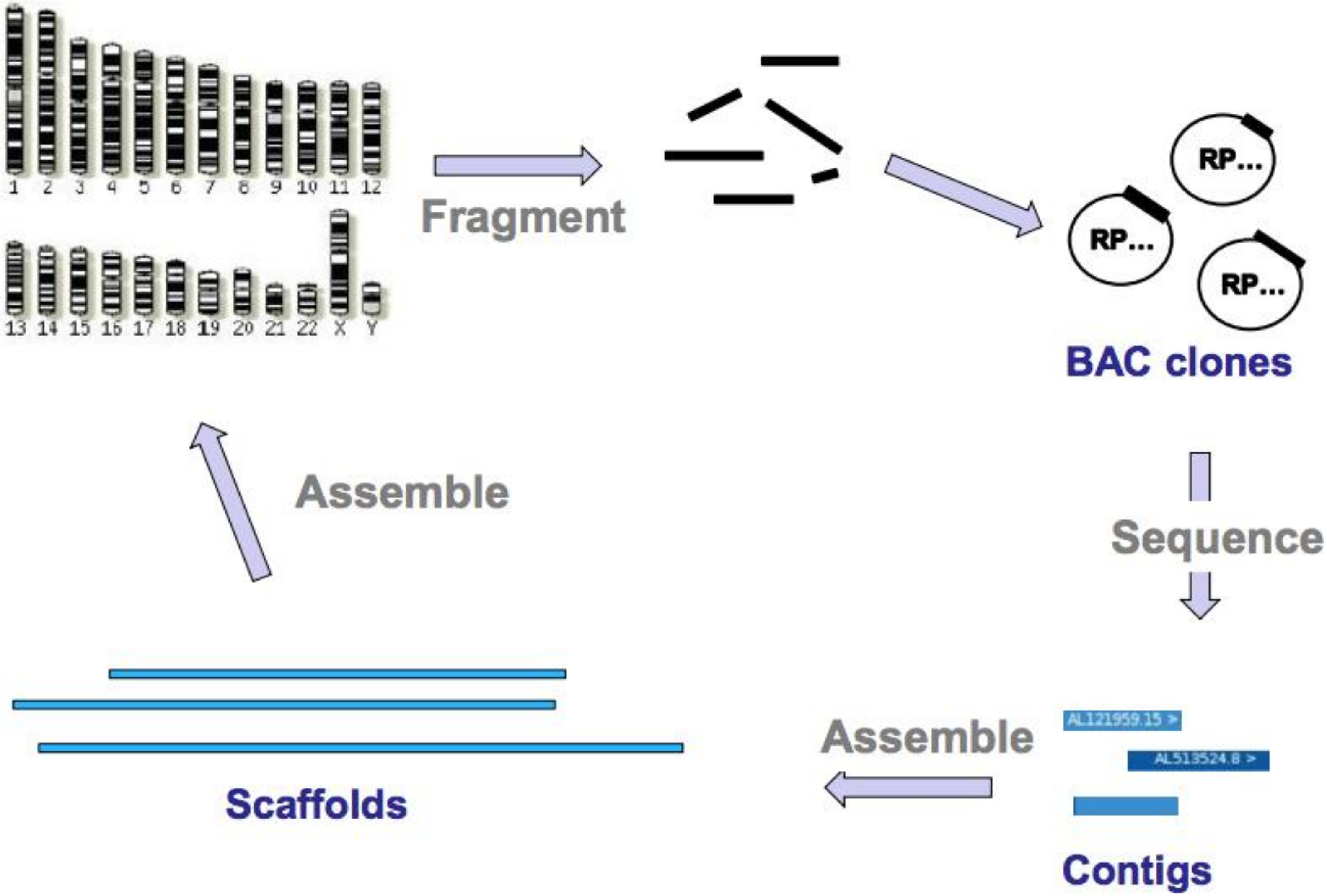
Genome browsers provide access to large amounts of sequence data via a graphical user interface. They use a visual, high-level overview of complex data in a form that can be grasped at a glance and provide the means to explore the data in increasing resolution from megabase scales down to the level of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

Genome Sequencing



Fragment

BAC clones

Sequence

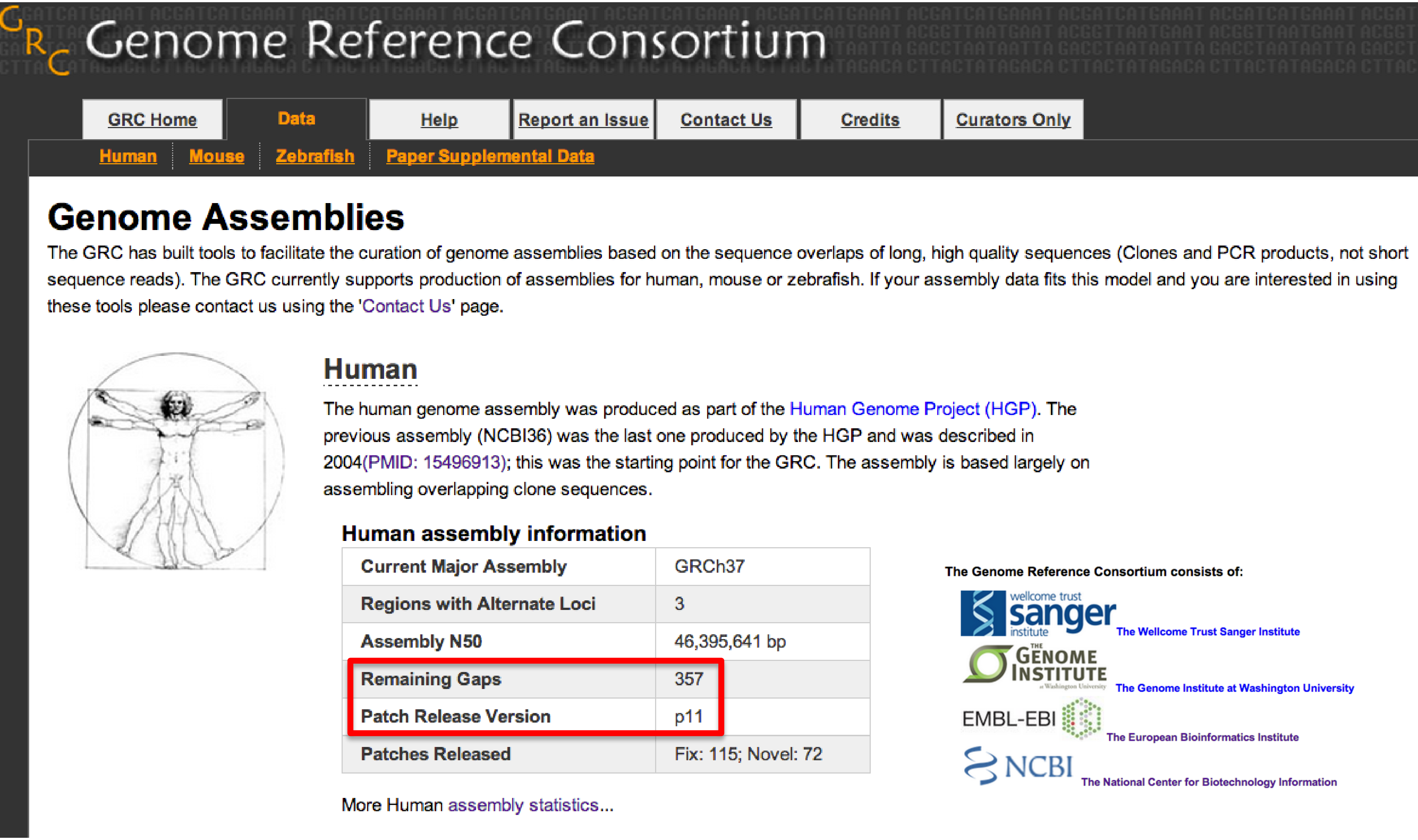
Contigs

Assemble

Scaffolds

Assemble

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 short sequence reads). The GRC 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 us using the 'Contact Us' page.

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	GRCh37
Regions with Alternate Loci	3
Assembly N50	46,395,641 bp
Remaining Gaps	357
Patch Release Version	p11
Patches Released	Fix: 115; Novel: 72

More Human assembly statistics...

The Genome Reference Consortium consists of:

wellcome trust sanger institute
THE GENOME INSTITUTE at Washington University
EMBL-EBI
NCBI

The Wellcome Trust Sanger Institute

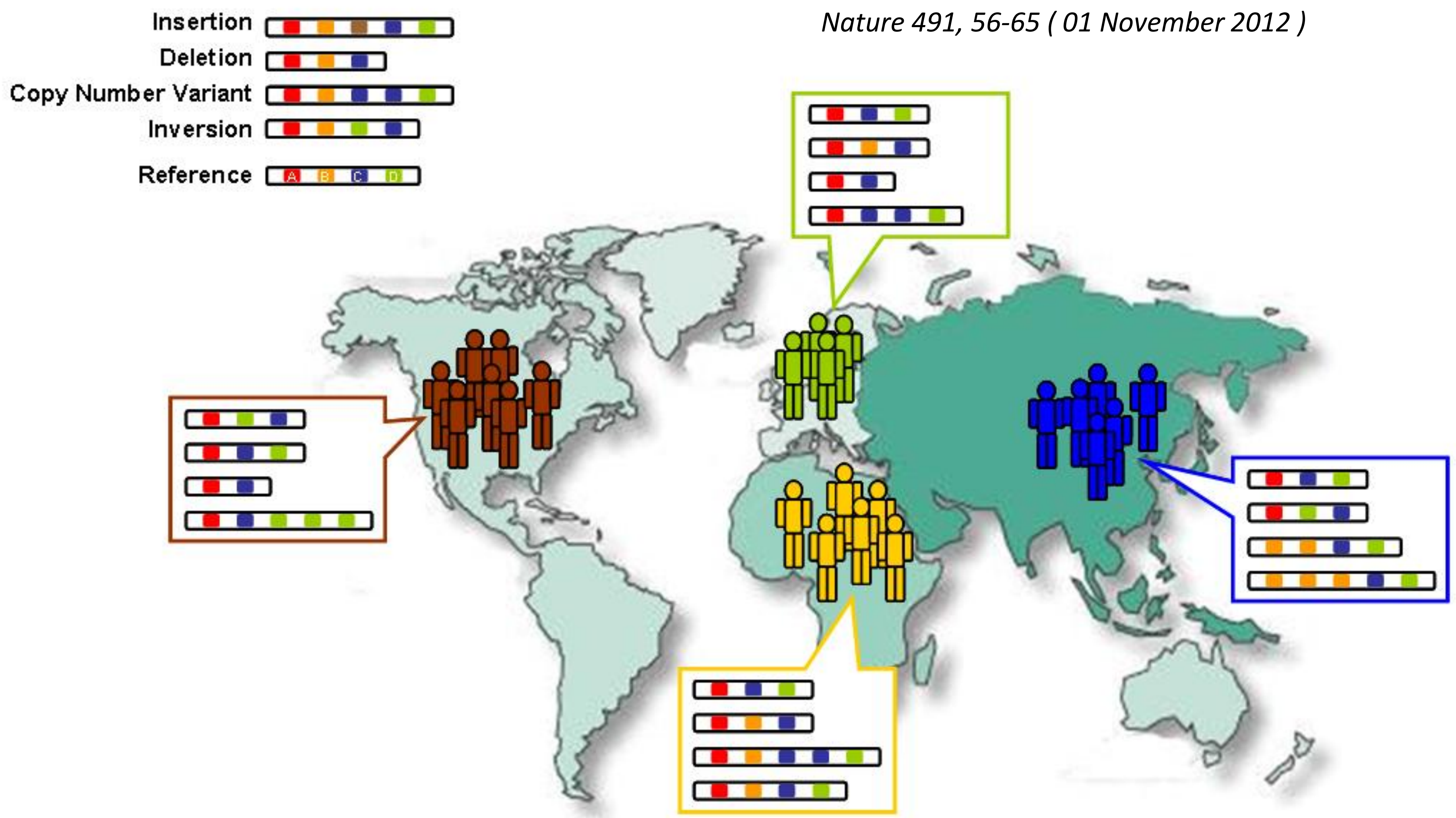
The Genome Institute at Washington University

The European Bioinformatics Institute

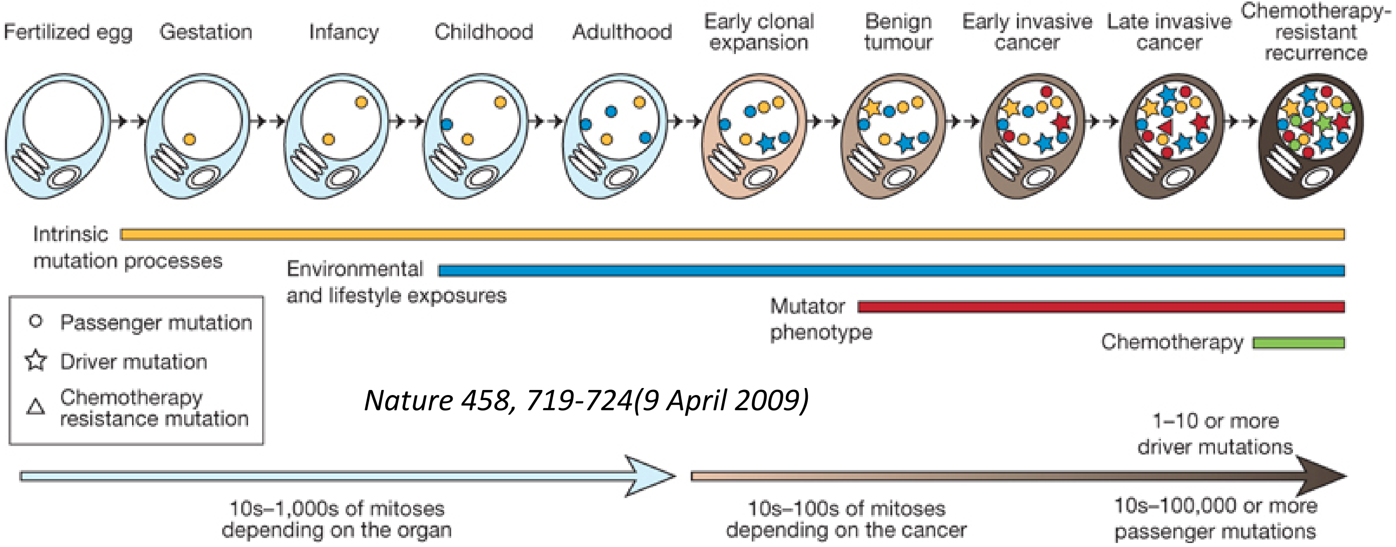
The National Center for Biotechnology Information

1000 Genomes

A Deep Catalog of Human Genetic Variation

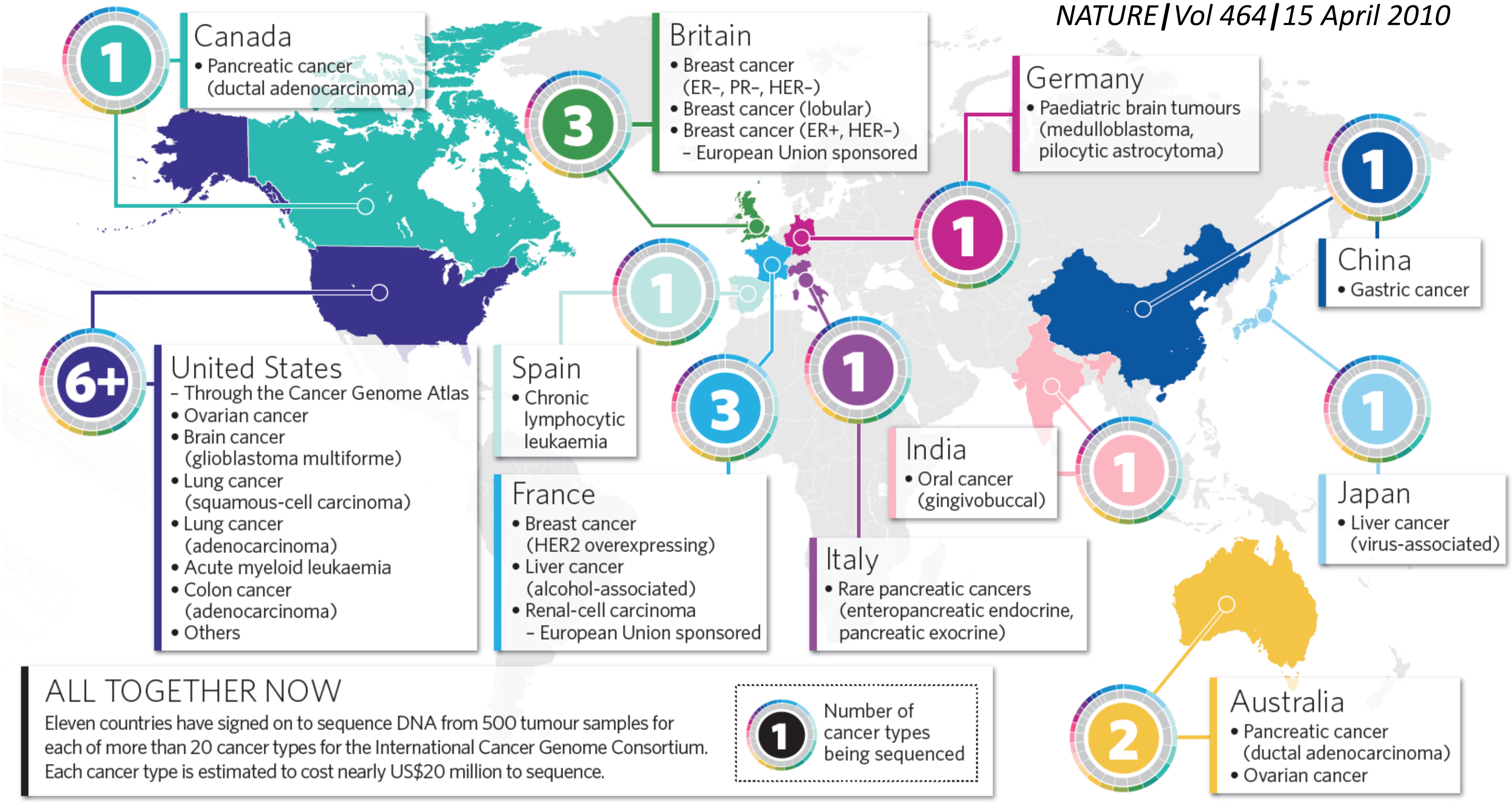


Nature 491, 56-65 (01 November 2012)


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.

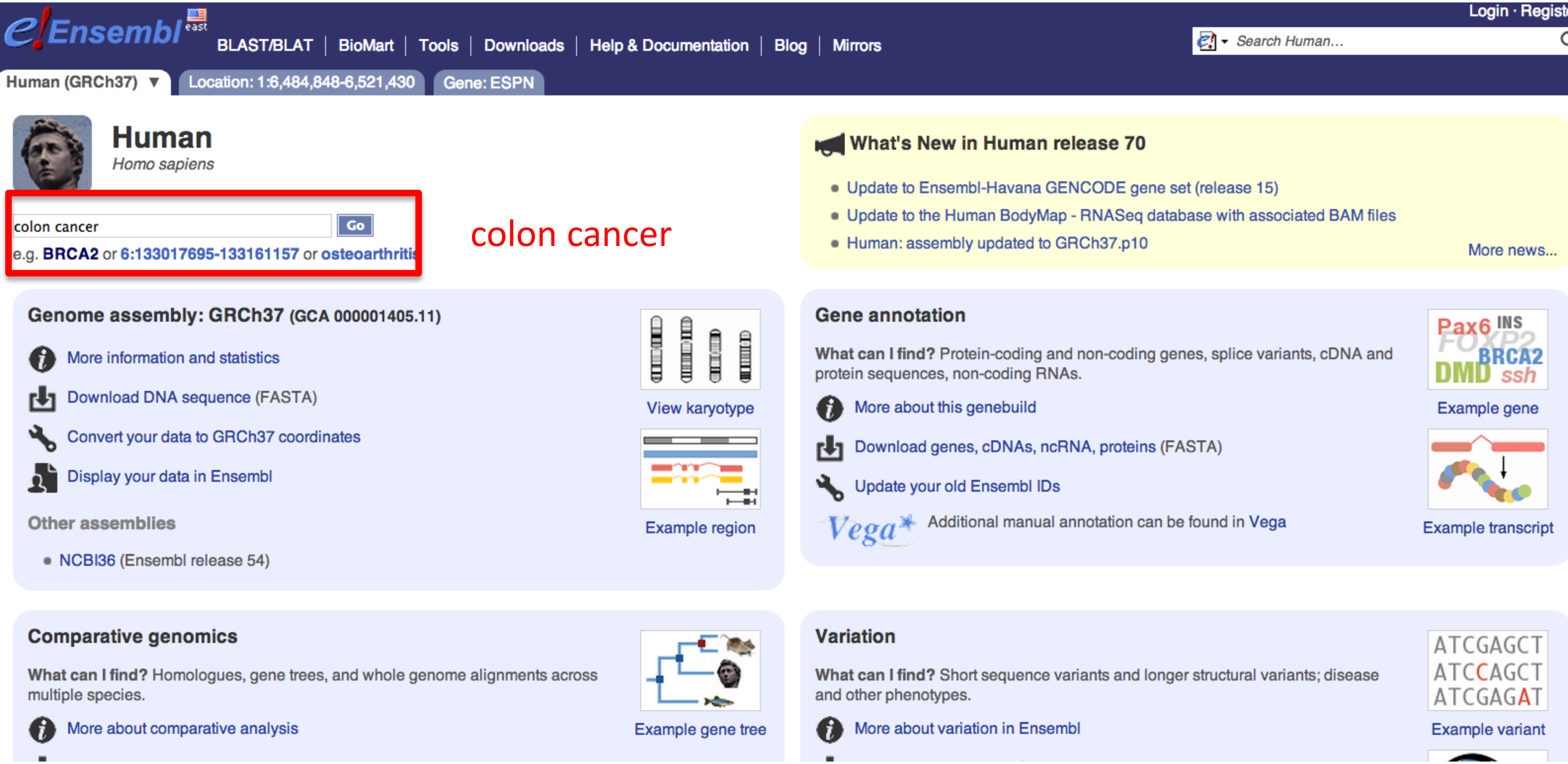
- Ensembl Genome Browsers: http://www.ensemblgenomes.org
- NCBI Map Viewer: http://www.ncbi.nlm.nih.gov/mapview/
- UCSC Genome Browser: http://genome.ucsc.edu

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/



Human

Homo sapiens

e.g. **BRCA2** or **6:133017695-133161157** or **osteoarthritis**

colon cancer

What's New in Human release 70

- Update to Ensembl-Havana GENCODE gene set (release 15)
- Update to the Human BodyMap - RNASeq database with associated BAM files
- Human: assembly updated to GRCh37.p10

More news...

Genome assembly: GRCh37 (GCA 000001405.11)

More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh37 coordinates

Display your data in Ensembl

Other assemblies

- NCBI36 (Ensembl release 54)

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Vega Additional manual annotation can be found in Vega

Comparative genomics

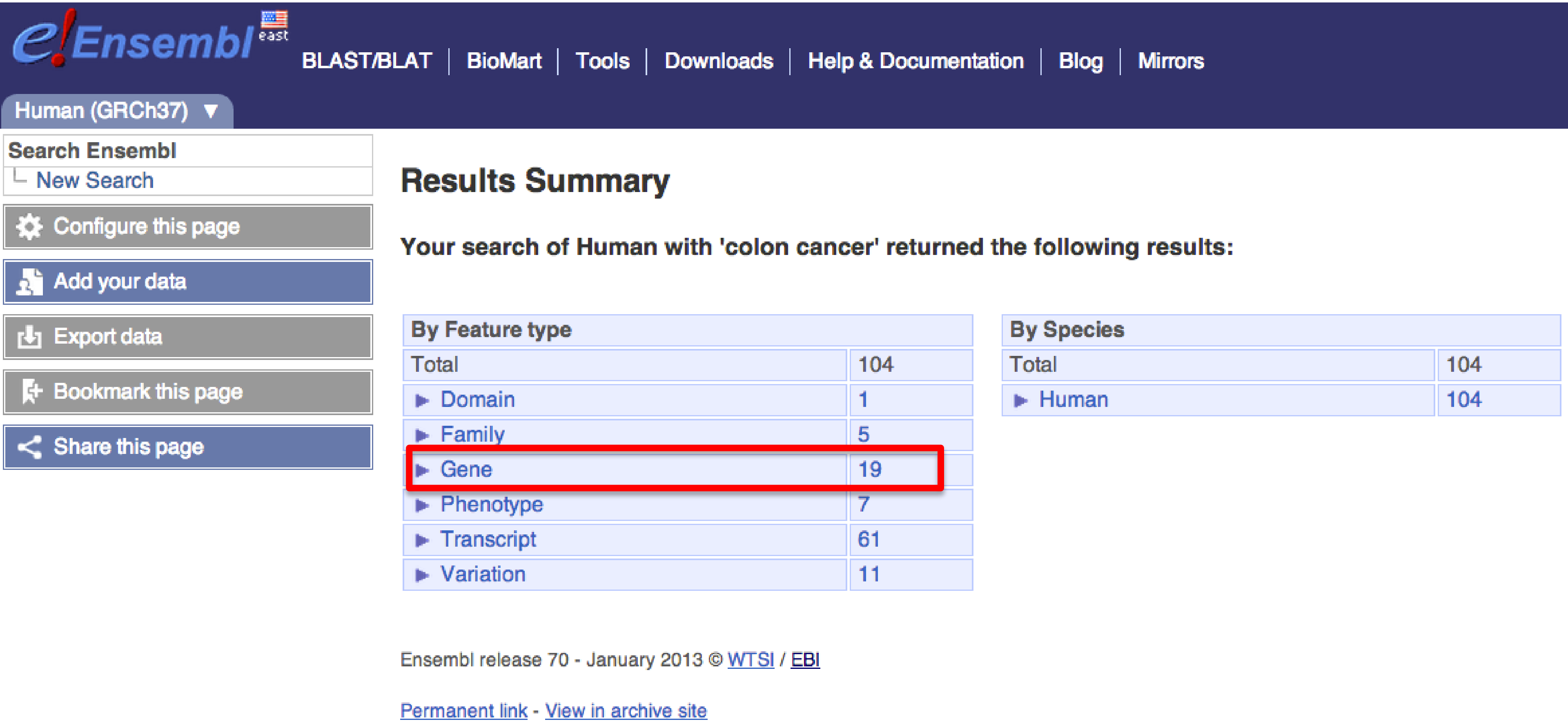
What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes.

More about variation in Ensembl



Results Summary

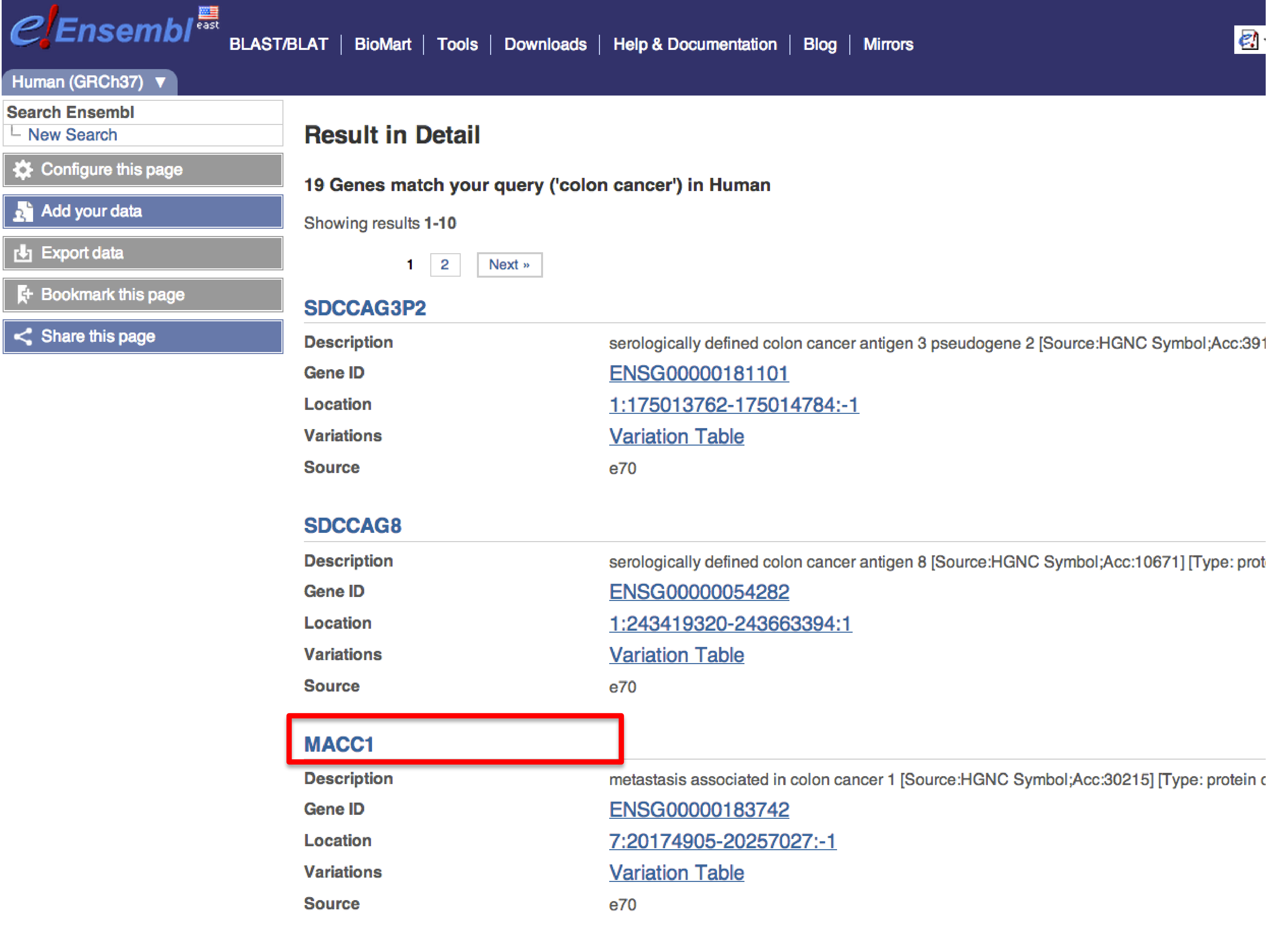
Your search of Human with 'colon cancer' returned the following results:

By Feature type	
Total	104
▶ Domain	1
▶ Family	5
▶ Gene	19
▶ Phenotype	7
▶ Transcript	61
▶ Variation	11

By Species	
Total	104
▶ Human	104

Ensembl release 70 - January 2013 © WTSI / EBI

Permanent link - View in archive site



Result in Detail

19 Genes match your query ('colon cancer') in Human

Showing results **1-10**

SDCCAG3P2

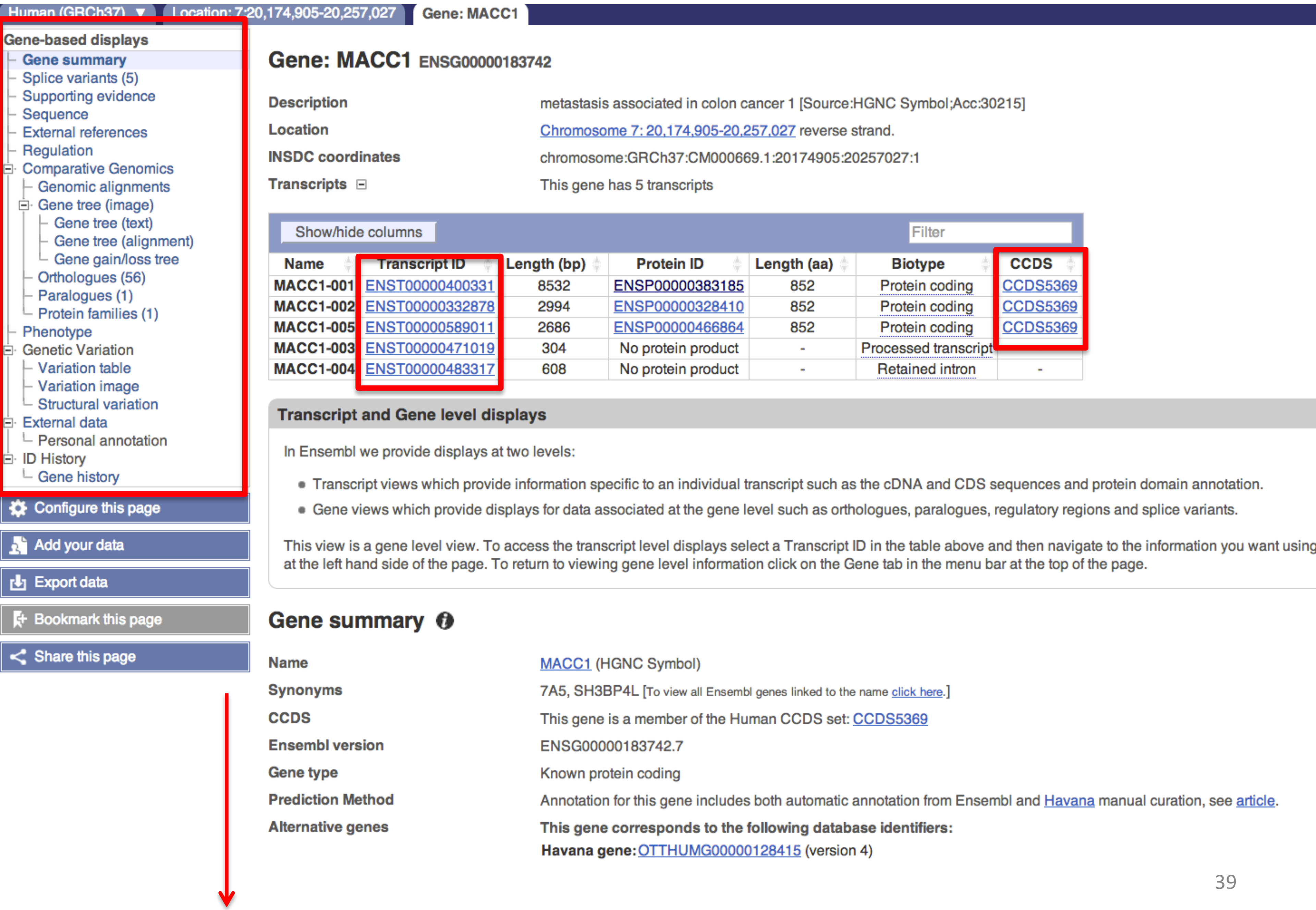
Description	serologically defined colon cancer antigen 3 pseudogene 2 [Source:HGNC Symbol;Acc:391
Gene ID	ENSG00000181101
Location	1:175013762-175014784:-1
Variations	Variation Table
Source	e70

SDCCAG8

Description	serologically defined colon cancer antigen 8 [Source:HGNC Symbol;Acc:10671] [Type: prot
Gene ID	ENSG00000054282
Location	1:243419320-243663394:1
Variations	Variation Table
Source	e70

MACC1

Description	metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215] [Type: protein c
Gene ID	ENSG00000183742
Location	7:20174905-20257027:-1
Variations	Variation Table
Source	e70



Gene: MACC1 ENSG00000183742

Description	metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]
Location	Chromosome 7: 20,174,905-20,257,027 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000669.1:20174905:20257027:1
Transcripts ⊟	This gene has 5 transcripts

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
MACC1-001	ENST00000400331	8532	ENSP00000383185	852	Protein coding	CCDS5369
MACC1-002	ENST00000332878	2994	ENSP00000328410	852	Protein coding	CCDS5369
MACC1-005	ENST00000589011	2686	ENSP00000466864	852	Protein coding	CCDS5369
MACC1-003	ENST00000471019	304	No protein product	-	Processed transcript	-
MACC1-004	ENST00000483317	608	No protein product	-	Retained intron	-

Transcript and Gene level displays

In Ensembl we provide displays at two levels:

- Transcript views which provide information specific to an individual transcript such as the cDNA and CDS sequences and protein domain annotation.
- Gene views which provide displays for data associated at the gene level such as orthologues, paralogues, regulatory regions and splice variants.

This view is a gene level view. To access the transcript level displays select a Transcript ID in the table above and then navigate to the information you want using at the left hand side of the page. To return to viewing gene level information click on the Gene tab in the menu bar at the top of the page.

Gene summary

Name	MACC1 (HGNC Symbol)
Synonyms	7A5, SH3BP4L [To view all Ensembl genes linked to the name click here.]
CCDS	This gene is a member of the Human CCDS set: CCDS5369
Ensembl version	ENSG00000183742.7
Gene type	Known protein coding
Prediction Method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.
Alternative genes	**This gene corresponds to the following database identifiers:** **Havana gene:** OTTHUMG00000128415 (version 4)

A consensus set of protein coding sequences

Consensus CDS protein set

CCDS Database

CCDS

EBI • NCBI • UCSC • WTSI

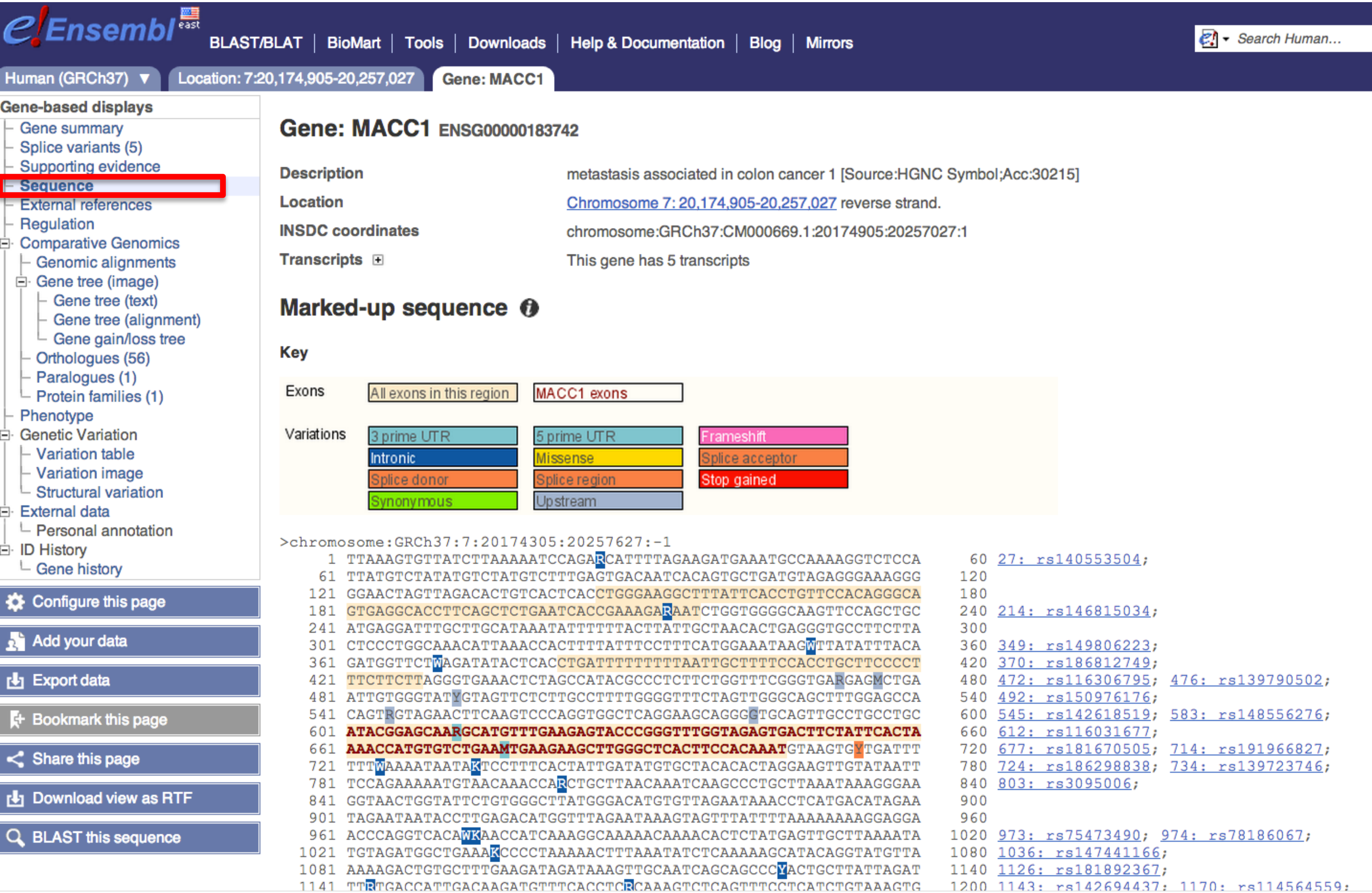
- Reaching a consensus coding sequence set for human and mouse.
- 26,473 (human)
  22,187 (mouse) (*as of Sept 2011)
- If you see a "CCDS ID", the coding sequence is agreed upon.

Genome Res. 2009 Jul;19(7):1316-23. Epub 2009 Jun 4

VEGA/Havana (human, mouse, z-fish)

- **Automatic annotation pipeline: Gene building all at once (whole genome)**
  Ensembl

- **Manual curation: reviewed by experts**
  VEGA: Vertebrate Genome Annotation
  Havana



Gene: MACC1 ENSG00000183742

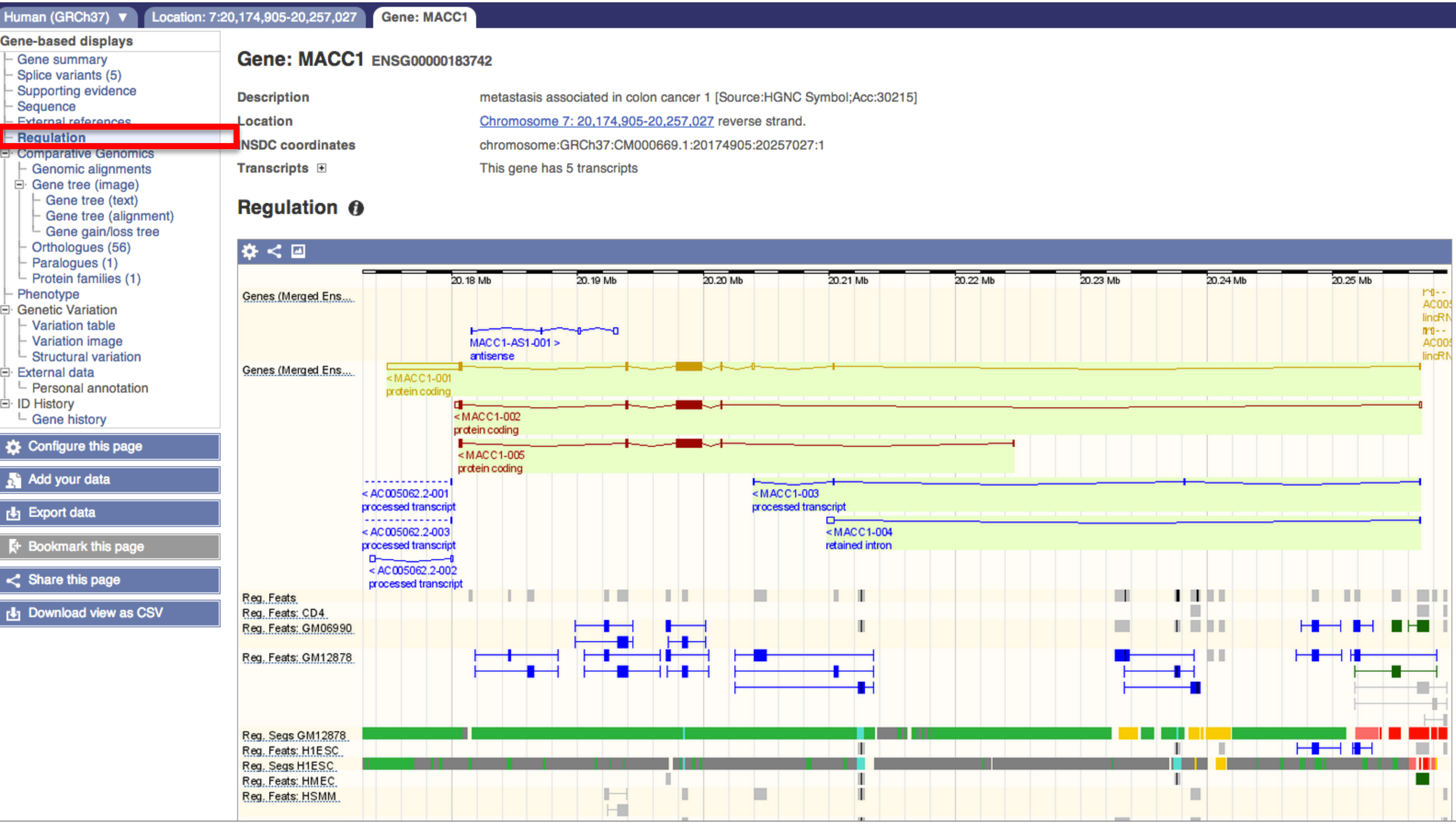
Description	metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]
Location	Chromosome 7: 20,174,905-20,257,027 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000669.1:20174905:20257027:1
Transcripts ⊞	This gene has 5 transcripts

Marked-up sequence

Key

Exons	All exons in this region	MACC1 exons	
Variations	3 prime UTR	5 prime UTR	Frameshift
	Intronic	Missense	Splice acceptor
	Splice donor	Splice region	Stop gained
	Synonymous	Upstream	

```
>chromosome:GRCh37:7:20174305:20257627:-1
   1 TTAAAGTGTTATCTTAAAAATCCAGARCATTTTAGAAGATGAAATGCCAAAAGGTCTCCA     60 27: rs140553504;
  61 TTATGTCTATATGTCTATGTCTTTGAGTGACAATCACAGTGCTGATGTAGAGGGAAAGGG    120
 121 GGAACTAGTTAGACACTGTCACTCACCTGGGAAGGCTTTATTCACCTGTTCCACAGGGCA    180
 181 GTGAGGCACCTTCAGCTCTGAATCACCGAAAGARAATCTGGTGGGGCAAGTTCCAGCTGC    240 214: rs146815034;
 241 ATGAGGATTTGCTTGCATAAATATTTTTTACTTATTGCTAACACTGAGGGTGCCTTCTTA    300
 301 CTCCCTGGCAAACATTAAACCACTTTTATTTCCTTTCATGGAAATAAGWTTATATTTACA    360 349: rs149806223;
 361 GATGGTTCTWAGATATACTCACCTGATTTTTTTTTAATTGCTTTTCCACCTGCTTCCCCT    420 370: rs186812749;
 421 TTCTTCTTAGGGTGAAACTCTAGCCATACGCCCTCTTCTGGTTTCGGGTGARGAGMCTGA    480 472: rs116306795; 476: rs139790502;
 481 ATTGTGGGTATYGTAGTTCTCTTGCCTTTTGGGGTTTCTAGTTGGGCAGCTTTGGAGCCA    540 492: rs150976176;
 541 CAGTRGTAGAACTTCAAGTCCCAGGTGGCTCAGGAAGCAGGGGTGCAGTTGCCTGCCTGC    600 545: rs142618519; 583: rs148556276;
 601 ATACGGAGCAARGCATGTTTGAAGAGTACCCGGGTTTGGTAGAGTGACTTCTATTCACTA    660 612: rs116031677;
 661 AAACCATGTGTCTGAAMTGAAGAAGCTTGGGCTCACTTCCACAAATGTAAGTGYTGATTT    720 677: rs181670505; 714: rs191966827;
 721 TTTWAAAATAATAKTCCTTTCACTATTGATATGTGCTACACACTAGGAAGTTGTATAATT    780 724: rs186298838; 734: rs139723746;
 781 TCCAGAAAAATGTAACAAACCARCTGCTTAACAAATCAAGCCCTGCTTAAATAAAGGGAA    840 803: rs3095006;
 841 GGTAACTGGTATTCTGTGGGCTTATGGGACATGTGTTAGAATAAACCTCATGACATAGAA    900
 901 TAGAATAATACCTTGAGACATGGTTTAGAATAAAGTAGTTTATTTTAAAAAAAAGGAGGA    960
 961 ACCCAGGTCACAWKAACCATCAAAGGCAAAAACAAAACACTCTATGAGTTGCTTAAAATA   1020 973: rs75473490; 974: rs78186067;
1021 TGTAGATGGCTGAAAKCCCCTAAAAACTTTAAATATCTCAAAAAGCATACAGGTATGTTA   1080 1036: rs147441166;
1081 AAAAGACTGTGCTTTGAAGATAGATAAAGTTGCAATCAGCAGCCCYACTGCTTATTAGAT   1140 1126: rs181892367;
1141 TTRTGACCATTGACAAGATGTTTCACCTCRCAAAGTCTCAGTTTCCTCATCTGTAAAGTG   1200 1143: rs142694437; 1170: rs114564559;
```



Gene: MACC1 ENSG00000183742

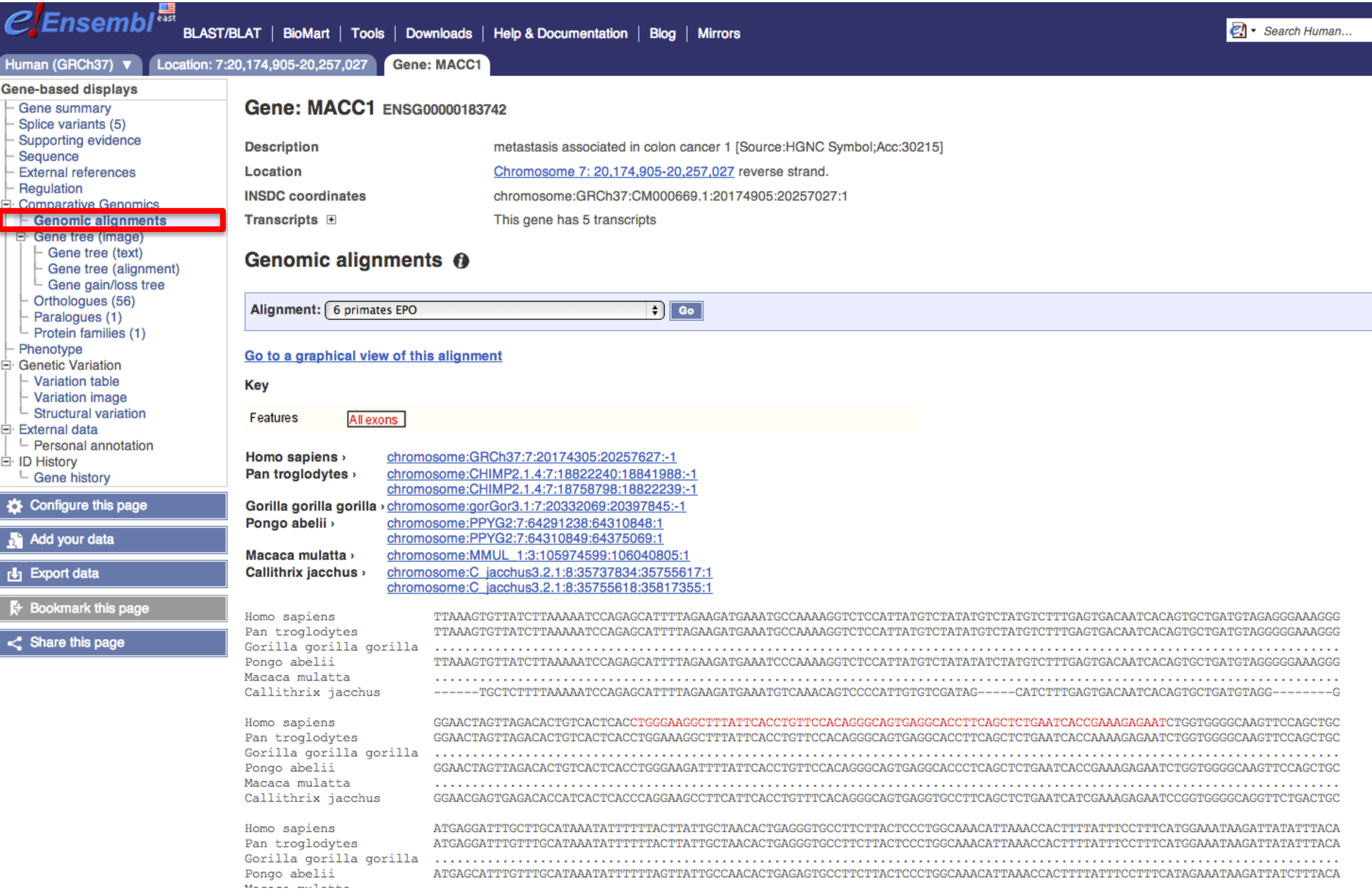
Description metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]

Location Chromosome 7: 20,174,905-20,257,027 reverse strand.

INSDC coordinates chromosome:GRCh37:CM000669.1:20174905:20257027:1

Transcripts ⊞ This gene has 5 transcripts

Regulation


e!Ensembl east
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Search Human…
Human (GRCh37) ▼
Location: 7:20,174,905-20,257,027
Gene: MACC1

Gene-based displays

- Gene summary
- Splice variants (5)
- Supporting evidence
- Sequence
- External references
- Regulation
- Comparative Genomics
  - Genomic alignments
  - Gene tree (image)
    - Gene tree (text)
    - Gene tree (alignment)
    - Gene gain/loss tree
  - Orthologues (56)
  - Paralogues (1)
  - Protein families (1)
- Phenotype
- Genetic Variation
  - Variation table
  - Variation image
  - Structural variation
- External data
  - Personal annotation
- ID History
  - Gene history

Gene: MACC1 ENSG00000183742

Description	metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]
Location	Chromosome 7: 20,174,905-20,257,027 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000669.1:20174905:20257027:1
Transcripts ⊞	This gene has 5 transcripts

Genomic alignments

Alignment: 6 primates EPO

Go to a graphical view of this alignment

Key

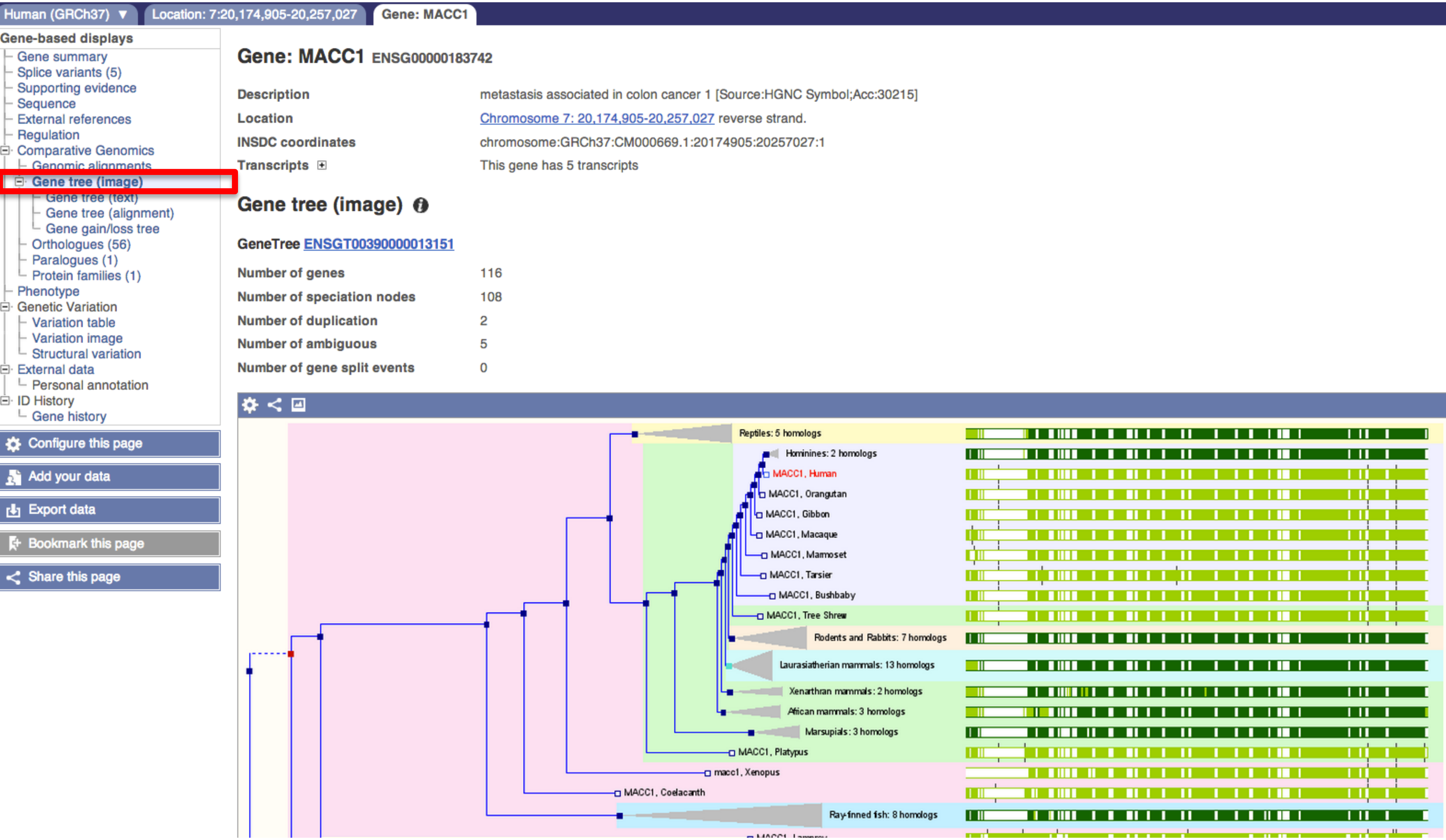
Features All exons

Homo sapiens ›	chromosome:GRCh37:7:20174305:20257627:-1
Pan troglodytes ›	chromosome:CHIMP2.1.4:7:18822240:18841988:-1 chromosome:CHIMP2.1.4:7:18758798:18822239:-1
Gorilla gorilla gorilla ›	chromosome:gorGor3.1:7:20332069:20397845:-1
Pongo abelii ›	chromosome:PPYG2:7:64291238:64310848:1 chromosome:PPYG2:7:64310849:64375069:1
Macaca mulatta ›	chromosome:MMUL_1:3:105974599:106040805:1
Callithrix jacchus ›	chromosome:C_jacchus3.2.1:8:35737834:35755617:1 chromosome:C_jacchus3.2.1:8:35755618:35817355:1

```
Homo sapiens             TTAAAGTGTTATCTTAAAAATCCAGAGCATTTTAGAAGATGAAATGCCAAAAGGTCTCCATTATGTCTATATGTCTATGTCTTTGAGTGACAATCACAGTGCTGATGTAGAGGGAAAGGG
Pan troglodytes          TTAAAGTGTTATCTTAAAAATCCAGAGCATTTTAGAAGATGAAATGCCAAAAGGTCTCCATTATGTCTATATGTCTATGTCTTTGAGTGACAATCACAGTGCTGATGTAGGGGGAAAGGG
Gorilla gorilla gorilla  ........................................................................................................................
Pongo abelii             TTAAAGTGTTATCTTAAAAATCCAGAGCATTTTAGAAGATGAAATCCCAAAAGGTCTCCATTATGTCTATATATCTATGTCTTTGAGTGACAATCACAGTGCTGATGTAGGGGGAAAGGG
Macaca mulatta           ........................................................................................................................
Callithrix jacchus       ------TGCTCTTTTAAAAATCCAGAGCATTTTAGAAGATGAAATGTCAAACAGTCCCCATTGTGTCGATAG-----CATCTTTGAGTGACAATCACAGTGCTGATGTAGG--------G

Homo sapiens             GGAACTAGTTAGACACTGTCACTCACCTGGGAAGGCTTTATTCACCTGTTCCACAGGGCAGTGAGGCACCTTCAGCTCTGAATCACCGAAAGAGAATCTGGTGGGGCAAGTTCCAGCTGC
Pan troglodytes          GGAACTAGTTAGACACTGTCACTCACCTGGAAAGGCTTTATTCACCTGTTCCACAGGGCAGTGAGGCACCTTCAGCTCTGAATCACCAAAAGAGAATCTGGTGGGGCAAGTTCCAGCTGC
Gorilla gorilla gorilla  ........................................................................................................................
Pongo abelii             GGAACTAGTTAGACACTGTCACTCACCTGGGAAGATTTTATTCACCTGTTCCACAGGGCAGTGAGGCACCCTCAGCTCTGAATCACCGAAAGAGAATCTGGTGGGGCAAGTTCCAGCTGC
Macaca mulatta           ........................................................................................................................
Callithrix jacchus       GGAACGAGTGAGACACCATCACTCACCCAGGAAGCCTTCATTCACCTGTTTCACAGGGCAGTGAGGTGCCTTCAGCTCTGAATCATCGAAAGAGAATCCGGTGGGGCAGGTTCTGACTGC

Homo sapiens             ATGAGGATTTGCTTGCATAAATATTTTTTACTTATTGCTAACACTGAGGGTGCCTTCTTACTCCCTGGCAAACATTAAACCACTTTTATTTCCTTTCATGGAAATAAGATTATATTTACA
Pan troglodytes          ATGAGGATTTGTTTGCATAAATATTTTTTACTTATTGCTAACACTGAGGGTGCCTTCTTACTCCCTGGCAAACATTAAACCACTTTTATTTCCTTTCATGGAAATAAGATTATATTTACA
Gorilla gorilla gorilla  ........................................................................................................................
Pongo abelii             ATGAGCATTTGTTTGCATAAATATTTTTTAGTTATTGCCAACACTGAGAGTGCCTTCTTACTCCCTGGCAAACATTAAACCACTTTTATTTCCTTTCATAGAAATAAGATTATCTTTACA
```



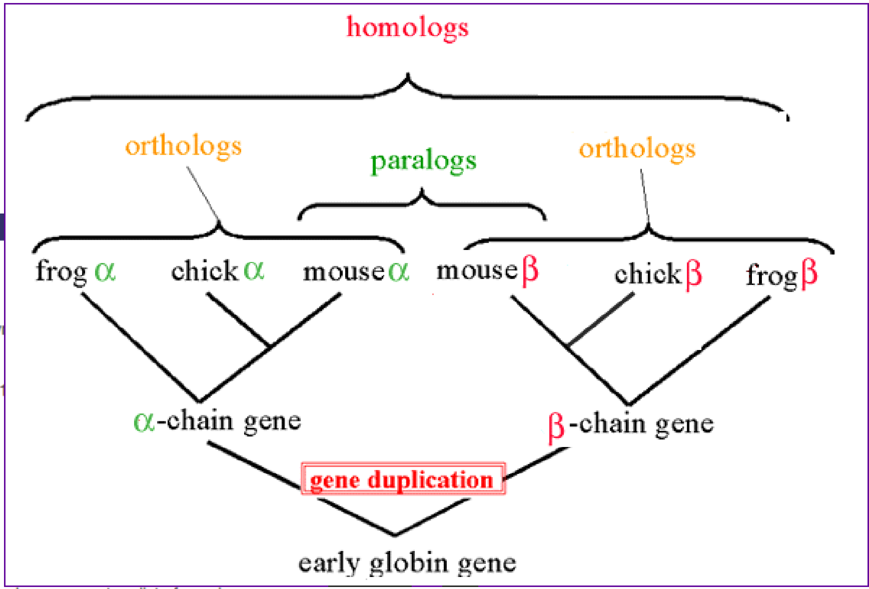
Gene: MACC1 ENSG00000183742

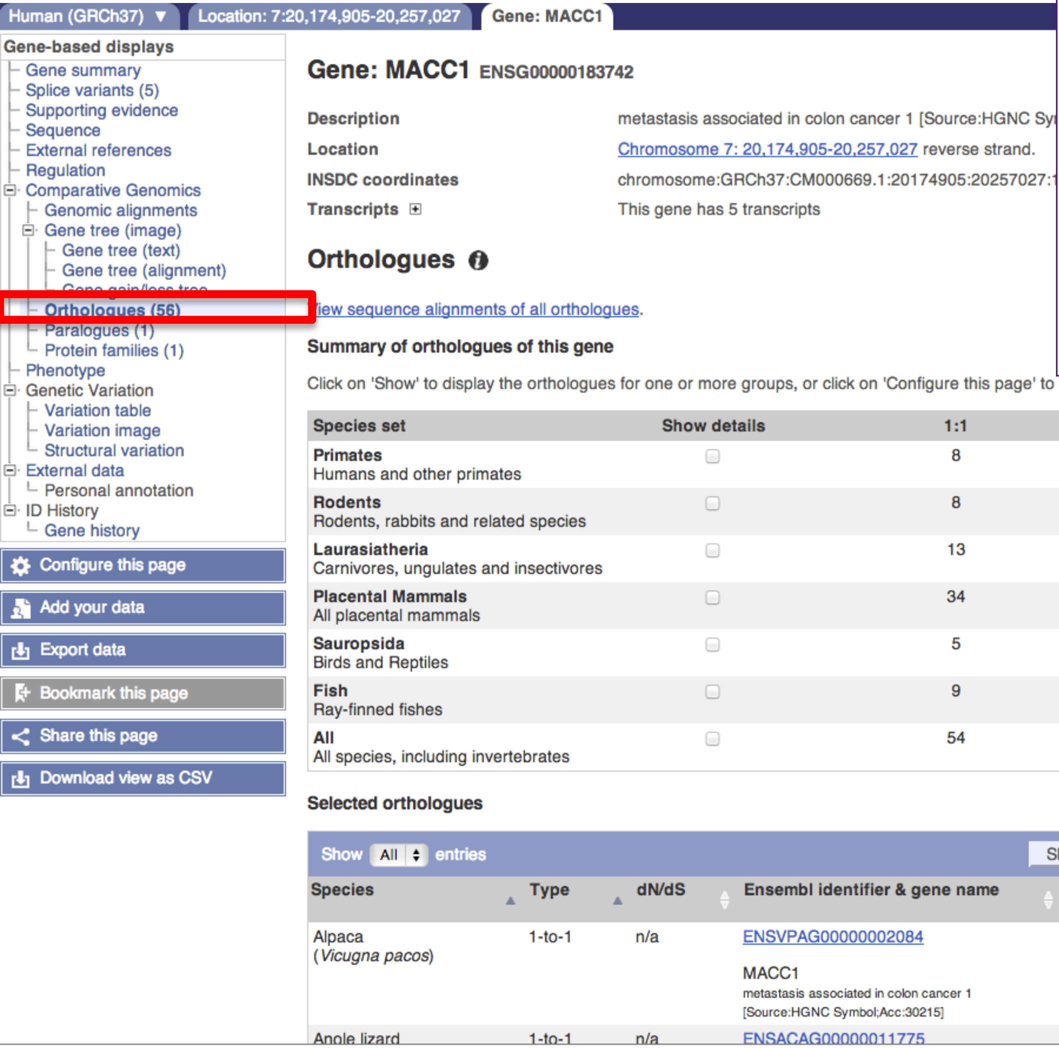
Description	metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]
Location	Chromosome 7: 20,174,905-20,257,027 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000669.1:20174905:20257027:1
Transcripts ⊞	This gene has 5 transcripts

Gene tree (image)

GeneTree ENSGT00390000013151

Number of genes	116
Number of speciation nodes	108
Number of duplication	2
Number of ambiguous	5
Number of gene split events	0





Gene-based displays

- Gene summary
- Splice variants (5)
- Supporting evidence
- Sequence
- External references
- Regulation
- Comparative Genomics
  - Genomic alignments
  - Gene tree (image)
    - Gene tree (text)
    - Gene tree (alignment)
    - Gene gain/loss tree
  - Orthologues (56)
  - Paralogues (1)
  - Protein families (1)
- Phenotype
- Genetic Variation
  - Variation table
  - Variation image
  - Structural variation
- External data
  - Personal annotation
- ID History
  - Gene history

Gene: MACC1 ENSG00000183742

Description	metastasis associated in colon cancer 1 [Source:HGNC Sy
Location	Chromosome 7: 20,174,905-20,257,027 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000669.1:20174905:20257027:1
Transcripts ⊞	This gene has 5 transcripts

Orthologues

iew sequence alignments of all orthologues.

Summary of orthologues of this gene

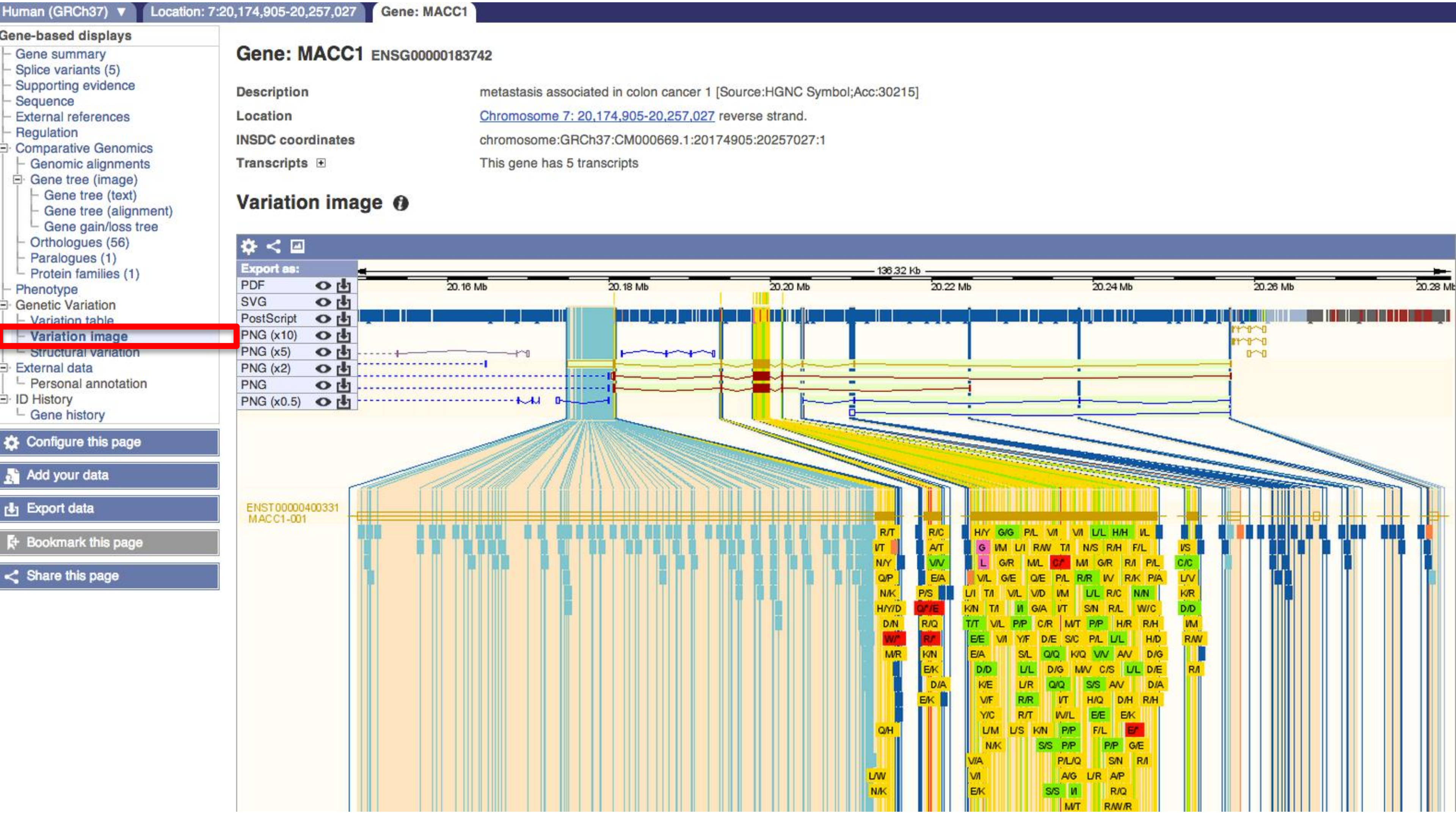
Click on 'Show' to display the orthologues for one or more groups, or click on 'Configure this page' to choose a custom list of species

Species set	Show details	1:1	1:many	many:many	No orthologues
Primates Humans and other primates	□	8	0	0	1
Rodents Rodents, rabbits and related species	□	8	0	0	0
Laurasiatheria Carnivores, ungulates and insectivores	□	13	0	0	0
Placental Mammals All placental mammals	□	34	0	0	1
Sauropsida Birds and Reptiles	□	5	0	0	0
Fish Ray-finned fishes	□	9	0	0	0
All All species, including invertebrates	□	54	2	0	4

Selected orthologues

Show All entries — Show/hide columns — Filter

Species	Type	dN/dS	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Alpaca (*Vicugna pacos*)	1-to-1	n/a	ENSVPAG00000002084 MACC1 metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	GeneScaffold_1891:823408-842968:-1	84	84
Anole lizard	1-to-1	n/a	ENSACAG00000011775	• Region Comparison	6:29083541-29105343:-1	61	61



Human (GRCh37) ▼ | Location: 7:20,174,905-20,257,027 | Gene: MACC1

Gene-based displays

- Gene summary
- Splice variants (5)
- Supporting evidence
- Sequence
- External references
- Regulation
- Comparative Genomics
  - Genomic alignments
  - Gene tree (image)
    - Gene tree (text)
    - Gene tree (alignment)
    - Gene gain/loss tree
  - Orthologues (56)
  - Paralogues (1)
  - Protein families (1)
- 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

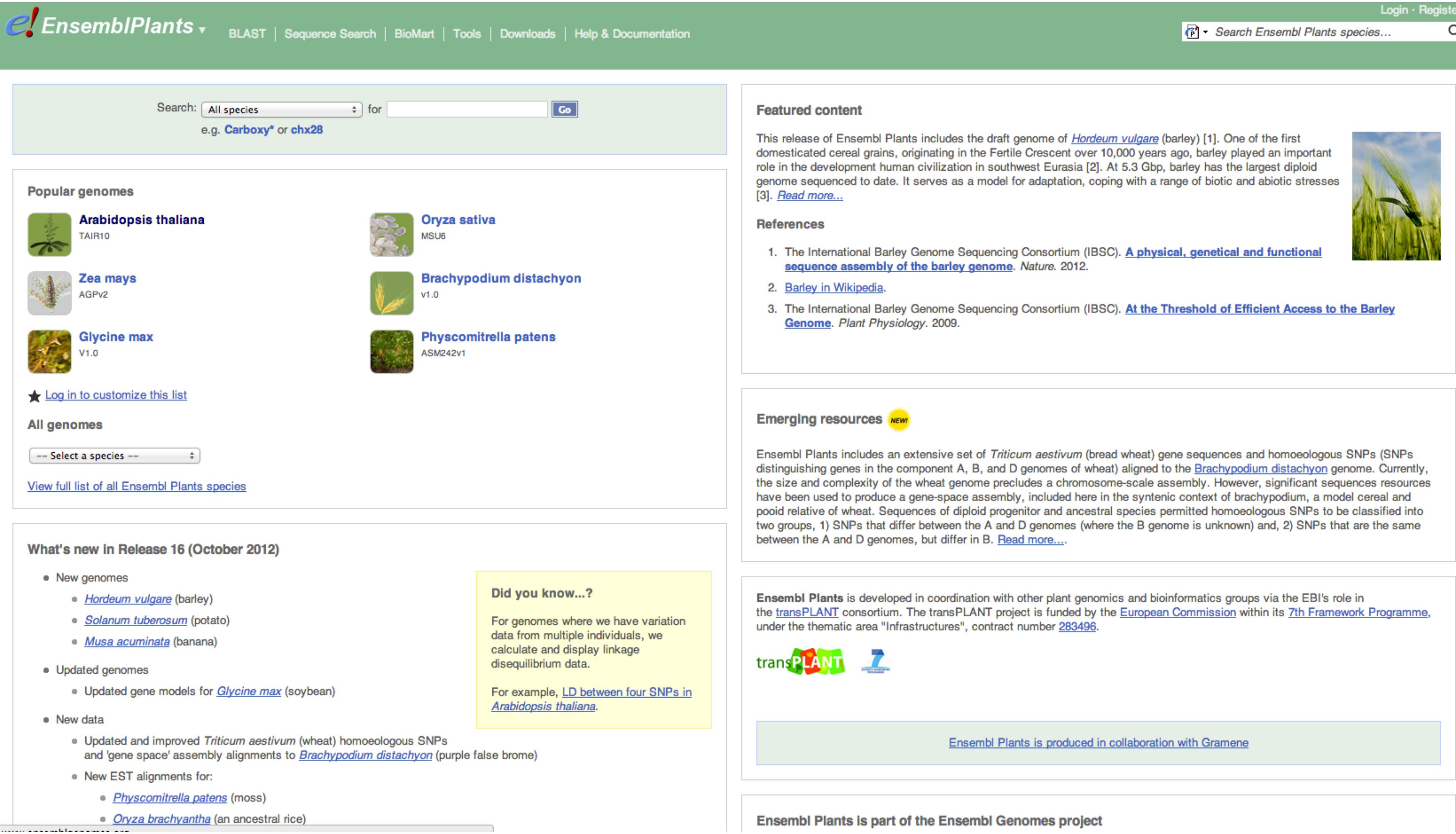
Bookmark this page

Share this page

Gene: MACC1 ENSG00000183742

Description	metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]
Location	Chromosome 7: 20,174,905-20,257,027 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000669.1:20174905:20257027:1
Transcripts ⊞	This gene has 5 transcripts

Variation image


EnsemblPlants
BLAST | Sequence Search | BioMart | Tools | Downloads | Help & Documentation
Login · Registe
Search Ensembl Plants species…
Search: All species for Go
e.g. Carboxy* or chx28

Popular genomes

Arabidopsis thaliana TAIR10	**Oryza sativa** MSU6
Zea mays AGPv2	**Brachypodium distachyon** v1.0
Glycine max V1.0	**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 16 (October 2012)

- New genomes
  - *Hordeum vulgare* (barley)
  - *Solanum tuberosum* (potato)
  - *Musa acuminata* (banana)
- Updated genomes
  - Updated gene models for *Glycine max* (soybean)
- New data
  - Updated and improved *Triticum aestivum* (wheat) homoeologous SNPs and 'gene space' assembly alignments to *Brachypodium distachyon* (purple false brome)
  - New EST alignments for:
    - *Physcomitrella patens* (moss)
    - *Oryza brachyantha* (an ancestral rice)

Did you know...?

For genomes where we have variation data from multiple individuals, we calculate and display linkage disequilibrium data.

For example, LD between four SNPs in *Arabidopsis thaliana*.

Featured content

This release of Ensembl Plants includes the draft genome of *Hordeum vulgare* (barley) [1]. One of the first domesticated cereal grains, originating in the Fertile Crescent over 10,000 years ago, barley played an important role in the development human civilization in southwest Eurasia [2]. At 5.3 Gbp, barley has the largest diploid genome sequenced to date. It serves as a model for adaptation, coping with a range of biotic and abiotic stresses [3]. *Read more...*

References

1. The International Barley Genome Sequencing Consortium (IBSC). **A physical, genetical and functional sequence assembly of the barley genome**. *Nature*. 2012.
2. Barley in Wikipedia.
3. The International Barley Genome Sequencing Consortium (IBSC). **At the Threshold of Efficient Access to the Barley Genome**. *Plant Physiology*. 2009.

Emerging resources NEW!

Ensembl Plants includes an extensive set of *Triticum aestivum* (bread wheat) gene sequences and homoeologous SNPs (SNPs distinguishing genes in the component A, B, and D genomes of wheat) aligned to the Brachypodium distachyon genome. Currently, the size and complexity of the wheat genome precludes a chromosome-scale assembly. However, significant sequences resources have been used to produce a gene-space assembly, included here in the syntenic context of brachypodium, a model cereal and pooid relative of wheat. Sequences of diploid progenitor and ancestral species permitted homoeologous SNPs to be classified into two groups, 1) SNPs that differ between the A and D genomes (where the B genome is unknown) and, 2) SNPs that are the same between the A and D genomes, but differ in B. Read more....

Ensembl Plants is developed in coordination with other plant genomics and bioinformatics groups via the EBI's role in the transPLANT consortium. The transPLANT project is funded by the European Commission within its 7th Framework Programme, under the thematic area "Infrastructures", contract number 283496.

Ensembl Plants is produced in collaboration with Gramene

Ensembl Plants is part of the Ensembl Genomes project

Next lecture: ExPASy and DTU tools