Browsing Genomes with Ensembl

www.ensembl.org
www.ensemblgenomes.org

Coursebook v72

http://tinyurl.com/Camb0613

Cambridge – 21st June 2013

TABLE OF CONTENTS

Introduction to Ensembl

Getting started with Ensembl
www.ensembl.org

Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute that annotates **chordate** genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl 'compara' team. Most annotation is updated every two months, leading to increasing Ensembl versions (such as version 70), however the gene sets are determined less frequently. A sister browser at www.ensemblgenomes.org is set up to access non-chordates, namely bacteria, plants, fungi, metazoa, and protists.

Ensembl provides genes and other **annotation** such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on protein and mRNA evidence in **UniProtKB** and **NCBI RefSeq** databases, along with manual annotation from the **VEGA/Havana** group. All the data are freely available and can be accessed via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interfaces (**Perl APIs**). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the **BioMart** web interface, which can extract information from the Ensembl databases without the need for programming knowledge by the user.

Synopsis – What can I do with Ensembl?

- View genes with other annotation along the chromosome.
- View alternative transcripts (i.e. splice variants) for a given gene.
- Explore homologues and phylogenetic trees across more than 60 species for any gene.
- Compare whole genome alignments and conserved regions across species.
- View microarray sequences that match to Ensembl genes.
- View ESTs, clones, mRNA and proteins for any chromosomal region.
- Examine single nucleotide polymorphisms (SNPs) for a gene or chromosomal region.
- View SNPs across strains (rat, mouse), populations (human), or breeds (dog).
- View positions and sequence of mRNAs and proteins that align with Ensembl genes.
- Upload your own data.
- Use BLAST, or BLAT against any Ensembl genome.
- Export sequence or create a table of gene information with BioMart.
- Determine how your variants affect genes and transcripts using the Variant Effect Predictor.
- Share Ensembl views with your colleagues and collaborators.

Need more help?

- Check Ensembl documentation
- Watch video tutorials on YouTube
- View the FAQs
- Try some exercises
- Read some publications
- Go to our online course

Stay in touch!

- Email the team with comments or questions at helpdesk@ensembl.org
- Follow the Ensembl blog
- Sign up to a mailing list

Further reading

Flicek, P. *et al*
Ensembl 2013
Nucleic Acids Res. Advanced Access (Database Issue)
http://www.ncbi.nlm.nih.gov/pubmed/23203987

Ensembl Methods Series
http://www.biomedcentral.com/series/ENSEMBL2010

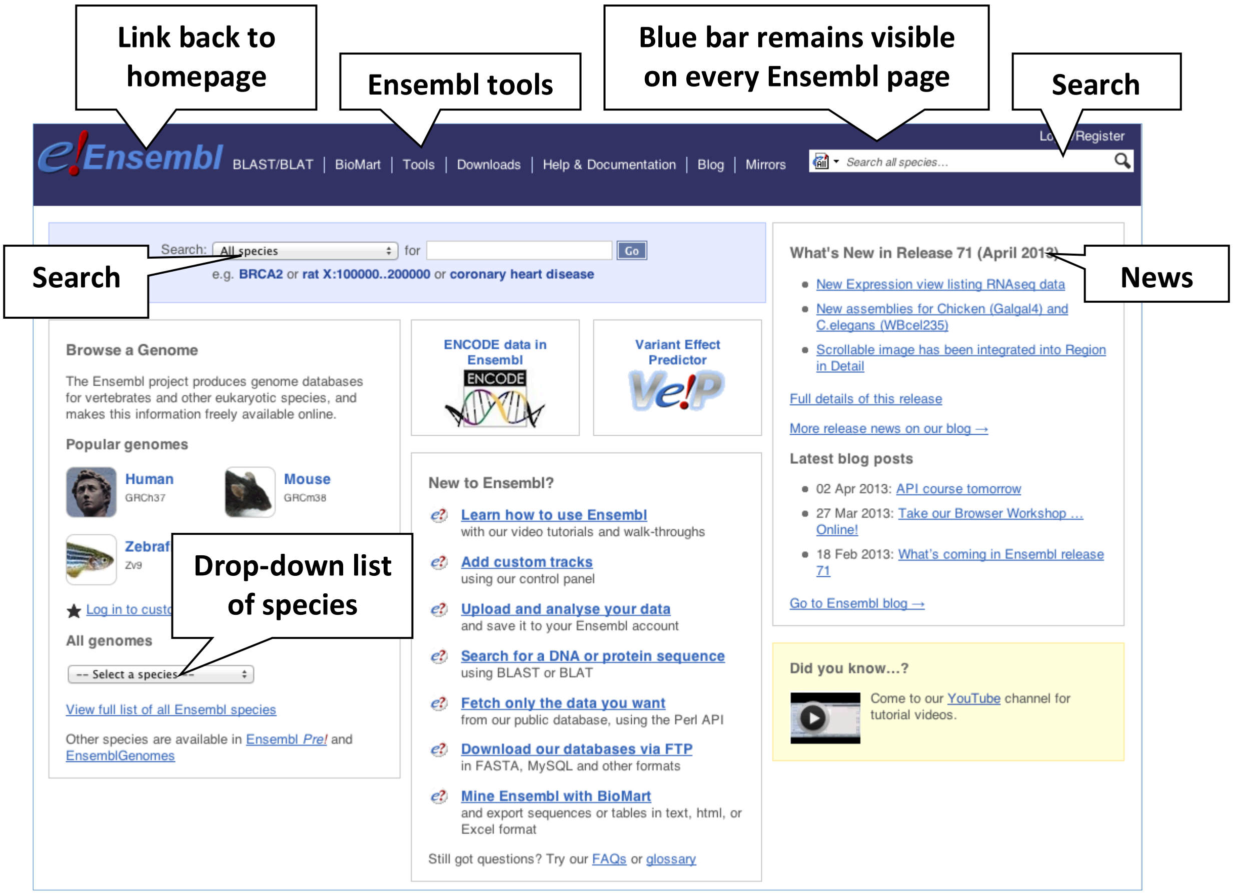
Xosé M. Fernández-Suárez and Michael K. Schuster
Using the Ensembl Genome Server to Browse Genomic Sequence Data.
UNIT 1.15 in Current Protocols in Bioinformatics, Jun 2010.

Giulietta M Spudich and Xosé M Fernández-Suárez
Touring Ensembl: A practical guide to genome browsing
BMC Genomics 2010, 11:295 (11 May 2010)

Exploring the Ensembl genome browser

Demo: Ensembl species

The front page of Ensembl is found at ensembl.org. It contains lots of information and links to help you navigate Ensembl:



Link back to homepage

Ensembl tools

Blue bar remains visible on every Ensembl page

Search

Search

News

Drop-down list of species

e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Lo /Register
Search all species...
Search: All species for Go

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

Zebraf Zv9

Log in to custo

All genomes

-- Select a species --

View full list of all Ensembl species

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

ENCODE data in Ensembl

Variant Effect Predictor

New to Ensembl?

- Learn how to use Ensembl with our video tutorials and walk-throughs
- Add custom tracks using our control panel
- Upload and analyse your data and save it to your Ensembl account
- Search for a DNA or protein sequence using BLAST or BLAT
- Fetch only the data you want from our public database, using the Perl API
- Download our databases via FTP in FASTA, MySQL and other formats
- Mine Ensembl with BioMart and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs or glossary

What's New in Release 71 (April 2013)

- New Expression view listing RNAseq data
- New assemblies for Chicken (Galgal4) and C.elegans (WBcel235)
- Scrollable image has been integrated into Region in Detail

Full details of this release

More release news on our blog →

Latest blog posts

- 02 Apr 2013: API course tomorrow
- 27 Mar 2013: Take our Browser Workshop ... Online!
- 18 Feb 2013: What's coming in Ensembl release 71

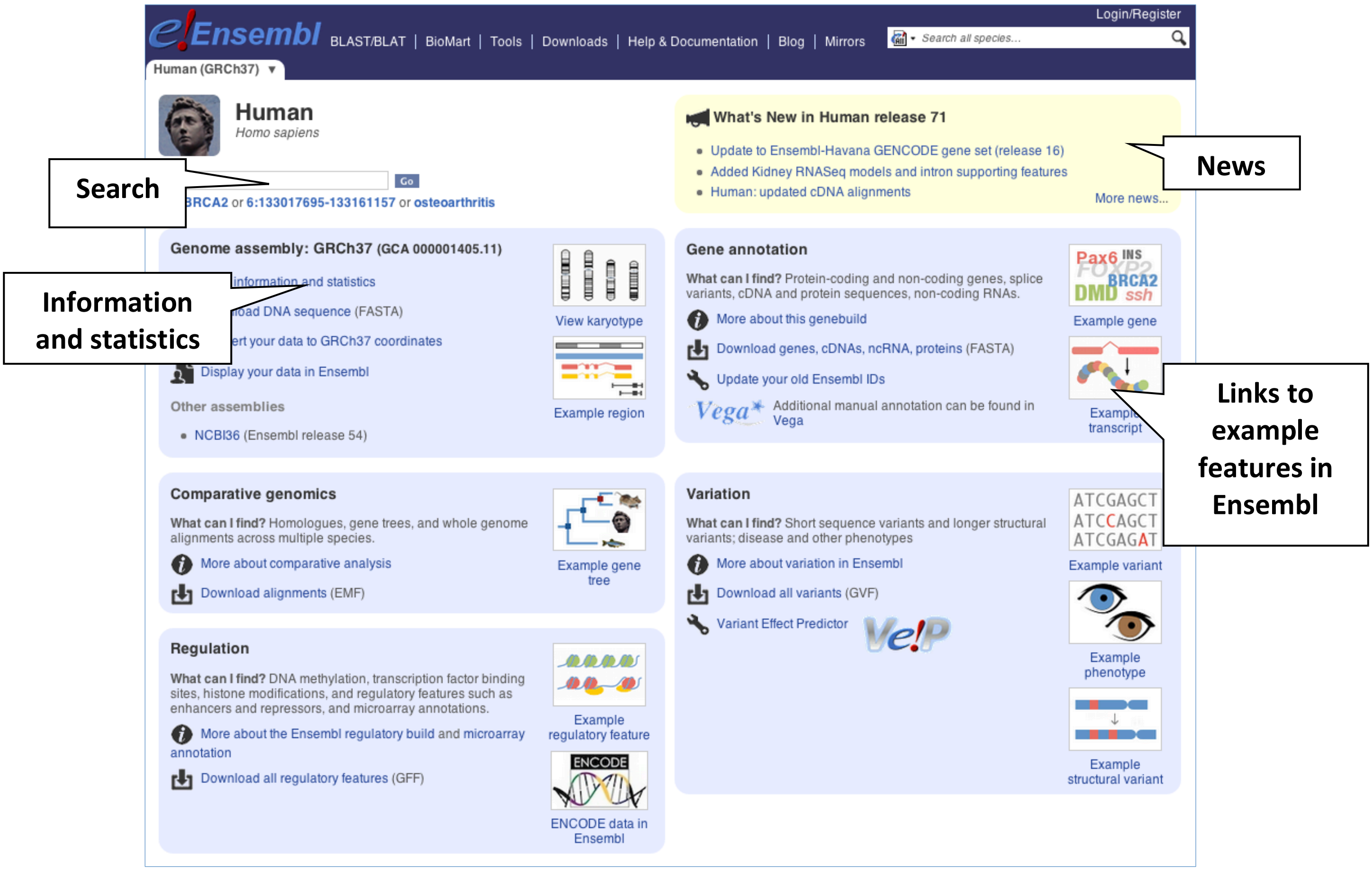
Go to Ensembl blog →

Did you know...?

Come to our YouTube channel for tutorial videos.

Click on View full list of all Ensembl species.

Click on the common name of your species of interest to go to the species homepage. We'll click on Human.


Search
News
Information and statistics
Links to example features in Ensembl
Login/Register
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Search all species...
Human (GRCh37) ▼
View karyotype
Example region
Example gene
Example transcript
Example gene tree
Example regulatory feature
ENCODE data in Ensembl
Example variant
Example phenotype
Example structural variant

Human
Homo sapiens

Go
BRCA2 or 6:133017695-133161157 or osteoarthritis

What's New in Human release 71

- Update to Ensembl-Havana GENCODE gene set (release 16)
- Added Kidney RNASeq models and intron supporting features
- Human: updated cDNA alignments

More news...

Genome assembly: GRCh37 (GCA 000001405.11)

information and statistics
load DNA sequence (FASTA)
ert 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
Download alignments (EMF)

Variation

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

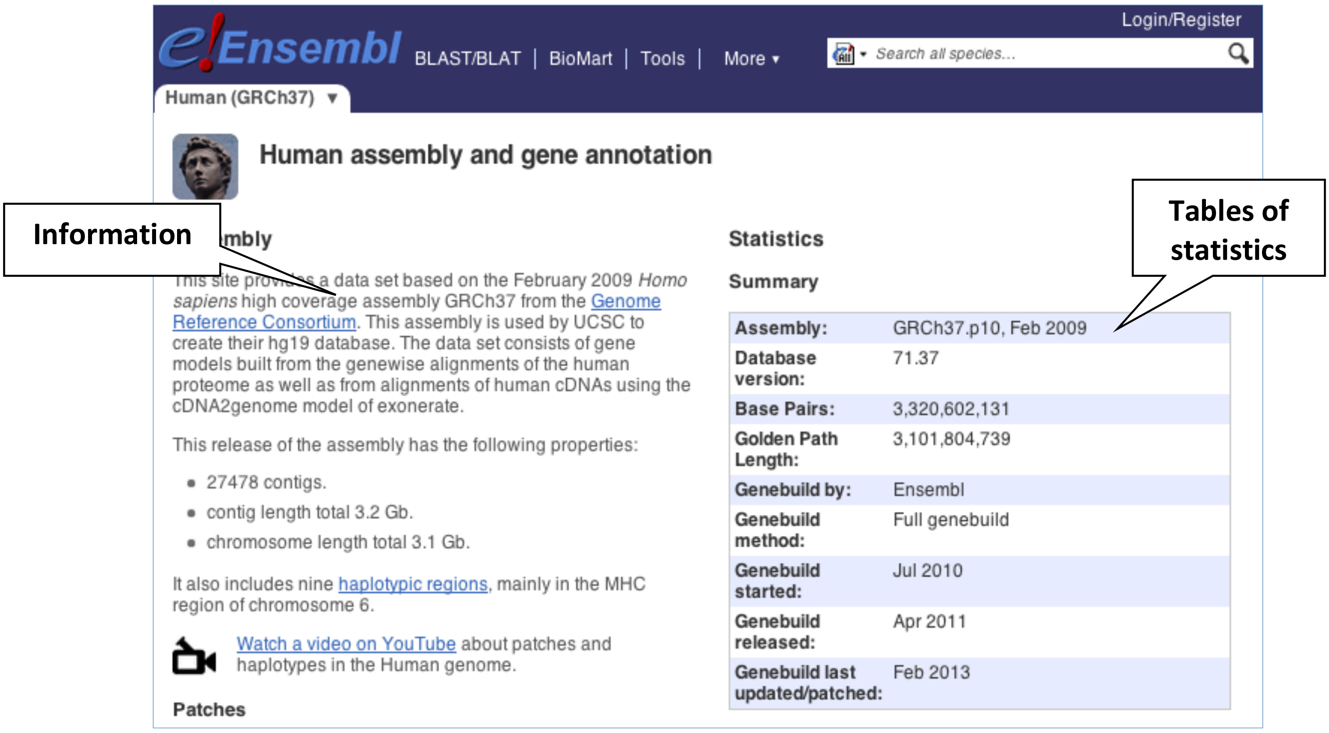
More about variation in Ensembl
Download all variants (GVF)
Variant Effect Predictor VEP

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

More about the Ensembl regulatory build and microarray annotation
Download all regulatory features (GFF)

To find out more about the genome assembly and genebuild, click on More information and statistics.



Human assembly and gene annotation

mbly

This site provides a data set based on the February 2009 *Homo sapiens* high coverage assembly GRCh37 from the Genome Reference Consortium. This assembly is used by UCSC to create their hg19 database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

It also includes nine haplotypic regions, mainly in the MHC region of chromosome 6.

Watch a video on YouTube about patches and haplotypes in the Human genome.

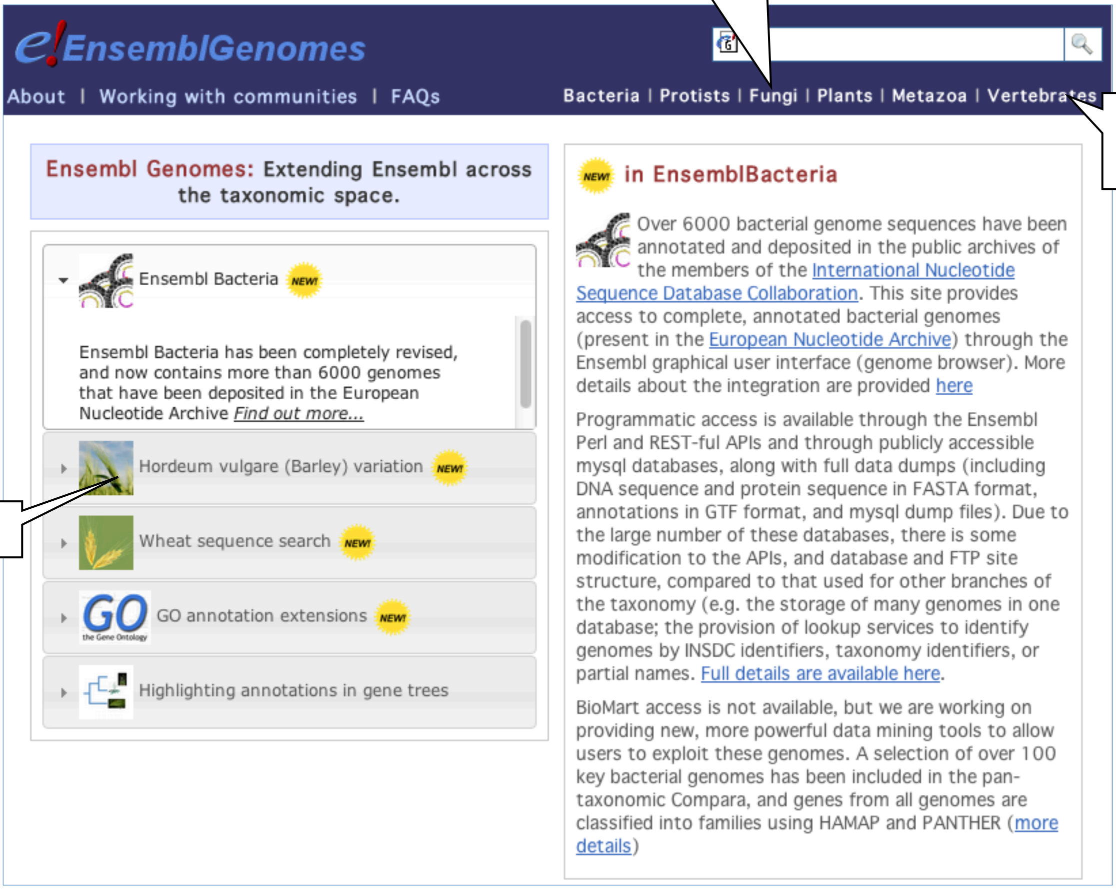
Patches

Statistics

Summary

Assembly:	GRCh37.p10, Feb 2009
Database version:	71.37
Base Pairs:	3,320,602,131
Golden Path Length:	3,101,804,739
Genebuild by:	Ensembl
Genebuild method:	Full genebuild
Genebuild started:	Jul 2010
Genebuild released:	Apr 2011
Genebuild last updated/patched:	Feb 2013

Let’s take a look at the Ensembl Genomes homepage at ensemblgenomes.org.


EnsemblGenomes
About | Working with communities | FAQs
Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates

Links to the taxa-specific sites

Link back to Ensembl

News

Ensembl Genomes: Extending Ensembl across the taxonomic space.

Ensembl Bacteria NEW!

Ensembl Bacteria has been completely revised, and now contains more than 6000 genomes that have been deposited in the European Nucleotide Archive *Find out more...*

Hordeum vulgare (Barley) variation NEW!

Wheat sequence search NEW!

GO annotation extensions NEW!

Highlighting annotations in gene trees

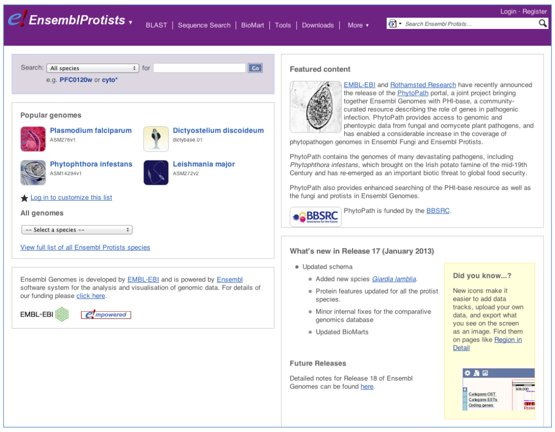
NEW! in EnsemblBacteria

Over 6000 bacterial genome sequences have been annotated and deposited in the public archives of the members of the International Nucleotide Sequence Database Collaboration. This site provides access to complete, annotated bacterial genomes (present in the European Nucleotide Archive) through the Ensembl graphical user interface (genome browser). More details about the integration are provided here

Programmatic access is available through the Ensembl Perl and REST-ful APIs and through publicly accessible mysql databases, along with full data dumps (including DNA sequence and protein sequence in FASTA format, annotations in GTF format, and mysql dump files). Due to the large number of these databases, there is some modification to the APIs, and database and FTP site structure, compared to that used for other branches of the taxonomy (e.g. the storage of many genomes in one database; the provision of lookup services to identify genomes by INSDC identifiers, taxonomy identifiers, or partial names. Full details are available here.

BioMart access is not available, but we are working on providing new, more powerful data mining tools to allow users to exploit these genomes. A selection of over 100 key bacterial genomes has been included in the pan-taxonomic Compara, and genes from all genomes are classified into families using HAMAP and PANTHER (more details)

Click on the different taxa to see their homepages. Each one is colour-coded.



Search: All species for
e.g. PFC0120w or cyto*

Popular genomes

Plasmodium falciparum ASM276v1
Dictyostelium discoideum dictybase.01
Phytophthora infestans ASM14294v1
Leishmania major ASM272v2

Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl Protists species

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please click here.

Featured content

EMBL-EBI and Rothamsted Research have recently announced the release of the PhytoPath portal, a joint project bringing together Ensembl Genomes with PHI-base, a community-curated resource describing the role of genes in pathogenic infection. PhytoPath provides access to genomic and phentoypic data from fungal and oomycete plant pathogens, and has enabled a considerable increase in the coverage of phytopathogen genomes in Ensembl Fungi and Ensembl Protists.

PhytoPath contains the genomes of many devastating pathogens, including *Phytophthora infestans*, which brought on the Irish potato famine of the mid-19th Century and has re-emerged as an important biotic threat to global food security.

PhytoPath also provides enhanced searching of the PHI-base resource as well as the fungi and protists in Ensembl Genomes.

PhytoPath is funded by the BBSRC.

What's new in Release 17 (January 2013)

- Updated schema
  - Added new spcies *Giardia lamblia*.
  - Protein features updated for all the protist species.
  - Minor internal fixes for the comparative genomics database
  - Updated BioMarts

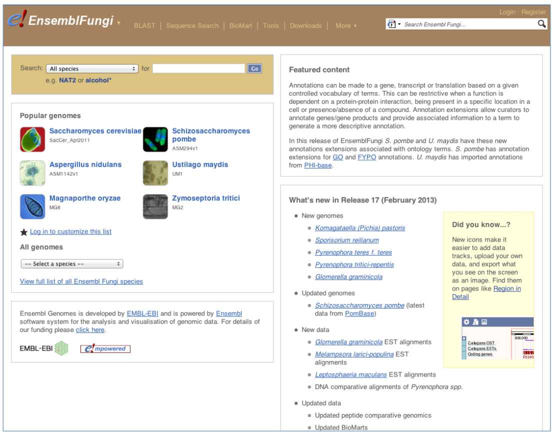
Future Releases

Detailed notes for Release 18 of Ensembl Genomes can be found here.

Did you know...?

New icons make it easier to add data tracks, upload your own data, and export what you see on the screen as an image. Find them on pages like Region in Detail

Protists



Search: All species for
e.g. NAT2 or alcohol*

Popular genomes

Saccharomyces cerevisiae SacCer_Apr2011
Schizosaccharomyces pombe ASM294v1
Aspergillus nidulans ASM1142v1
Ustilago maydis UM1
Magnaporthe oryzae MG8
Zymoseptoria tritici MG2

Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl Fungi species

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please click here.

Featured content

Annotations can be made to a gene, transcript or translation based on a given controlled vocabulary of terms. This can be restrictive when a function is dependent on a protein-protein interaction, being present in a specific location in a cell or presence/absence of a compound. Annotation extensions allow curators to annotate genes/gene products and provide associated information to a term to generate a more descriptive annotation.

In this release of EnsemblFungi *S. pombe* and *U. maydis* have these new annotations extensions associated with ontology terms. *S. pombe* has annotation extensions for GO and FYPO annotations. *U. maydis* has imported annotations from PHI-base.

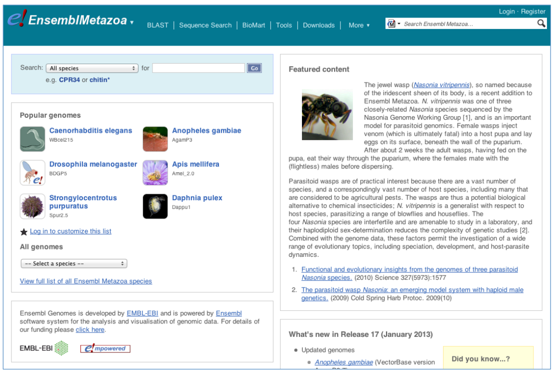
What's new in Release 17 (February 2013)

- New genomes
  - *Komagataella (Pichia) pastoris*
  - *Sporisorium reilianum*
  - *Pyrenophora teres f. teres*
  - *Pyrenophora tritici-repentis*
  - *Glomerella graminicola*
- Updated genomes
  - *Schizosaccharomyces pombe* (latest data from PomBase)
- New data
  - *Glomerella graminicola* EST alignments
  - *Melampsora larici-populina* EST alignments
  - *Leptosphaeria maculans* EST alignments
  - DNA comparative alignments of *Pyrenophora spp.*
- Updated data
  - Updated peptide comparative genomics
  - Updated BioMarts

Did you know...?

New icons make it easier to add data tracks, upload your own data, and export what you see on the screen as an image. Find them on pages like Region in Detail

Fungi



EnsemblMetazoa
BLAST | Sequence Search | BioMart | Tools | Downloads | More
Login · Register
Search Ensembl Metazoa…

Search: All species for Go
e.g. CPR34 or chitin*

Popular genomes

Caenorhabditis elegans WBcel215
Anopheles gambiae AgamP3
Drosophila melanogaster BDGP5
Apis mellifera Amel_2.0
Strongylocentrotus purpuratus Spur2.5
Daphnia pulex Dappu1

Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl Metazoa species

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please click here.

Featured content

The jewel wasp (*Nasonia vitripennis*), so named because of the iridescent sheen of its body, is a recent addition to Ensembl Metazoa. *N. vitripennis* was one of three closely-related *Nasonia* species sequenced by the Nasonia Genome Working Group [1], and is an important model for parasitoid genomics. Female wasps inject venom (which is ultimately fatal) into a host pupa and lay eggs on its surface, beneath the wall of the puparium. After about 2 weeks the adult wasps, having fed on the pupa, eat their way through the puparium, where the females mate with the (flightless) males before dispersing.

Parasitoid wasps are of practical interest because there are a vast number of species, and a correspondingly vast number of host species, including many that are considered to be agricultural pests. The wasps are thus a potential biological alternative to chemical insecticides; *N. vitripennis* is a generalist with respect to host species, parasitizing a range of blowflies and houseflies. The four *Nasonia* species are interfertile and are amenable to study in a laboratory, and their haplodiploid sex-determination reduces the complexity of genetic studies [2]. Combined with the genome data, these factors permit the investigation of a wide range of evolutionary topics, including speciation, development, and host-parasite dynamics.

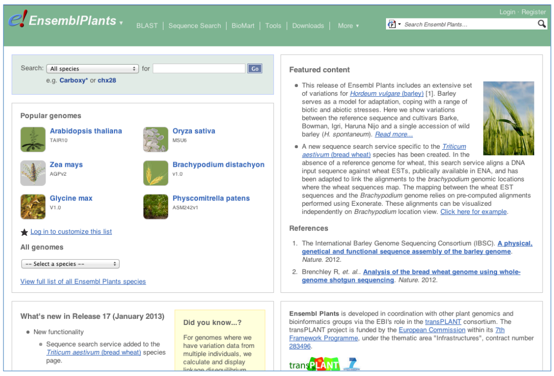
1. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. (2010) Science 327(5973):1577
2. The parasitoid wasp *Nasonia*: an emerging model system with haploid male genetics. (2009) Cold Spring Harb Protoc. 2009(10)

What's new in Release 17 (January 2013)

- Updated genomes
  - *Anopheles gambiae* (VectorBase version

Did you know...?

Metazoa



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 17 (January 2013)

- New functionality
  - Sequence search service added to the *Triticum aestivum* (bread wheat) species page.

Did you know...?

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

Featured content

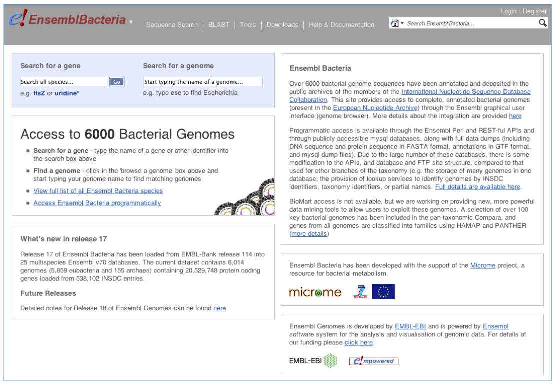
- This release of Ensembl Plants includes an extensive set of variations for *Hordeum vulgare* (barley) [1]. Barley serves as a model for adaptation, coping with a range of biotic and abiotic stresses. Here we show variations between the reference sequence and cultivars Barke, Bowman, Igri, Haruna Nijo and a single accession of wild barley (*H. spontaneum*). *Read more...*
- A new sequence search service specific to the *Triticum aestivum* (bread wheat) species has been created. In the absence of a reference genome for wheat, this search service aligns a DNA input sequence against wheat ESTs, publically available in ENA, and has been adapted to link the alignments to the *brachypodium* genomic locations where the wheat sequences map. The mapping between the wheat EST sequences and the *Brachypodium* genome relies on pre-computed alignments performed using Exonerate. These alignments can be visualized independently on *Brachypodium* location view. Click here for example.

References

1. The International Barley Genome Sequencing Consortium (IBSC). **A physical, genetical and functional sequence assembly of the barley genome**. *Nature.* 2012.
2. Brenchley R, *et. al.*. **Analysis of the bread wheat genome using whole-genome shotgun sequencing**. *Nature.* 2012.

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.

Plants



EnsemblBacteria
Sequence Search | BLAST | Tools | Downloads | Help & Documentation
Login · Register
Search Ensembl Bacteria…

Search for a gene

Search all species... Go
e.g. ftsZ or uridine*

Search for a genome

Start typing the name of a genome...
e.g. type esc to find Escherichia

Access to **6000** Bacterial Genomes

- **Search for a gene** - type the name of a gene or other identifier into the search box above
- **Find a genome** - click in the 'browse a genome' box above and start typing your genome name to find matching genomes
- View full list of all Ensembl Bacteria species
- Access Ensembl Bacteria programmatically

What's new in release 17

Release 17 of Ensembl Bacteria has been loaded from EMBL-Bank release 114 into 25 multispecies Ensembl v70 databases. The current dataset contains 6,014 genomes (5,859 eubacteria and 155 archaea) containing 20,529,748 protein coding genes loaded from 538,102 INSDC entries.

Future Releases

Detailed notes for Release 18 of Ensembl Genomes can be found here.

Ensembl Bacteria

Over 6000 bacterial genome sequences have been annotated and deposited in the public archives of the members of the International Nucleotide Sequence Database Collaboration. This site provides access to complete, annotated bacterial genomes (present in the European Nucleotide Archive) through the Ensembl graphical user interface (genome browser). More details about the integration are provided here

Programmatic access is available through the Ensembl Perl and REST-ful APIs and through publicly accessible mysql databases, along with full data dumps (including DNA sequence and protein sequence in FASTA format, annotations in GTF format, and mysql dump files). Due to the large number of these databases, there is some modification to the APIs, and database and FTP site structure, compared to that used for other branches of the taxonomy (e.g. the storage of many genomes in one database; the provision of lookup services to identify genomes by INSDC identifiers, taxonomy identifiers, or partial names. Full details are available here.

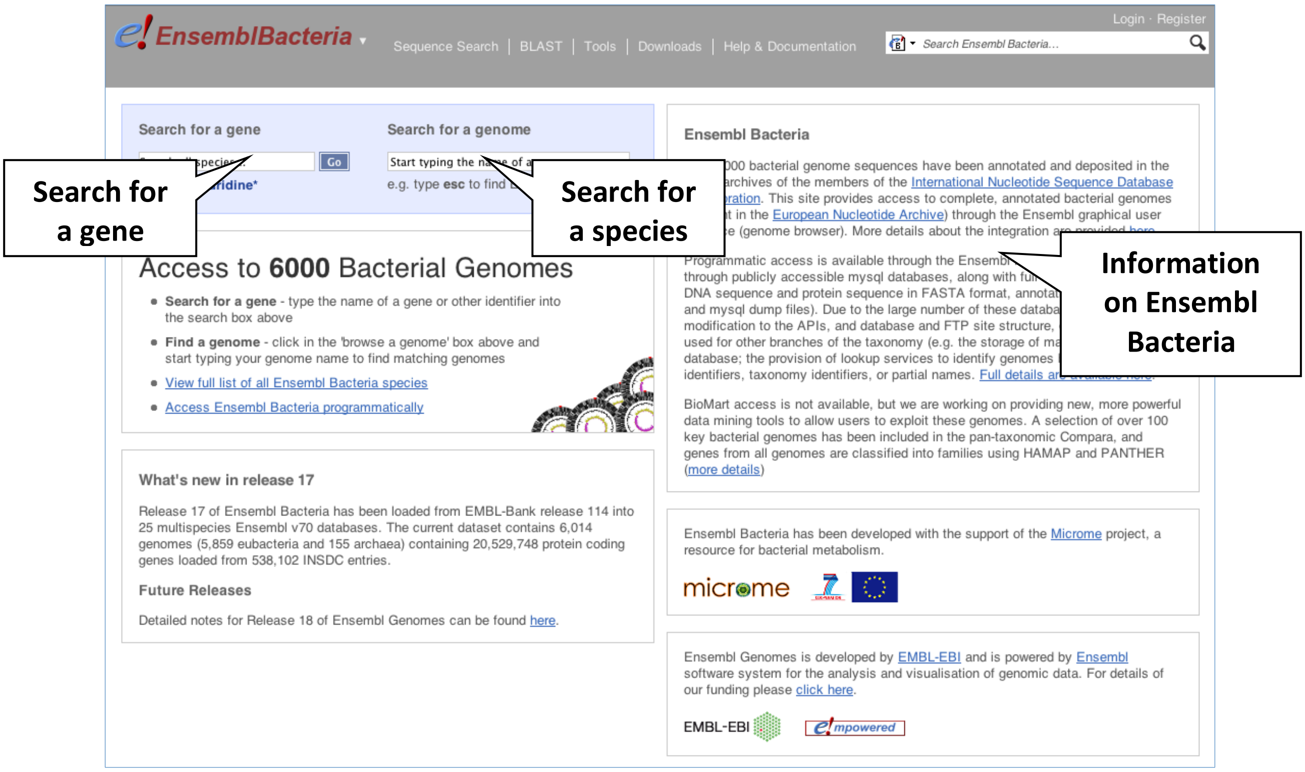
BioMart access is not available, but we are working on providing new, more powerful data mining tools to allow users to exploit these genomes. A selection of over 100 key bacterial genomes has been included in the pan-taxonomic Compara, and genes from all genomes are classified into families using HAMAP and PANTHER (more details)

Ensembl Bacteria has been developed with the support of the Microme project, a resource for bacterial metabolism.

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please click here.

Bacteria

You can navigate most of the taxa in the same way as you would with Ensembl, but Ensembl Bacteria has a large number of genomes, so needs slightly different methods. Let's look at it in more detail.



Search for a gene
Search for a species
Information on Ensembl Bacteria
EnsemblBacteria
Sequence Search | BLAST | Tools | Downloads | Help & Documentation
Login · Register
Search Ensembl Bacteria…

Search for a gene

Search all species... Go
e.g. ftsZ or uridine*

Search for a genome

Start typing the name of a genome...
e.g. type esc to find Escherichia

Access to **6000** Bacterial Genomes

- **Search for a gene** - type the name of a gene or other identifier into the search box above
- **Find a genome** - click in the 'browse a genome' box above and start typing your genome name to find matching genomes
- View full list of all Ensembl Bacteria species
- Access Ensembl Bacteria programmatically

What's new in release 17

Release 17 of Ensembl Bacteria has been loaded from EMBL-Bank release 114 into 25 multispecies Ensembl v70 databases. The current dataset contains 6,014 genomes (5,859 eubacteria and 155 archaea) containing 20,529,748 protein coding genes loaded from 538,102 INSDC entries.

Future Releases

Detailed notes for Release 18 of Ensembl Genomes can be found here.

Ensembl Bacteria

Over 6000 bacterial genome sequences have been annotated and deposited in the public archives of the members of the International Nucleotide Sequence Database Collaboration. This site provides access to complete, annotated bacterial genomes (present in the European Nucleotide Archive) through the Ensembl graphical user interface (genome browser). More details about the integration are provided here

Programmatic access is available through the Ensembl Perl and REST-ful APIs and through publicly accessible mysql databases, along with full data dumps (including DNA sequence and protein sequence in FASTA format, annotations in GTF format, and mysql dump files). Due to the large number of these databases, there is some modification to the APIs, and database and FTP site structure, compared to that used for other branches of the taxonomy (e.g. the storage of many genomes in one database; the provision of lookup services to identify genomes by INSDC identifiers, taxonomy identifiers, or partial names. Full details are available here.

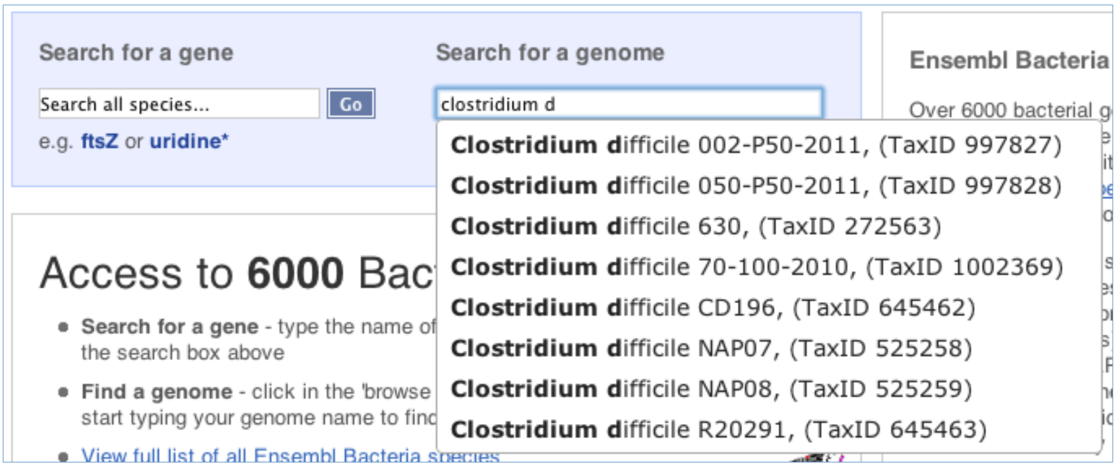
BioMart access is not available, but we are working on providing new, more powerful data mining tools to allow users to exploit these genomes. A selection of over 100 key bacterial genomes has been included in the pan-taxonomic Compara, and genes from all genomes are classified into families using HAMAP and PANTHER (more details)

Ensembl Bacteria has been developed with the support of the Microme project, a resource for bacterial metabolism.

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please click here.

There's no full species list for bacteria as it would be hard to navigate with the number of species. To find a species, start to type the species name into the species search box. A drop down list will appear with possible species.

For example, to find a substrain of *Clostridium difficile* type in Clostridium d.



Search for a gene

Search all species... Go

e.g. **ftsZ** or **uridine***

Search for a genome

clostridium d

- **Clostridium d**ifficile 002-P50-2011, (TaxID 997827)
- **Clostridium d**ifficile 050-P50-2011, (TaxID 997828)
- **Clostridium d**ifficile 630, (TaxID 272563)
- **Clostridium d**ifficile 70-100-2010, (TaxID 1002369)
- **Clostridium d**ifficile CD196, (TaxID 645462)
- **Clostridium d**ifficile NAP07, (TaxID 525258)
- **Clostridium d**ifficile NAP08, (TaxID 525259)
- **Clostridium d**ifficile R20291, (TaxID 645463)

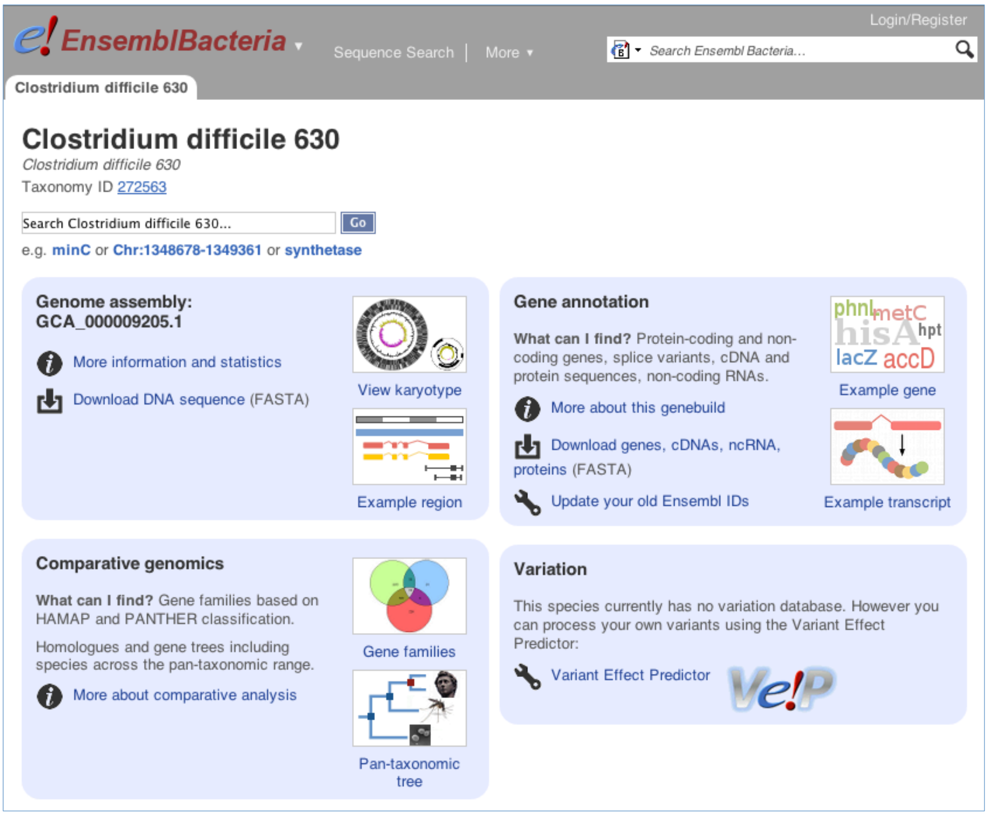
Ensembl Bacteria

Over 6000 bacterial g

Access to **6000** Bac

- **Search for a gene** - type the name of the search box above
- **Find a genome** - click in the 'browse start typing your genome name to find
- View full list of all Ensembl Bacteria species

The drop down contains various strains of *Clostridium difficile*. Let's choose Clostridium difficile 630. This will take us to another species homepage, where we can explore various features.



Clostridium difficile 630

Clostridium difficile 630

Taxonomy ID 272563

Search Clostridium difficile 630... Go

e.g. **minC** or **Chr:1348678-1349361** or **synthetase**

Genome assembly: GCA_000009205.1

More information and statistics

Download DNA sequence (FASTA)

View karyotype

Example region

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

Example gene

Example transcript

Comparative genomics

What can I find? Gene families based on HAMAP and PANTHER classification.

Homologues and gene trees including species across the pan-taxonomic range.

More about comparative analysis

Gene families

Pan-taxonomic tree

Variation

This species currently has no variation database. However you can process your own variants using the Variant Effect Predictor:

Variant Effect Predictor

Exercises: Ensembl species

Exercise 1 – Panda

(a) Go to the species homepage for Panda. What is the name of the genome assembly for Panda?

(b) Click on More information and statistics. How long is the Panda genome (in bp)? How many genes have been annotated?

Exercise 2 – Zebrafish

(a) What's new in release 71 for zebrafish?

(b) What previous release is available for zebrafish?

Exercise 3 – Mosquitos

(a) Go to Ensembl Metazoa. How many species of the genus *Anopheles* are there?

(b) Who published the genome sequence for *Anopheles gambiae*?

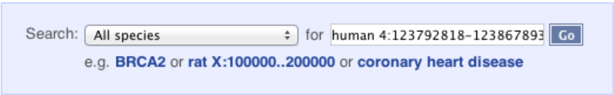
Exercise 4 – Bacteria

Go to Ensembl Bacteria and find the species *Belliella baltica*. How many coding and non-coding genes does it have?

Demo: The Region in detail view

Start at the Ensembl front page, ensembl.org. You can search for a region by typing it into a search box, but you have to specify the species.

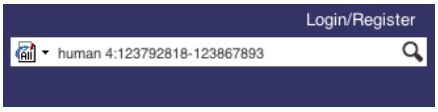
Type (or copy and paste) human 4:123792818-123867893 into either search box.



Search: All species for human 4:123792818-123867893 Go

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

or



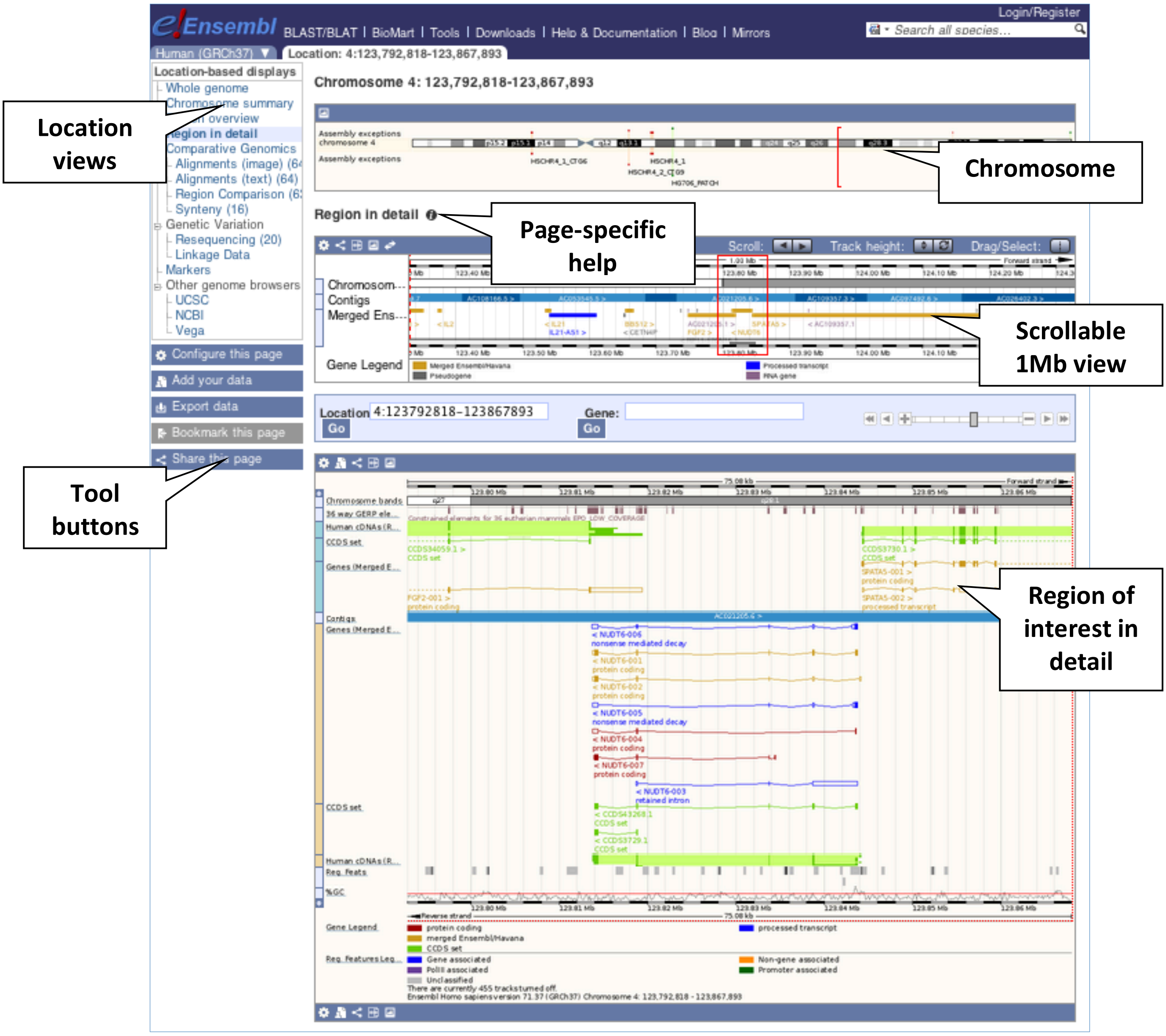
human 4:123792818-123867893

Press Enter or click Go to jump directly to the **Region in detail** Page.

Click on the button ⓘ to view page-specific help.

The help pages provide links to Frequently Asked Questions, a Glossary, Video Tutorials, and a form to Contact HelpDesk.

There is a help video on this page at http://youtu.be/tTKEvgPUq94.



Location: 4:123,792,818-123,867,893

Chromosome 4: 123,792,818-123,867,893

Location 4:123792818-123867893 Go

There are currently 455 tracks turned off.

Ensembl Homo sapiens version 71.37 (GRCh37) Chromosome 4: 123,792,818 - 123,867,893

Location views

Chromosome

Page-specific help

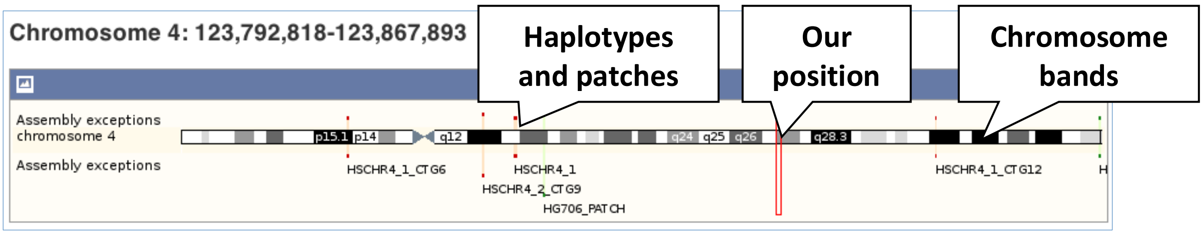
Scrollable 1Mb view

Tool buttons

Region of interest in detail

The Region in detail page is made up of three images, let's look at each one on detail.

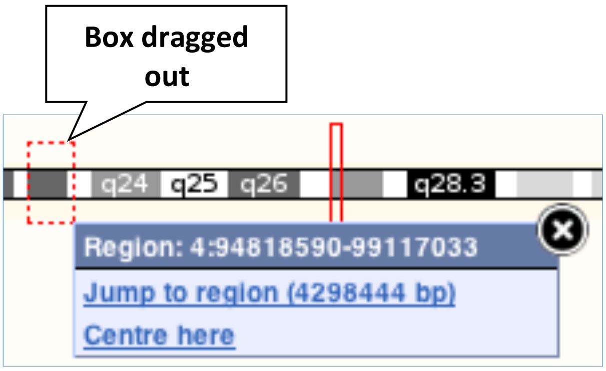
The first image shows the chromosome:


Haplotypes and patches

Our position

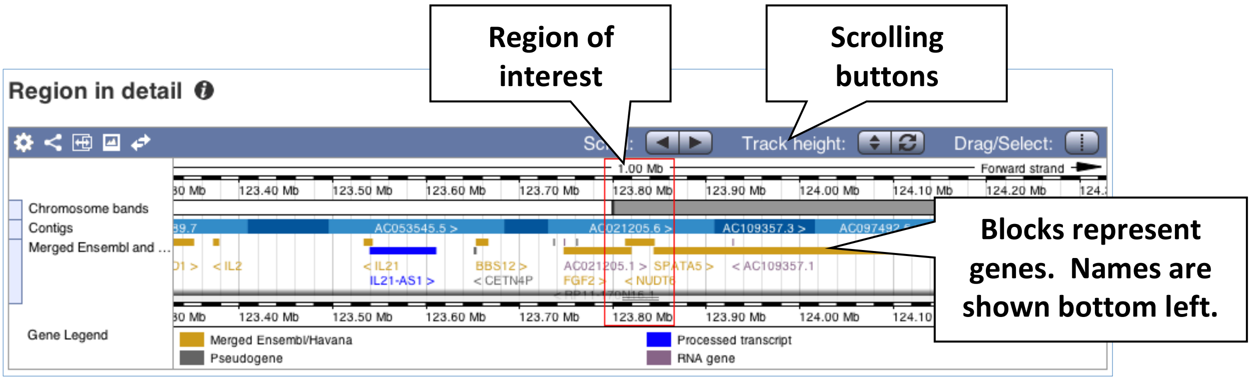
Chromosome bands

You can jump to a different region by dragging out a box in this image. Drag out a box on the chromosome, a pop-up menu will appear.


Box dragged out

If you wanted to move to the region, you could click on Jump to region (###bp). For now, we'll close the pop-up by clicking on the X on the corner.

The second image shows a 1Mb region around our selected region. This view allows you to scroll back and forth along the chromosome.

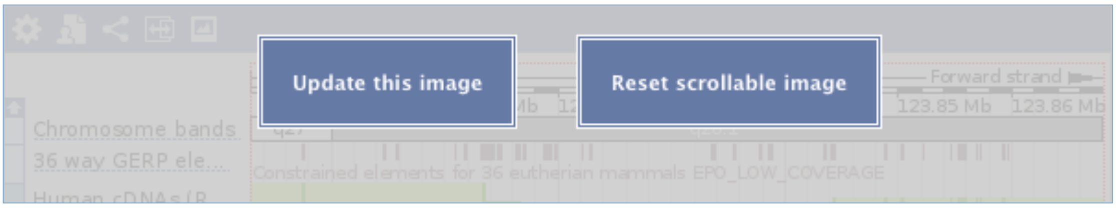

Region of interest

Scrolling buttons

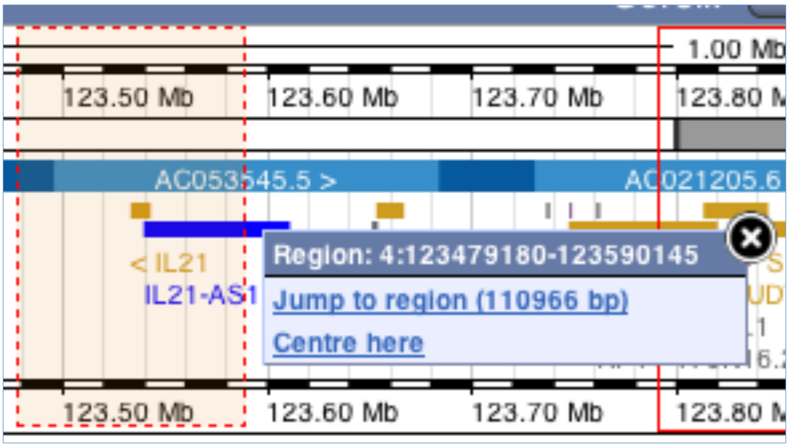
Blocks represent genes. Names are shown bottom left.

At the moment the gene track is set to a fixed height. Click on the Automatic track height button ⬍ to expand the image to include all possible data in the track.

Scroll along the chromosome by clicking and dragging within the image. As you do this you'll see the image below grey out and two blue buttons appear. Clicking on Update this image would jump the lower image to the region central to the scrollable image. We want to go back to where we started, so we'll click on Reset scrollable image.

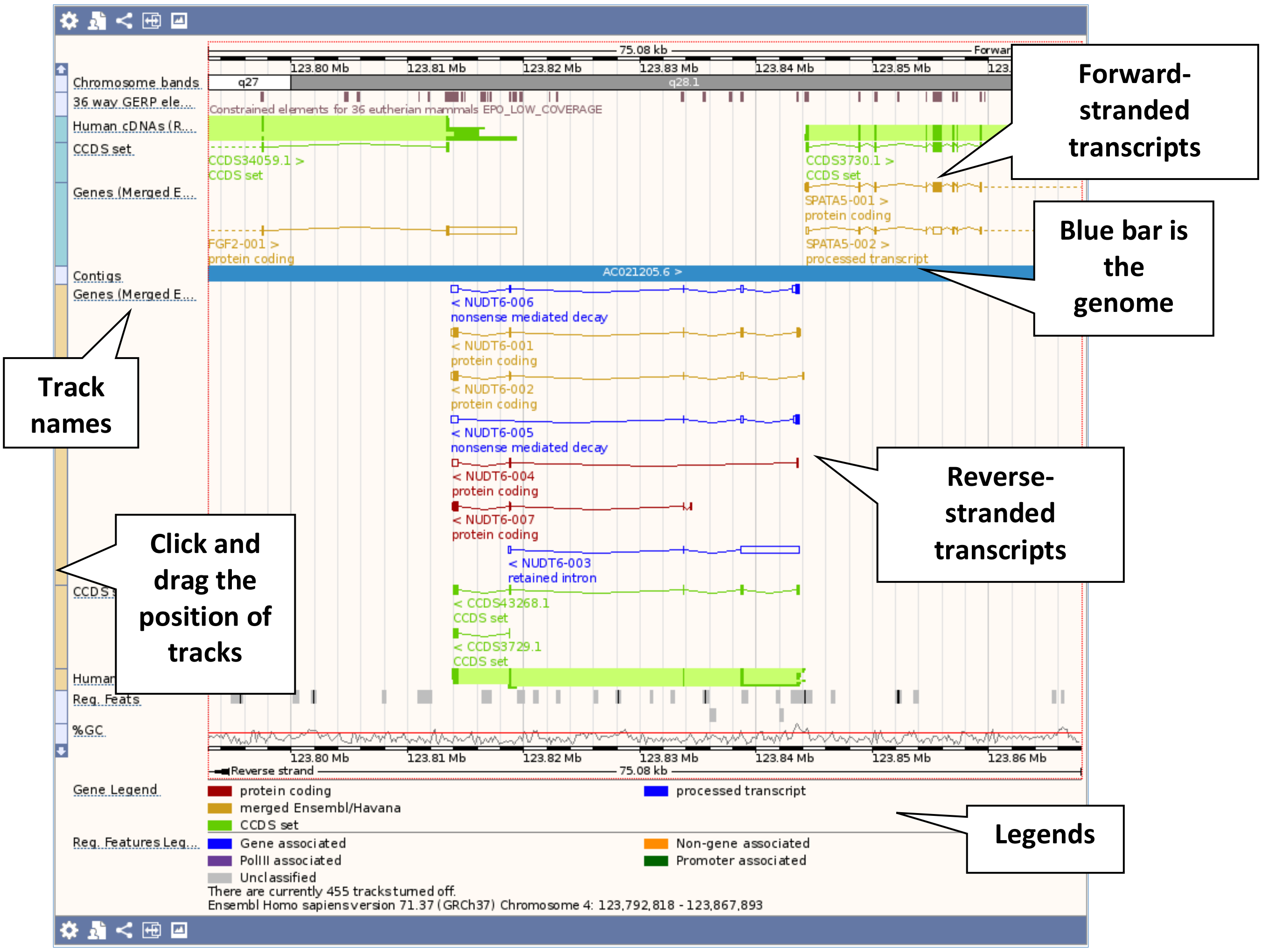


You can also drag out and jump to a region. Either hold down shift and drag in the image, or click on the Drag/Select button ↔ to change the action of your mouse click, and drag out a box.



Click on the X to close the pop-up menu.

The third image is a detailed, configurable view of the region.



Gene Legend
🟥 protein coding
🟧 merged Ensembl/Havana
🟩 CCDS set
🟦 processed transcript

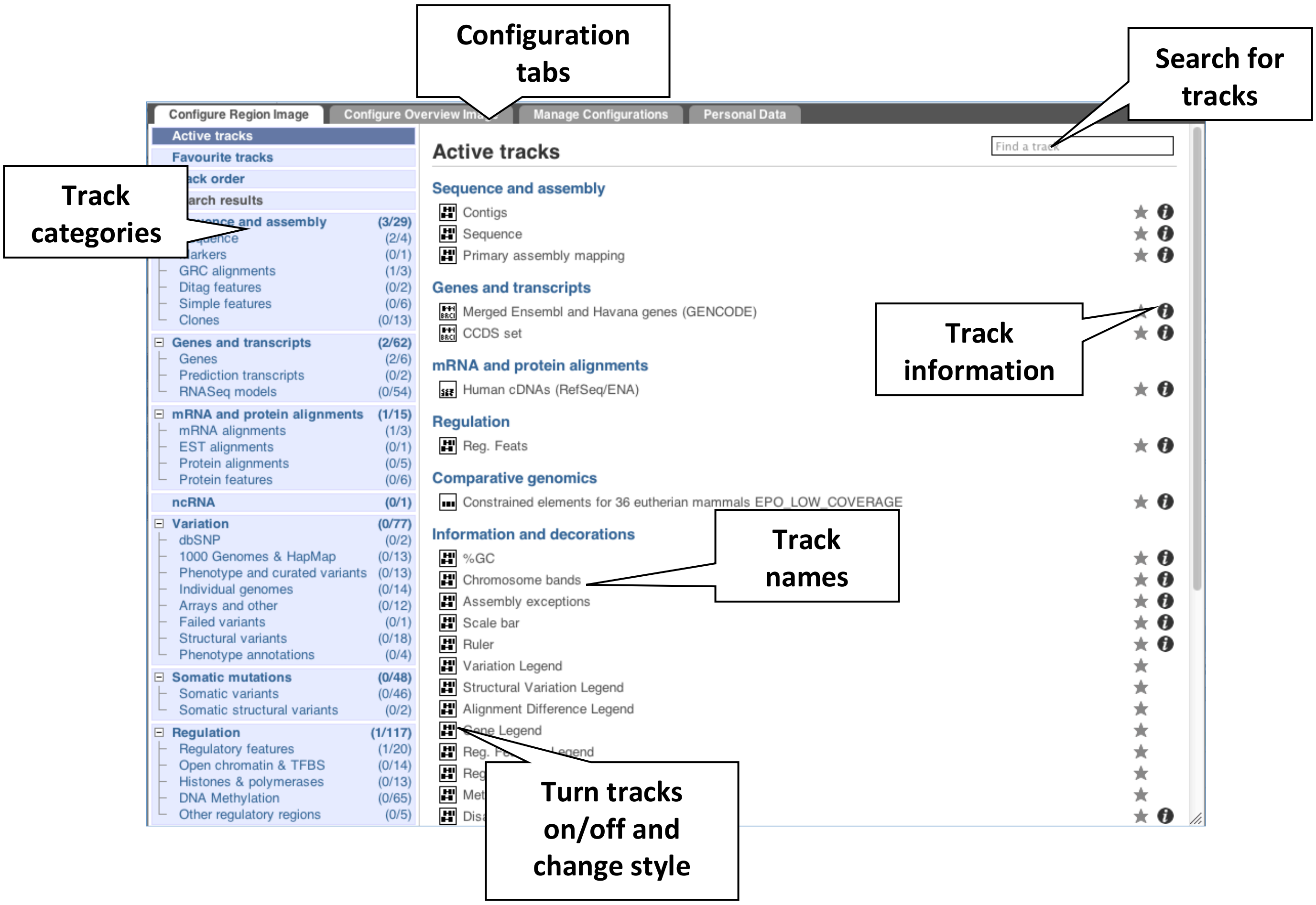
Reg. Features Leg...
🟦 Gene associated
🟪 PolII associated
⬜ Unclassified
🟧 Non-gene associated
🟩 Promoter associated

There are currently 455 tracks turned off.
Ensembl Homo sapiens version 71.37 (GRCh37) Chromosome 4: 123,792,818 - 123,867,893

We can edit what we see on this page by clicking on the blue Configure this page menu at the left.

This will open a menu that allows you to change the image.

You can put some tracks on in different styles; more details are in this FAQ: http://www.ensembl.org/Help/Faq?id=335.



Configuration tabs

Search for tracks

Track categories

Track information

Track names

Turn tracks on/off and change style

Active tracks
Favourite tracks
ack order
arch results

ence and assembly	(3/29)
quence	(2/4)
arkers	(0/1)
GRC alignments	(1/3)
Ditag features	(0/2)
Simple features	(0/6)
Clones	(0/13)
Genes and transcripts	(2/62)
Genes	(2/6)
Prediction transcripts	(0/2)
RNASeq models	(0/54)
mRNA and protein alignments	(1/15)
mRNA alignments	(1/3)
EST alignments	(0/1)
Protein alignments	(0/5)
Protein features	(0/6)
ncRNA	(0/1)
Variation	(0/77)
dbSNP	(0/2)
1000 Genomes & HapMap	(0/13)
Phenotype and curated variants	(0/13)
Individual genomes	(0/14)
Arrays and other	(0/12)
Failed variants	(0/1)
Structural variants	(0/18)
Phenotype annotations	(0/4)
Somatic mutations	(0/48)
Somatic variants	(0/46)
Somatic structural variants	(0/2)
Regulation	(1/117)
Regulatory features	(1/20)
Open chromatin & TFBS	(0/14)
Histones & polymerases	(0/13)
DNA Methylation	(0/65)
Other regulatory regions	(0/5)

Active tracks

Sequence and assembly
- Contigs
- Sequence
- Primary assembly mapping

Genes and transcripts
- Merged Ensembl and Havana genes (GENCODE)
- CCDS set

mRNA and protein alignments
- Human cDNAs (RefSeq/ENA)

Regulation
- Reg. Feats

Comparative genomics
- Constrained elements for 36 eutherian mammals EPO_LOW_COVERAGE

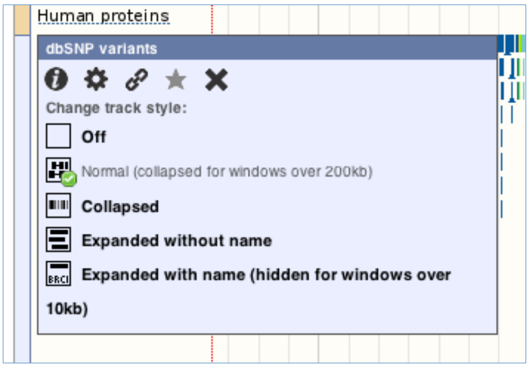
Information and decorations
- %GC
- Chromosome bands
- Assembly exceptions
- Scale bar
- Ruler
- Variation Legend
- Structural Variation Legend
- Alignment Difference Legend
- Gene Legend
- Reg. Fe Legend
- Reg
- Met
- Dis

Let's add some tracks to this image. Add:

- Human proteins – Labels
- dbSNP variants – Normal
- 1000 Genomes – AMR – Collapsed

Now click on the tick in the top left hand to save and close the menu. Alternatively, click anywhere outside of the menu. We can now see the tracks in the image.

We can also change the way the tracks appear by hovering over the track name then the cog wheel to open a menu. We can move tracks around by clicking and dragging on the bar to the left of the track name.



Human proteins

dbSNP variants

Change track style:
- □ Off
- Normal (collapsed for windows over 200kb) ✓
- Collapsed
- Expanded without name
- Expanded with name (hidden for windows over 10kb)

Now that you've got the view how you want it, you might like to show something you've found to a colleague or collaborator. Click on the Share this page button to generate a link. Email the link to someone else, so that they can see the same view as you, including all the tracks you've added. These links contain the Ensembl release number, so if a new release or even assembly comes out, your link will just take you to the archive site for the release it was made on.

To return this to the default view, go to Configure this page and select Reset configuration at the bottom of the menu.

Exercises: The Region in Detail view

Exercise 5 – Exploring a genomic region in human

(a) Go to the region from 32,448,000 to 33,198,000 bp on human chromosome 13. On which cytogenetic band is this region located? How many contigs make up this portion of the assembly (contigs are contiguous stretches of DNA sequence that have been assembled solely based on direct sequencing information)?

(b) Zoom in on the *BRCA2* gene.

(c) Are there any Tilepath clones that contain the complete *BRCA2* gene?

(d) Create a Share link for this display. Email it to yourself and open the link.

(e) Export the genomic sequence of the region you are looking at in FASTA format.

(f) Turn off all tracks you added to the Region in detail page.

Exercise 6 – Exploring patches and haplotypes in human

(a) Go to the region 6:112294691-112624977 in human. What is the green highlighted region? (Tip: you can search for help terms in the Ensembl search boxes.)

(b) Can you see the patches in the chromosome view? Drag out a box to jump to a region containing the patch labelled HG27_patch. What are the coordinates of the patch?

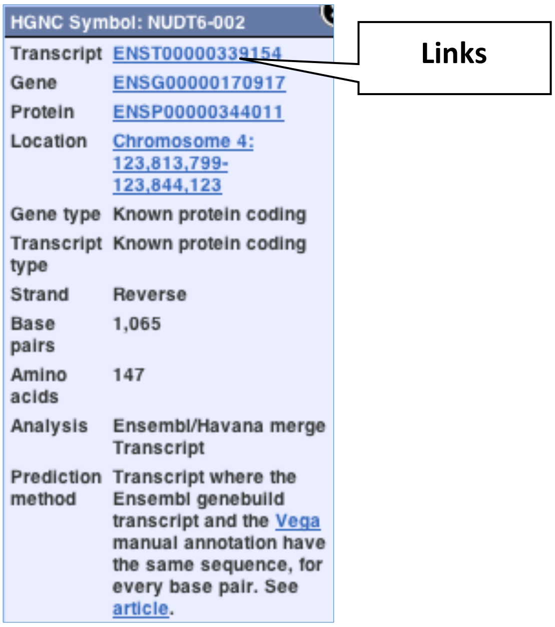
(c) Can you compare this patch with the reference? What has changed between this patch and the sequence it replaced?

(d) Go back to the previous view and scroll to the right in the 1Mb view until you reach a red highlighted region. What is this?

Genes and transcripts

Demo: The gene tab

If you click on any one of the transcripts in the Region in detail image, a pop-up menu will appear, allowing you to jump directly to that gene or transcript.

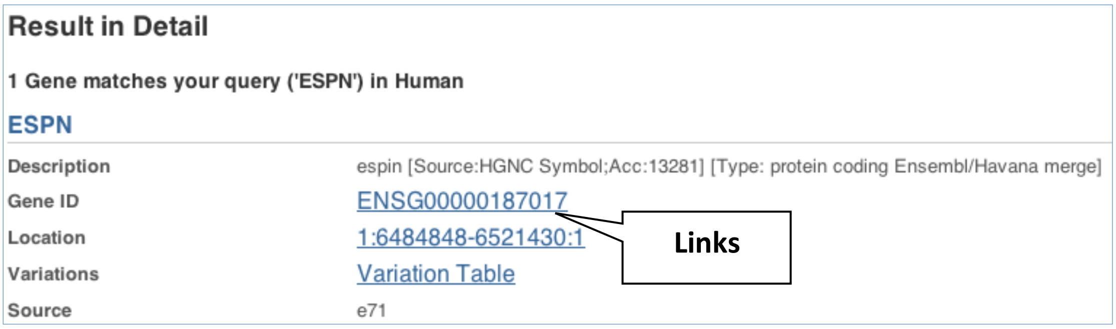

HGNC Symbol: NUDT6-002

Transcript	ENST00000339154
Gene	ENSG00000170917
Protein	ENSP00000344011
Location	Chromosome 4: 123,813,799-123,844,123
Gene type	Known protein coding
Transcript type	Known protein coding
Strand	Reverse
Base pairs	1,065
Amino acids	147
Analysis	Ensembl/Havana merge Transcript
Prediction method	Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See article.

Another way to go to a gene of interest is to search directly for it.

We're going to look at the human *ESPN* gene. This gene encodes a multifunctional actin-bundling protein with a major role in mediating sensory transduction in various mechanosensory and chemosensory cells. Mutations in this gene are associated with deafness (http://tinyurl.com/espn-ncbi-gene).

From ensembl.org, type *ESPN* into the search bar and click the Go button. Click on Gene and Human to find the hits.

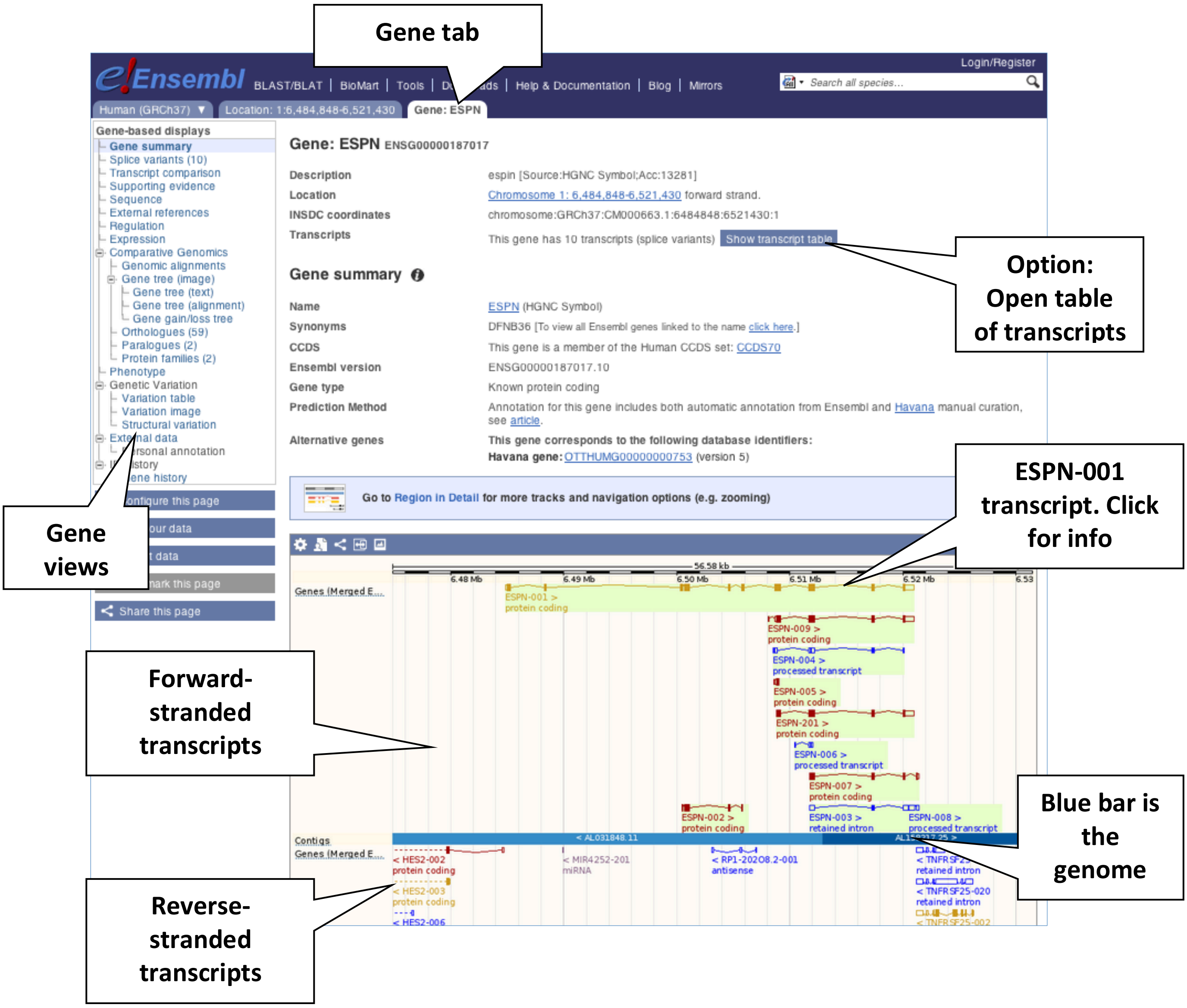

Result in Detail

1 Gene matches your query ('ESPN') in Human

ESPN

Description	espin [Source:HGNC Symbol;Acc:13281] [Type: protein coding Ensembl/Havana merge]
Gene ID	ENSG00000187017
Location	1:6484848-6521430:1
Variations	Variation Table
Source	e71

Click on the gene name or Ensembl ID. The **Gene tab** should open:



Gene tab

e!Ensembl BLAST/BLAT | BioMart | Tools | Do ads | Help & Documentation | Blog | Mirrors

Login/Register

Search all species...

Human (GRCh37) ▼ | Location: 1:6,484,848-6,521,430 | Gene: ESPN

Gene-based displays
- Gene summary
- Splice variants (10)
- Transcript comparison
- Supporting evidence
- Sequence
- External references
- Regulation
- Expression
- Comparative Genomics
  - Genomic alignments
  - Gene tree (image)
    - Gene tree (text)
    - Gene tree (alignment)
    - Gene gain/loss tree
  - Orthologues (59)
  - Paralogues (2)
  - Protein families (2)
- Phenotype
- Genetic Variation
  - Variation table
  - Variation image
  - Structural variation
- External data
  - rsonal annotation
- istory
  - ene history

ontigure this page

our data

t data

nark this page

Share this page

Gene: ESPN ENSG00000187017

Description	espin [Source:HGNC Symbol;Acc:13281]
Location	Chromosome 1: 6,484,848-6,521,430 forward strand.
INSDC coordinates	chromosome:GRCh37:CM000663.1:6484848:6521430:1
Transcripts	This gene has 10 transcripts (splice variants) Show transcript table

Gene summary

Name	ESPN (HGNC Symbol)
Synonyms	DFNB36 [To view all Ensembl genes linked to the name click here.]
CCDS	This gene is a member of the Human CCDS set: CCDS70
Ensembl version	ENSG00000187017.10
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:** OTTHUMG00000000753 (version 5)

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

Option: Open table of transcripts

ESPN-001 transcript. Click for info

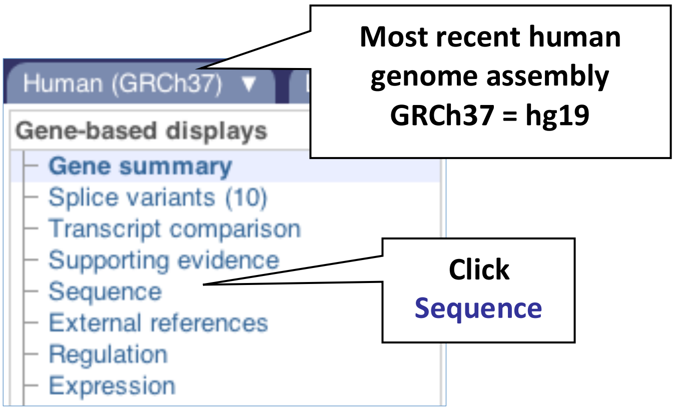
Gene views

Forward-stranded transcripts

Blue bar is the genome

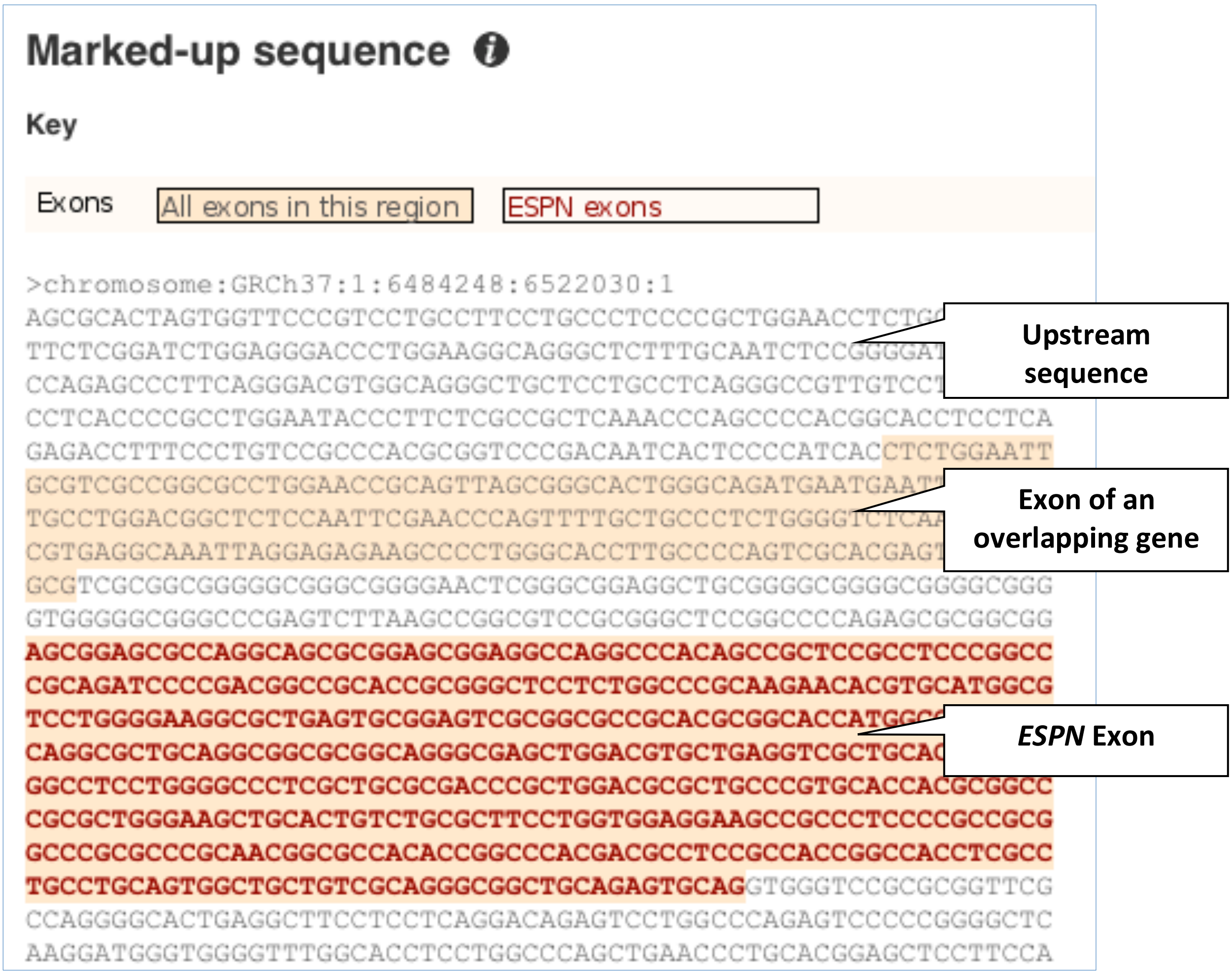
Reverse-stranded transcripts

Let's walk through some of the links in the left hand navigation column. How can we view the genomic sequence? Click Sequence at the left of the page.



Most recent human genome assembly GRCh37 = hg19

Click **Sequence**



Marked-up sequence

Key

Exons | All exons in this region | ESPN exons

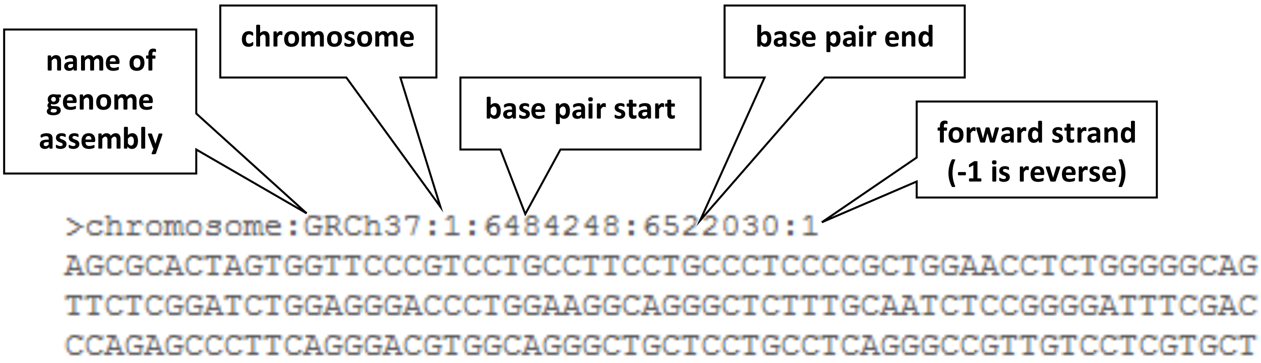
```
>chromosome:GRCh37:1:6484248:6522030:1
AGCGCACTAGTGGTTCCCGTCCTGCCTTCCTGCCCTCCCCGCTGGAACCTCTG
TTCTCGGATCTGGAGGGACCCTGGAAGGCAGGGCTCTTTGCAATCTCCGGGGA
CCAGAGCCCTTCAGGGACGTGGCAGGGCTGCTCCTGCCTCAGGGCCGTTGTCCT
CCTCACCCCGCCTGGAATACCCTTCTCGCCGCTCAAACCCAGCCCCACGGCACCTCCTCA
GAGACCTTTCCCTGTCCGCCCACGCGGTCCCGACAATCACTCCCCATCACCTCTGGAATT
GCGTCGCCGGCGCCTGGAACCGCAGTTAGCGGGCACTGGGCAGATGAATGAAT
TGCCTGGACGGCTCTCCAATTCGAACCCAGTTTTGCTGCCCTCTGGGGTCTCAA
CGTGAGGCAAATTAGGAGAGAAGCCCCTGGGCACCTTGCCCCAGTCGCACGAGT
GCGTCGCGGCGGGGGCGGGCGGGGAACTCGGGCGGAGGCTGCGGGGCGGGGCGGGGCGGG
GTGGGGGCGGGCCCGAGTCTTAAGCCGGCGTCCGCGGGCTCCGGCCCCAGAGCGCGGCGG
AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCC
CGCAGATCCCCGACGGCCGCACCGCGGGCTCCTCTGGCCCGCAAGAACACGTGCATGGCG
TCCTGGGGAAGGCGCTGAGTGCGGAGTCGCGGCGCCGCACGCGGCACCATGG
CAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCTGGACGTGCTGAGGTCGCTGCAC
GGCCTCCTGGGGCCCTCGCTGCGCGACCCGCTGGACGCGCTGCCCGTGCACCACGCGGCC
CGCGCTGGGAAGCTGCACTGTCTGCGCTTCCTGGTGGAGGAAGCCGCCCTCCCCGCCGCG
GCCCGCGCCCGCAACGGCGCCACACCGGCCCACGACGCCTCCGCCACCGGCCACCTCGCC
TGCCTGCAGTGGCTGCTGTCGCAGGGCGGCTGCAGAGTGCAGGTGGGTCCGCGCGGTTCG
CCAGGGGCACTGAGGCTTCCTCCTCAGGACAGAGTCCTGGCCCAGAGTCCCCCGGGCTC
AAGGATGGGTGGGGTTTGGCACCTCCTGGCCCAGCTGAACCCTGCACGGAGCTCCTTCCA
```

Upstream sequence

Exon of an overlapping gene

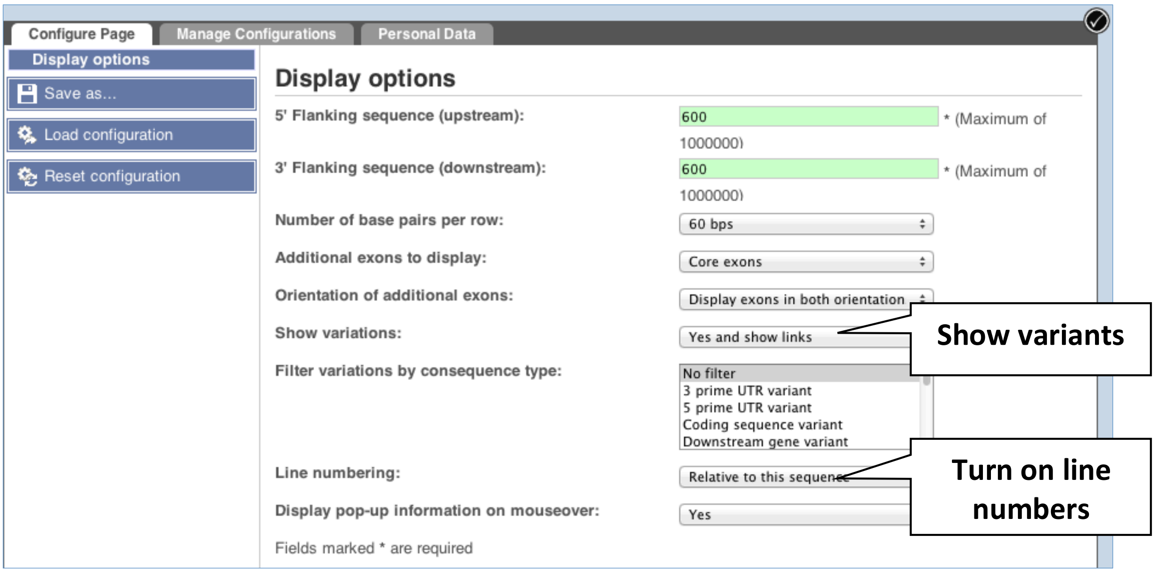
ESPN Exon

The sequence is shown in FASTA format. Take a look at the FASTA header:



```
>chromosome:GRCh37:1:6484248:6522030:1
AGCGCACTAGTGGTTCCCGTCCTGCCTTCCTGCCCTCCCCGCTGGAACCTCTGGGGGCAG
TTCTCGGATCTGGAGGGACCCTGGAAGGCAGGGCTCTTTGCAATCTCCGGGGATTTCGAC
CCAGAGCCCTTCAGGGACGTGGCAGGGCTGCTCCTGCCTCAGGGCCGTTGTCCTCGTGCT
```

Exons are highlighted within the genomic sequence. Variations can be added with the Configure this page link found at the left. Click on it now.



Display options

5' Flanking sequence (upstream): 600 * (Maximum of 1000000)

3' Flanking sequence (downstream): 600 * (Maximum of 1000000)

Number of base pairs per row: 60 bps

Additional exons to display: Core exons

Orientation of additional exons: Display exons in both orientation

Show variations: Yes and show links

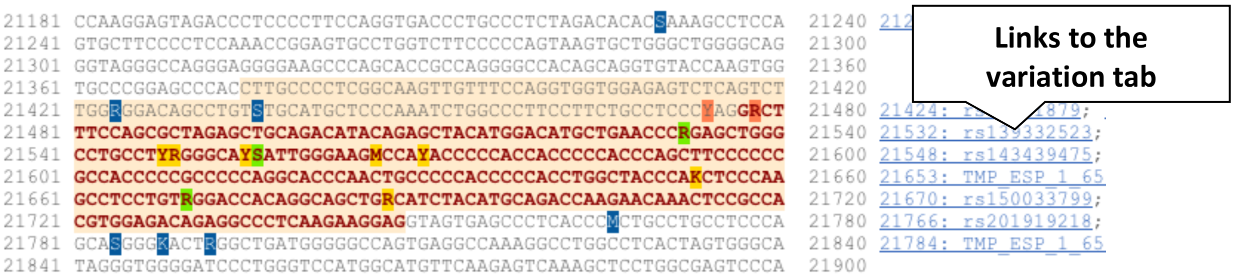
Filter variations by consequence type: No filter
3 prime UTR variant
5 prime UTR variant
Coding sequence variant
Downstream gene variant

Line numbering: Relative to this sequence

Display pop-up information on mouseover: Yes

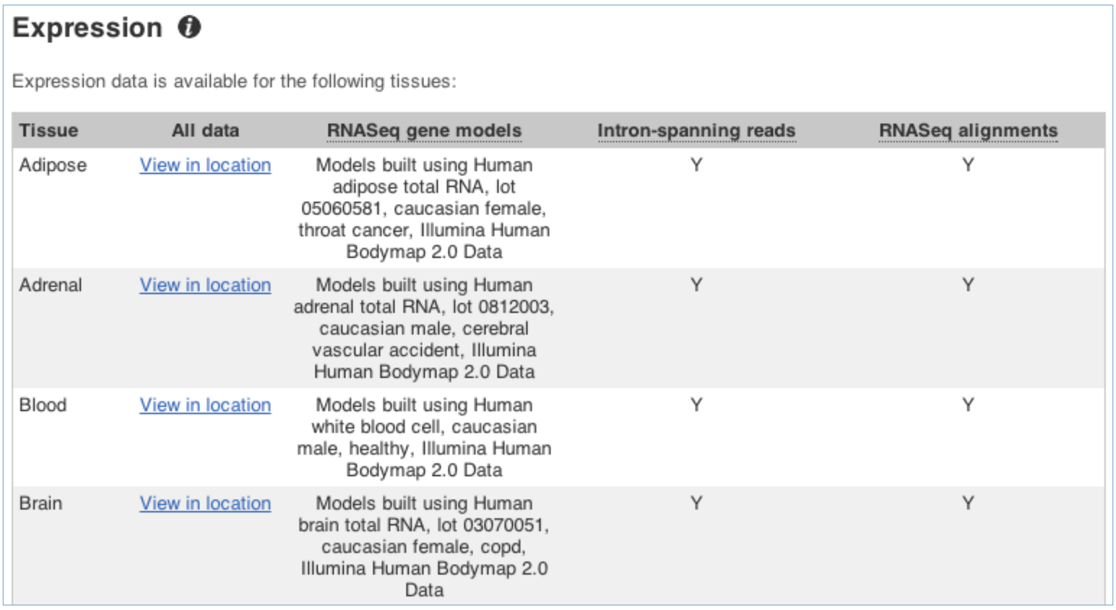
Fields marked * are required

Once you have selected changes (in this example, Show variations and Line numbering) click at the top right.



```
21181 CCAAGGAGTAGACCCTCCCCTTCCAGGTGACCCTGCCCTCTAGACACACSAAAGCCTCCA  21240 212
21241 GTGCTTCCCCTCCAAACCGGAGTGCCTGGTCTTCCCCCAGTAAGTGCTGGGCTGGGGCAG  21300
21301 GGTAGGGCCAGGGAGGGGAAGCCCAGCACCGCCAGGGGCCACAGCAGGTGTACCAAGTGG  21360
21361 TGCCCGGAGCCCACCTTGCCCCTCGGCAAGTTGTTTCCAGGTGGTGGAGAGTCTCAGTCT  21420
21421 TGGRGGACAGCCTGTSTGCATGCTCCCAAATCTGGCCCTTCCTTCTGCCTCCCYAGGRCT  21480 21424: rs 1879;
21481 TTCCAGCGCTAGAGCTGCAGACATACAGAGCTACATGGACATGCTGAACCCRGAGCTGGG  21540 21532: rs139332523;
21541 CCTGCCTYRGGGCAYSATTGGGAAGMCCAYACCCCCACCACCCCCACCCAGCTTCCCCCC  21600 21548: rs143439475;
21601 GCCACCCCCGCCCCCAGGCACCCAACTGCCCCCACCCCCACCTGGCTACCCAKCTCCCAA  21660 21653: TMP_ESP_1_65
21661 GCCTCCTGTRGGACCACAGGCAGCTGRCATCTACATGCAGACCAAGAACAAACTCCGCCA  21720 21670: rs150033799;
21721 CGTGGAGACAGAGGCCCTCAAGAAGGAGGTAGTGAGCCCTCACCCMCTGCCTGCCTCCCA  21780 21766: rs201919218;
21781 GCASGGGKACTRGGCTGATGGGGGCCAGTGAGGCCAAAGGCCTGGCCTCACTAGTGGGCA  21840 21784: TMP_ESP_1_65
21841 TAGGGTGGGGATCCCTGGGTCCATGGCATGTTCAAGAGTCAAAGCTCCTGGCGAGTCCCA  21900
```

Let’s look at where our gene is expressed. Click on Expression in the left-hand menu.



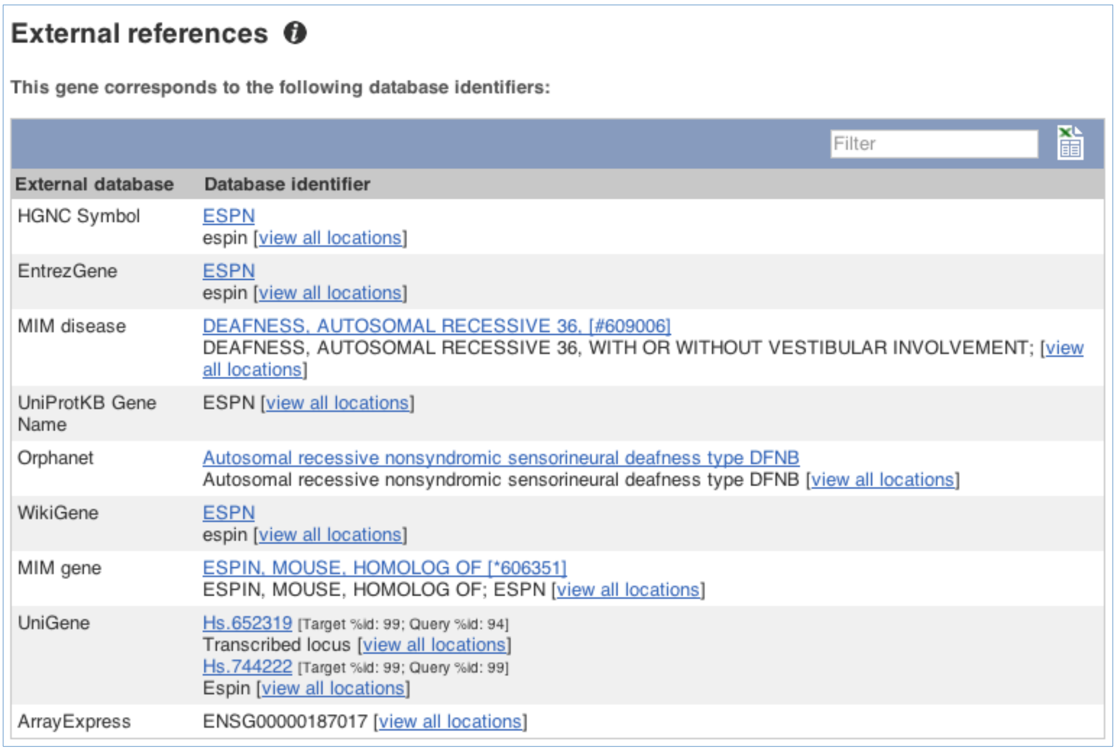
Expression

Expression data is available for the following tissues:

Tissue	All data	RNASeq gene models	Intron-spanning reads	RNASeq alignments
Adipose	View in location	Models built using Human adipose total RNA, lot 05060581, caucasian female, throat cancer, Illumina Human Bodymap 2.0 Data	Y	Y
Adrenal	View in location	Models built using Human adrenal total RNA, lot 0812003, caucasian male, cerebral vascular accident, Illumina Human Bodymap 2.0 Data	Y	Y
Blood	View in location	Models built using Human white blood cell, caucasian male, healthy, Illumina Human Bodymap 2.0 Data	Y	Y
Brain	View in location	Models built using Human brain total RNA, lot 03070051, caucasian female, copd, Illumina Human Bodymap 2.0 Data	Y	Y

Hover over the column titles for a pop-up definition.

Can our gene be found in other databases? Go up the left-hand menu to External references:



External references

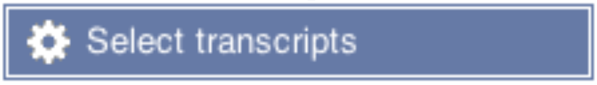
This gene corresponds to the following database identifiers:

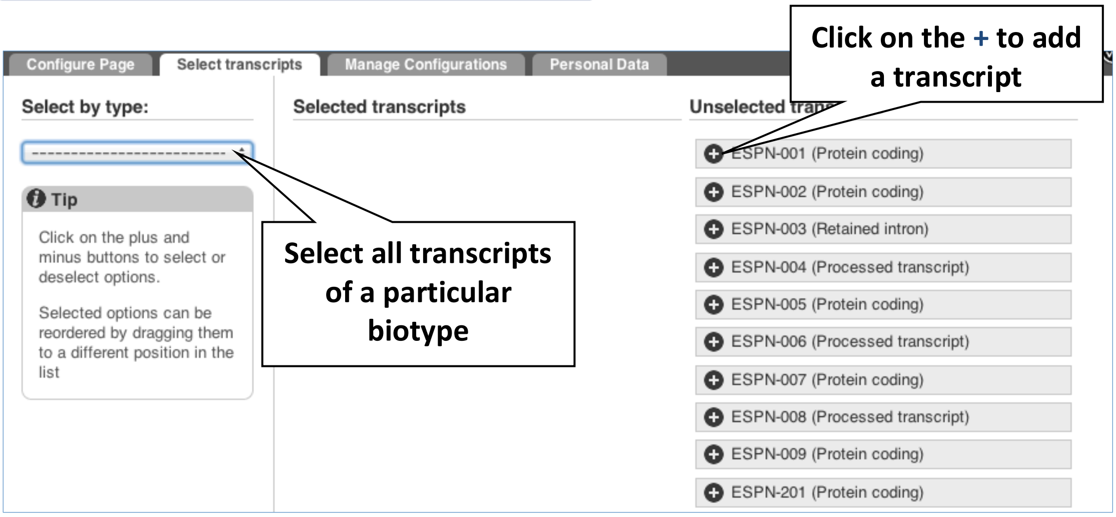
External database	Database identifier
HGNC Symbol	ESPN espin [view all locations]
EntrezGene	ESPN espin [view all locations]
MIM disease	DEAFNESS, AUTOSOMAL RECESSIVE 36, [#609006] DEAFNESS, AUTOSOMAL RECESSIVE 36, WITH OR WITHOUT VESTIBULAR INVOLVEMENT; [view all locations]
UniProtKB Gene Name	ESPN [view all locations]
Orphanet	Autosomal recessive nonsyndromic sensorineural deafness type DFNB Autosomal recessive nonsyndromic sensorineural deafness type DFNB [view all locations]
WikiGene	ESPN espin [view all locations]
MIM gene	ESPIN, MOUSE, HOMOLOG OF [*606351] ESPIN, MOUSE, HOMOLOG OF; ESPN [view all locations]
UniGene	Hs.652319 [Target %id: 99; Query %id: 94] Transcribed locus [view all locations] Hs.744222 [Target %id: 99; Query %id: 99] Espin [view all locations]
ArrayExpress	ENSG00000187017 [view all locations]

This contains links to the gene in other projects, such as Uniprot.

To find out more about the individual transcripts of this gene, click on Transcript comparison in the left-hand menu.

You must now choose the transcripts you'd like to see, click on the blue Select transcripts button.





Click on the plus and minus buttons to select or deselect options.

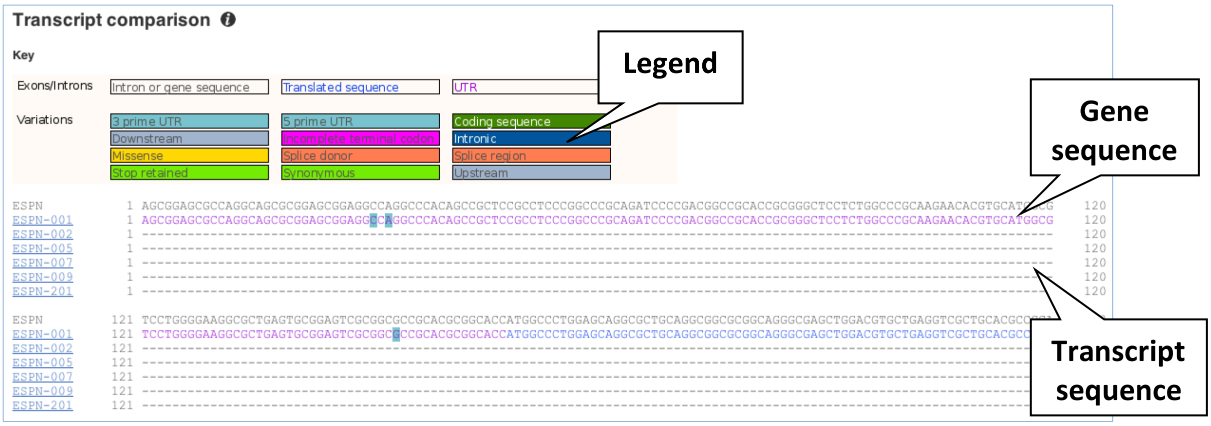
Selected options can be reordered by dragging them to a different position in the list

- ESPN-001 (Protein coding)
- ESPN-002 (Protein coding)
- ESPN-003 (Retained intron)
- ESPN-004 (Processed transcript)
- ESPN-005 (Protein coding)
- ESPN-006 (Processed transcript)
- ESPN-007 (Protein coding)
- ESPN-008 (Processed transcript)
- ESPN-009 (Protein coding)
- ESPN-201 (Protein coding)

Click on the + to add a transcript

Select all transcripts of a particular biotype

Let's select all the protein-coding transcripts, then close the menu.



```
ESPN       1 AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCCCGCAGATCCCCGACGGCCGCACCGCGGCTCCTCTGGCCCGCAAGAACACGTGCAT   120
ESPN-001   1 AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCCCGCAGATCCCCGACGGCCGCACCGCGGCTCCTCTGGCCCGCAAGAACACGTGCATGGCG 120
ESPN-002   1 ------------------------------------------------------------------------------------------------------------------------ 120
ESPN-005   1 ------------------------------------------------------------------------------------------------------------------------ 120
ESPN-007   1 ------------------------------------------------------------------------------------------------------------------------ 120
ESPN-009   1 ------------------------------------------------------------------------------------------------------------------------ 120
ESPN-201   1 ------------------------------------------------------------------------------------------------------------------------ 120

ESPN     121 TCCTGGGGAAGGCGCTGAGTGCGGAGTCGCGGCGCCGCACGCGGCACCATGGCCCTGGAGCAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCTGGACGTGCTGAGGTCGCTGCACGCC
ESPN-001 121 TCCTGGGGAAGGCGCTGAGTGCGGAGTCGCGGCGCCGCACGCGGCACCATGGCCCTGGAGCAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCTGGACGTGCTGAGGTCGCTGCACGCC
ESPN-002 121 --------------------------------------------------------------------------------------------------------------------
ESPN-005 121 --------------------------------------------------------------------------------------------------------------------
ESPN-007 121 --------------------------------------------------------------------------------------------------------------------
ESPN-009 121 --------------------------------------------------------------------------------------------------------------------
ESPN-201 121 --------------------------------------------------------------------------------------------------------------------
```

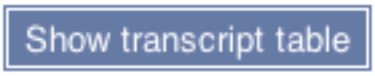
Legend

Gene sequence

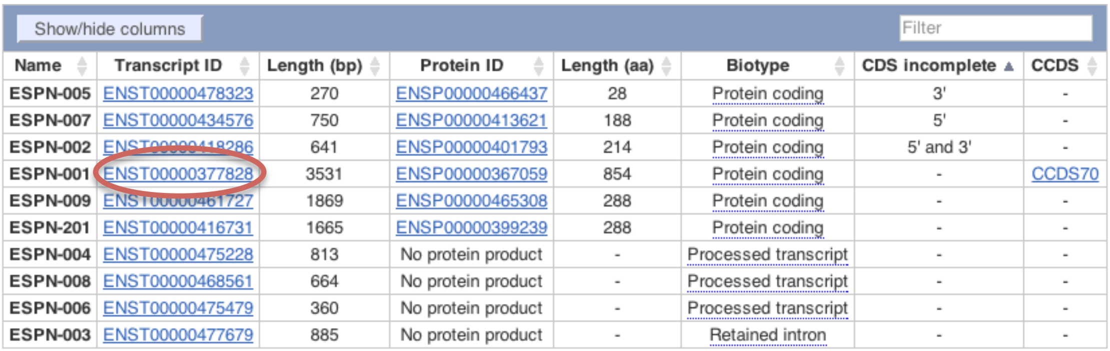
Transcript sequence

Demo: The transcript tab

Let's now explore one splice isoform. Click on Show transcript table at the top.



Click on the ID for the largest one, ESPN-001 (ENST00000377828).

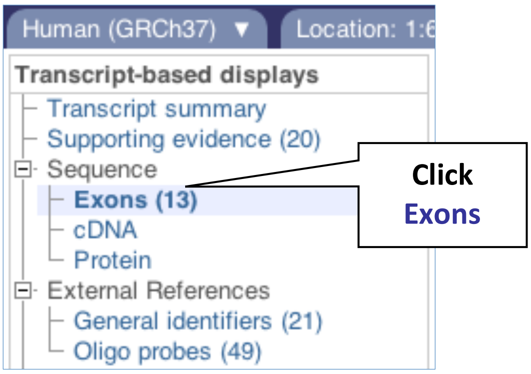


Show/hide columns

Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
ESPN-005	ENST00000478323	270	ENSP00000466437	28	Protein coding	3'	-
ESPN-007	ENST00000434576	750	ENSP00000413621	188	Protein coding	5'	-
ESPN-002	ENST00000418286	641	ENSP00000401793	214	Protein coding	5' and 3'	-
ESPN-001	ENST00000377828	3531	ENSP00000367059	854	Protein coding	-	CCDS70
ESPN-009	ENST00000461727	1869	ENSP00000465308	288	Protein coding	-	-
ESPN-201	ENST00000416731	1665	ENSP00000399239	288	Protein coding	-	-
ESPN-004	ENST00000475228	813	No protein product	-	Processed transcript	-	-
ESPN-008	ENST00000468561	664	No protein product	-	Processed transcript	-	-
ESPN-006	ENST00000475479	360	No protein product	-	Processed transcript	-	-
ESPN-003	ENST00000477679	885	No protein product	-	Retained intron	-	-

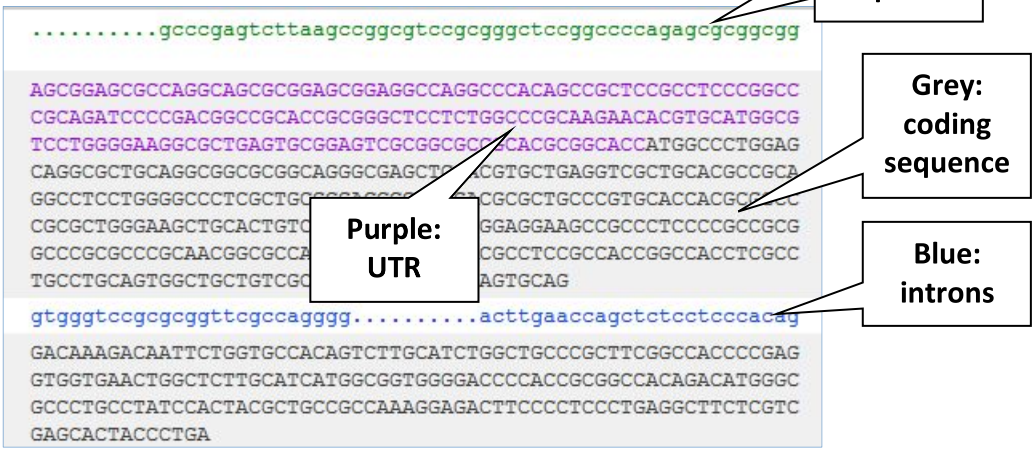
You are now in the Transcript tab for ESPN-001. The left hand navigation column provides several options for the transcript ESPN-001. Click on the Exons link, circled in the image below.



Transcript-based displays
- Transcript summary
- Supporting evidence (20)
- ⊟ Sequence
  - **Exons (13)**
  - cDNA
  - Protein
- ⊟ External References
  - General identifiers (21)
  - Oligo probes (49)

Click Exons

Green: flanking sequence



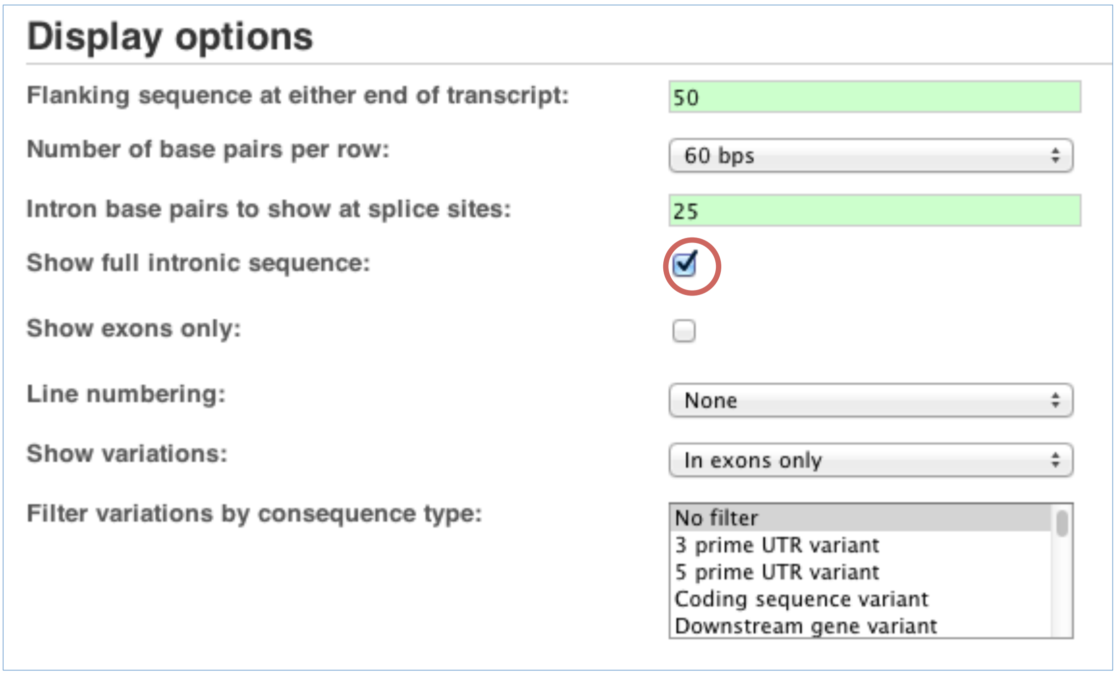
```
..........gcccgagtcttaagccggcgtccgcgggctccggccccagagcgcggcgg

AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCC
CGCAGATCCCCGACGGCCGCACCGCGGGCTCCTCTGGCCCGCAAGAACACGTGCATGGCG
TCCTGGGGAAGGCGCTGAGTGCGGAGTCGCGGCGCCGCACGCGGCACCATGGCCCTGGAG
CAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCT  ACGTGCTGAGGTCGCTGCACGCCGCA
GGCCTCCTGGGGCCCTCGCTG  CGCGCTGCCCGTGCACCACGC
CGCGCTGGGAAGCTGCACTGTC  GGAGGAAGCCGCCCTCCCCGCCGCG
GCCCGCGCCCGCAACGGCGCCA  CGCCTCCGCCACCGGCCACCTCGCC
TGCCTGCAGTGGCTGCTGTCGC  AGTGCAG
gtgggtccgcgcggttcgccagggg..........acttgaaccagctctcctcccacag
GACAAAGACAATTCTGGTGCCACAGTCTTGCATCTGGCTGCCCGCTTCGGCCACCCCGAG
GTGGTGAACTGGCTCTTGCATCATGGCGGTGGGGACCCCACCGCGGCCACAGACATGGGC
GCCCTGCCTATCCACTACGCTGCCGCCAAAGGAGACTTCCCCTCCCTGAGGCTTCTCGTC
GAGCACTACCCTGA
```

Purple: UTR

Grey: coding sequence

Blue: introns

You may want to change the display (for example, to show more flanking sequence, or to show full introns). In order to do so click on Configure this page and change the display options accordingly.



Display options

Flanking sequence at either end of transcript: 50

Number of base pairs per row: 60 bps

Intron base pairs to show at splice sites: 25

Show full intronic sequence: ☑

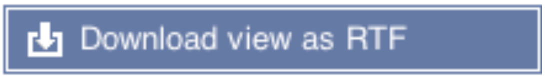
Show exons only: □

Line numbering: None

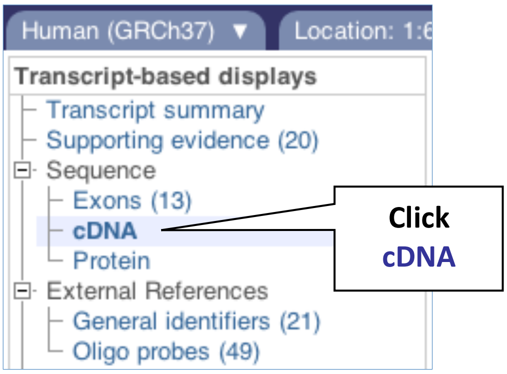
Show variations: In exons only

Filter variations by consequence type: No filter
3 prime UTR variant
5 prime UTR variant
Coding sequence variant
Downstream gene variant

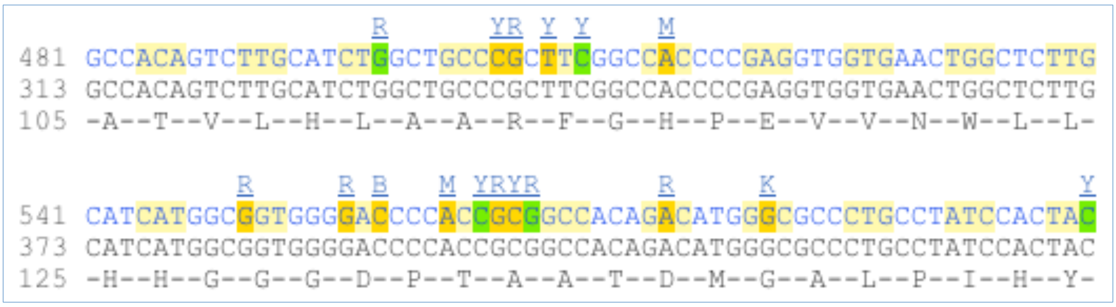
If you would like to export the sequence, including the colours, click Download view as RTF. A Rich Text Format document will be generated that can be opened in word processor such as MS Word.



Now click on the cDNA link to see the spliced transcript sequence.



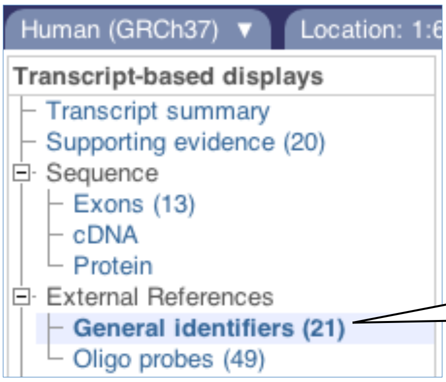
Click cDNA



```
                     R      YR Y Y    M
481 GCCACAGTCTTGCATCTGGCTGCCCGCTTCGGCCACCCCGAGGTGGTGAACTGGCTCTTG
313 GCCACAGTCTTGCATCTGGCTGCCCGCTTCGGCCACCCCGAGGTGGTGAACTGGCTCTTG
105 -A--T--V--L--H--L--A--A--R--F--G--H--P--E--V--V--N--W--L--L-

             R     R B   M YRYR       R     K                  Y
541 CATCATGGCGGTGGGGACCCCACCGCGGCCACAGACATGGGCGCCCTGCCTATCCACTAC
373 CATCATGGCGGTGGGGACCCCACCGCGGCCACAGACATGGGCGCCCTGCCTATCCACTAC
125 -H--H--G--G--G--D--P--T--A--A--T--D--M--G--A--L--P--I--H--Y-
```

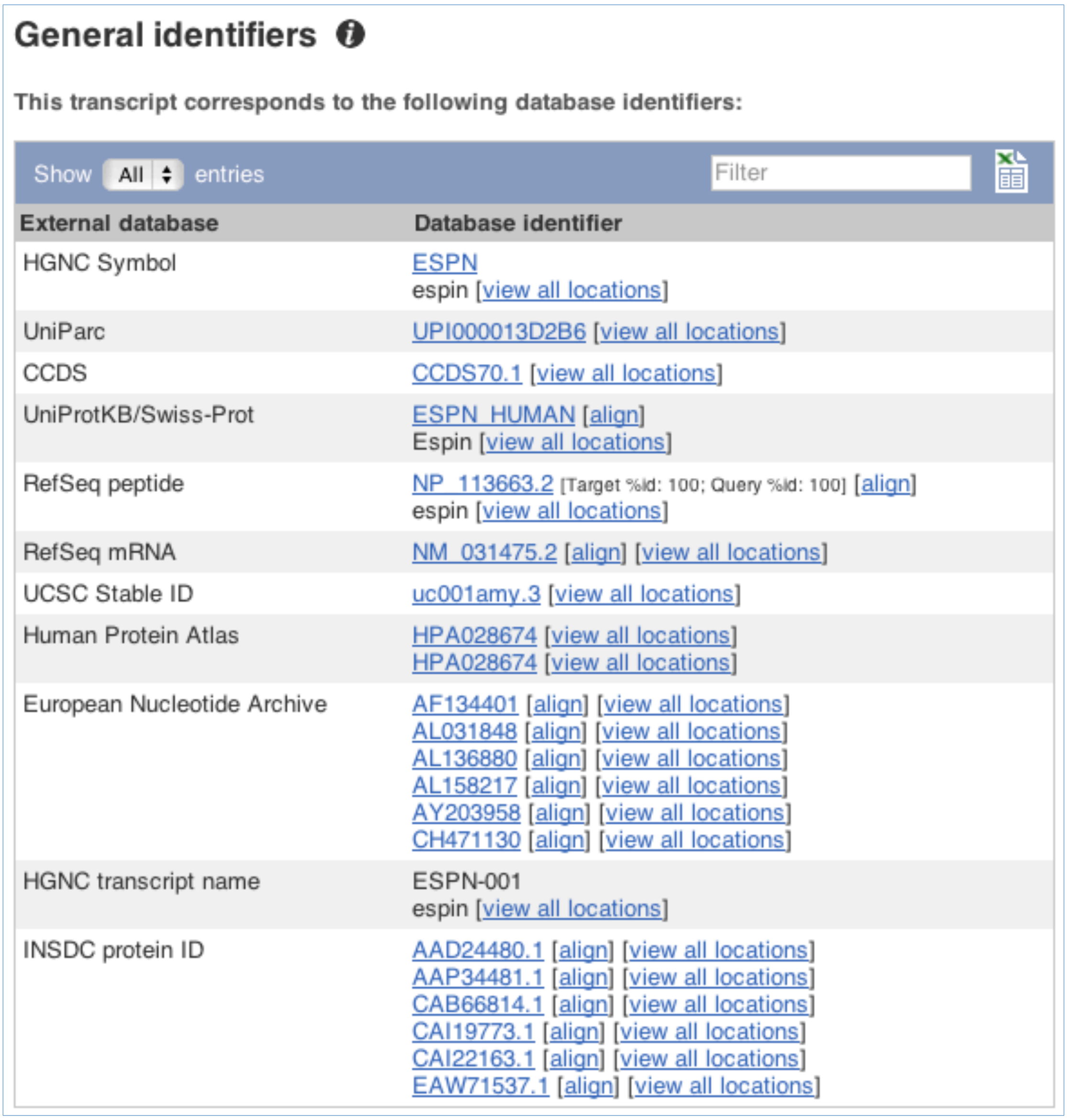
UnTranslated Regions (UTRs) are highlighted in dark yellow, codons are highlighted in light yellow, and exon sequence is shown in black or blue letters to show exon divides. Sequence variants are represented by highlighted nucleotides and clickable IUPAC codes are above the sequence.

Next, follow the General identifiers link at the left.



Click General Identifiers

This page shows information from other databases such as RefSeq, EntrezGene, OMIM, UniProtKB, and others, that matches to the Ensembl transcript and protein

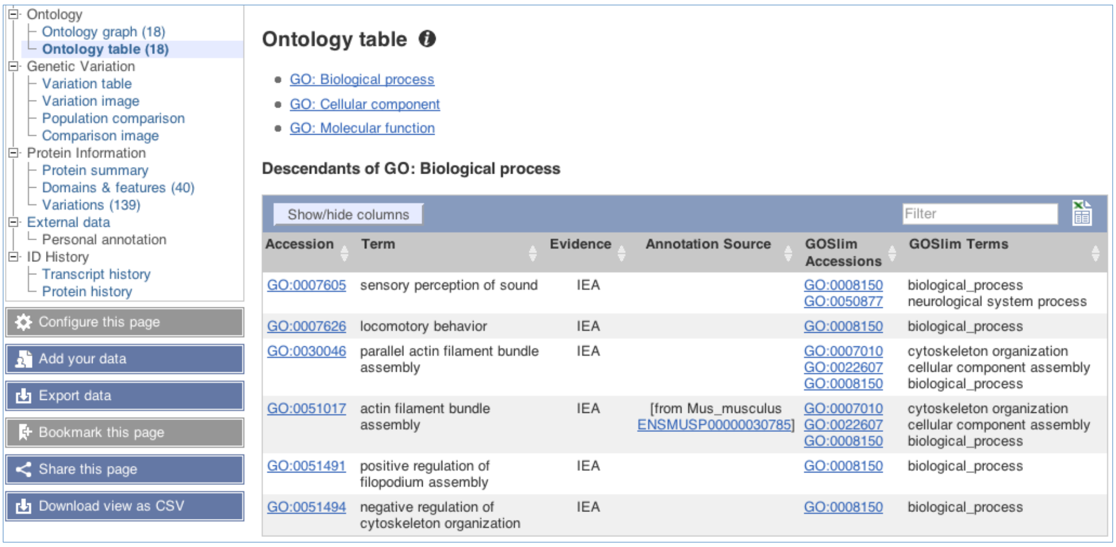


General identifiers

This transcript corresponds to the following database identifiers:

External database	Database identifier
HGNC Symbol	ESPN espin [view all locations]
UniParc	UPI000013D2B6 [view all locations]
CCDS	CCDS70.1 [view all locations]
UniProtKB/Swiss-Prot	ESPN_HUMAN [align] Espin [view all locations]
RefSeq peptide	NP_113663.2 [Target %id: 100; Query %id: 100] [align] espin [view all locations]
RefSeq mRNA	NM_031475.2 [align] [view all locations]
UCSC Stable ID	uc001amy.3 [view all locations]
Human Protein Atlas	HPA028674 [view all locations] HPA028674 [view all locations]
European Nucleotide Archive	AF134401 [align] [view all locations] AL031848 [align] [view all locations] AL136880 [align] [view all locations] AL158217 [align] [view all locations] AY203958 [align] [view all locations] CH471130 [align] [view all locations]
HGNC transcript name	ESPN-001 espin [view all locations]
INSDC protein ID	AAD24480.1 [align] [view all locations] AAP34481.1 [align] [view all locations] CAB66814.1 [align] [view all locations] CAI19773.1 [align] [view all locations] CAI22163.1 [align] [view all locations] EAW71537.1 [align] [view all locations]

Click on Ontology table to see GO terms from the Gene Ontology consortium. www.geneontology.org



- Ontology
  - Ontology graph (18)
  - **Ontology table (18)**
- Genetic Variation
  - Variation table
  - Variation image
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (40)
  - Variations (139)
- External data
  - Personal annotation
- ID History
  - Transcript history
  - Protein history

Ontology table

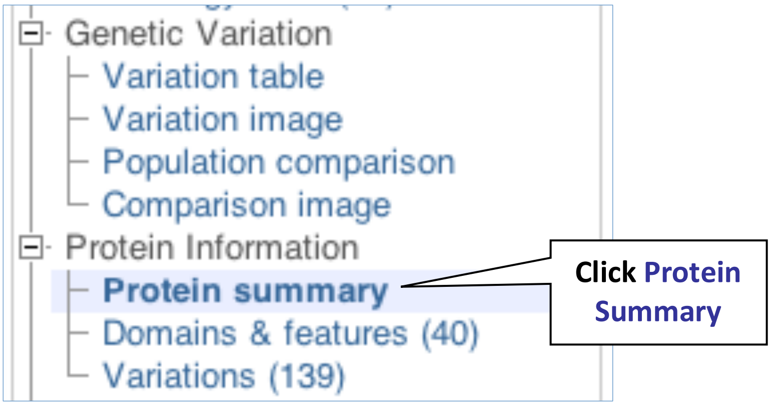
- GO: Biological process
- GO: Cellular component
- GO: Molecular function

Descendants of GO: Biological process

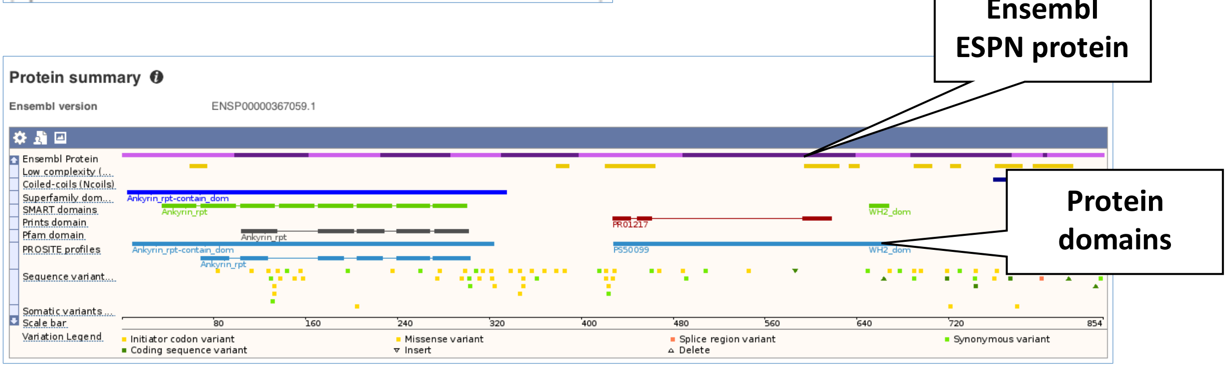
Accession	Term	Evidence	Annotation Source	GOSlim Accessions	GOSlim Terms
GO:0007605	sensory perception of sound	IEA		GO:0008150 GO:0050877	biological_process neurological system process
GO:0007626	locomotory behavior	IEA		GO:0008150	biological_process
GO:0030046	parallel actin filament bundle assembly	IEA		GO:0007010 GO:0022607 GO:0008150	cytoskeleton organization cellular component assembly biological_process
GO:0051017	actin filament bundle assembly	IEA	[from Mus_musculus ENSMUSP00000030785]	GO:0007010 GO:0022607 GO:0008150	cytoskeleton organization cellular component assembly biological_process
GO:0051491	positive regulation of filopodium assembly	IEA		GO:0008150	biological_process
GO:0051494	negative regulation of cytoskeleton organization	IEA		GO:0008150	biological_process

Click on the ⓘ to see a guide to the three-letter Evidence codes.

Now click on Protein summary to view domains from Pfam, PROSITE, Superfamily, InterPro, and more.



Click Protein Summary



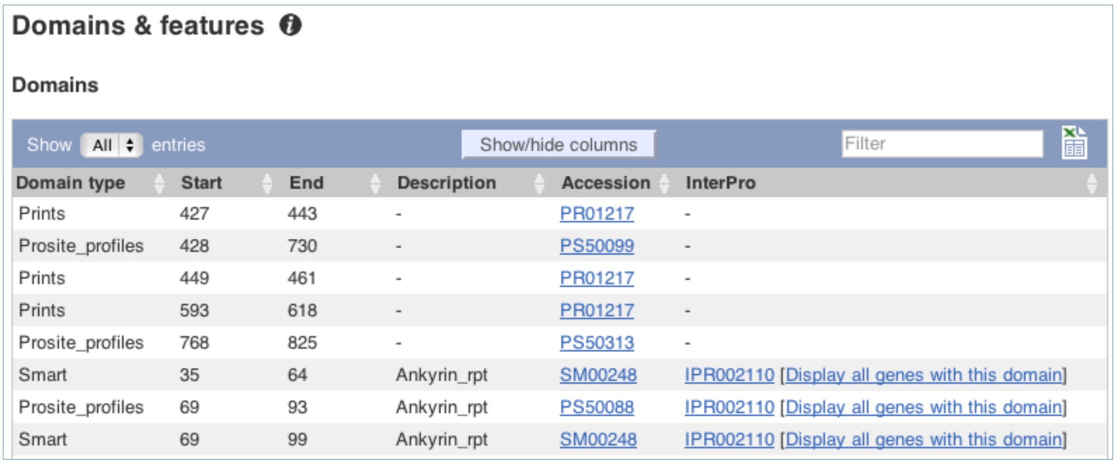
Ensembl ESPN protein

Protein domains

Protein summary ⓘ

Ensembl version ENSP00000367059.1

Clicking on Domains & features shows a table of this information.



Domains & features

Domains

Domain type	Start	End	Description	Accession	InterPro
Prints	427	443	-	PR01217	-
Prosite_profiles	428	730	-	PS50099	-
Prints	449	461	-	PR01217	-
Prints	593	618	-	PR01217	-
Prosite_profiles	768	825	-	PS50313	-
Smart	35	64	Ankyrin_rpt	SM00248	IPR002110 [Display all genes with this domain]
Prosite_profiles	69	93	Ankyrin_rpt	PS50088	IPR002110 [Display all genes with this domain]
Smart	69	99	Ankyrin_rpt	SM00248	IPR002110 [Display all genes with this domain]

Exercises: Genes and transcripts

Exercise 7 – Exploring the human *MYH9* gene

(a) Find the human *MYH9* (myosin, heavy chain 9, non-muscle) gene, and go to the Gene tab.

- On which chromosome and which strand of the genome is this gene located?
- How many transcripts (splice variants) are there?
- How many of these transcripts are protein coding?
- What is the longest transcript, and how long is the protein it encodes?
- Which transcript has a CCDS record associated with it?

? Why is the CCDS important – what does it tell us?

(b) Click on Phenotype at the left side of the page. Are there any diseases associated with this gene, according to MIM (Mendelian Inheritance in Man)?

(c) In the transcript table, click on the transcript ID for MYH9-001, and go to the Transcript tab.

- How many exons does it have?

- Are any of the exons completely or partially untranslated?
- Is there an associated sequence in UniProtKB/Swiss-Prot? Have a look at the General identifiers for this transcript.
- What are some functions of MYH9-001 according to the Gene Ontology consortium? Have a look at the Ontology table for this transcript.

(d) Are there microarray (oligo) probes that can be used to monitor ENST00000216181 expression?

Exercise 8 – Finding a gene associated with a phenotype

Phenylketonuria is a genetic disorder caused by an inability to metabolise phenylalanine. This results in an accumulation of phenylalanine causing seizures and mental retardation.

(a) Search for phenylketonuria from the Ensembl homepage. What gene is associated with this disorder?

(b) What tissues is this gene expressed in? Is this surprising, given the gene's role in disease? What is meant by "Intron-spanning reads" and "RNASeq alignments"?

(c) How many protein coding transcripts does this gene have? View all of these in the transcript comparison view.

(d) What is the MIM disease identifier for this gene? Does Orphanet list any other disorders associated with this gene?

Exercise 9 – Exploring a plant gene (*Vitis vinifera*, grape)

Start in http://plants.ensembl.org/index.html and select the *Vitis vinifera* genome.

(a) What GO: biological process terms are associated with the *MADS4* gene?

(b) Go to the transcript tab for the only transcript, Vv01s0010g03900.t01. How many exons does it have? Which one is the longest? How much of that is coding?

(c) What domains can be found in the protein product of this transcript? How many different groups agree with each of these domains?

BioMart

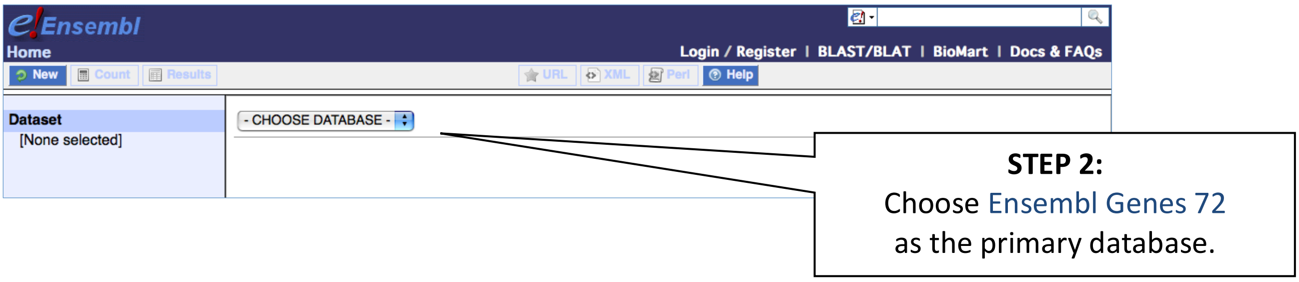
Demo: BioMart

Follow these instructions to guide you through BioMart to answer the following query:

> You have three questions about a set of human genes: *ESPN, MYH9, USH1C, CHD7, CISD2, THRB, DFNB31*
> *(*these are HGNC gene symbols. More details on the HUGO Gene Nomenclature Committee can be found on http://www.genenames.org)
>
> 1) What are the EntrezGene IDs for these genes?
>
> 2) Are there associated functions from the GO (gene ontology) project that might help describe their function?
>
> 3) What are their cDNA sequences?

Step 1: Click on *BioMart* in the top header of a www.ensembl.org page to go to: www.ensembl.org/biomart/martview

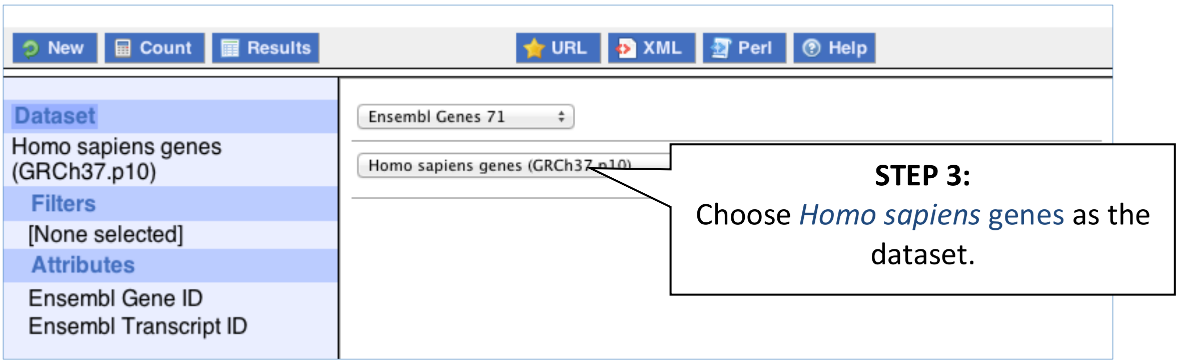
NOTE: These answers were determined using BioMart Ensembl 71.



Dataset
[None selected]

- CHOOSE DATABASE -

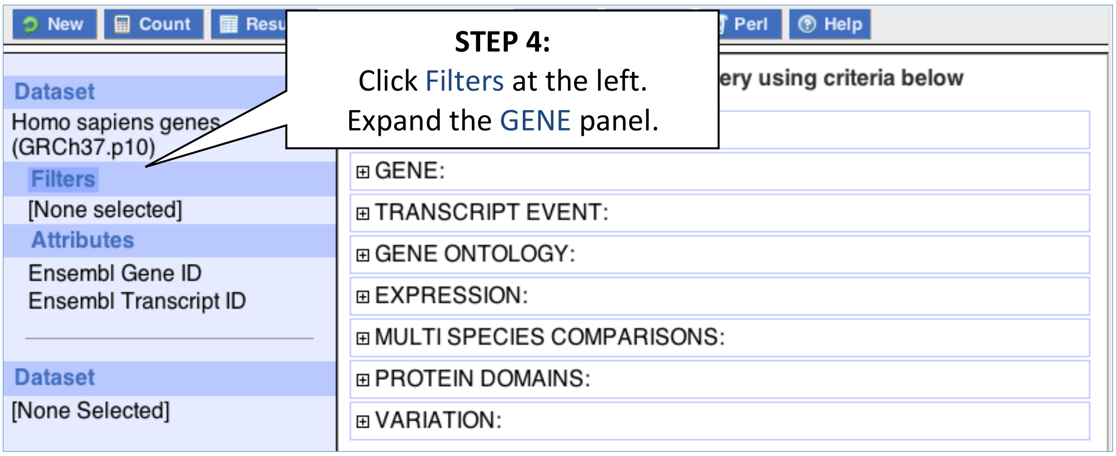
STEP 2:
Choose Ensembl Genes 72
as the primary database.



Dataset
Homo sapiens genes (GRCh37.p10)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Ensembl Genes 71
Homo sapiens genes (GRCh37.p10)

STEP 3:
Choose *Homo sapiens* genes as the dataset.



...ery using criteria below

Dataset

Homo sapiens genes (GRCh37.p10)

Filters

[None selected]

Attributes

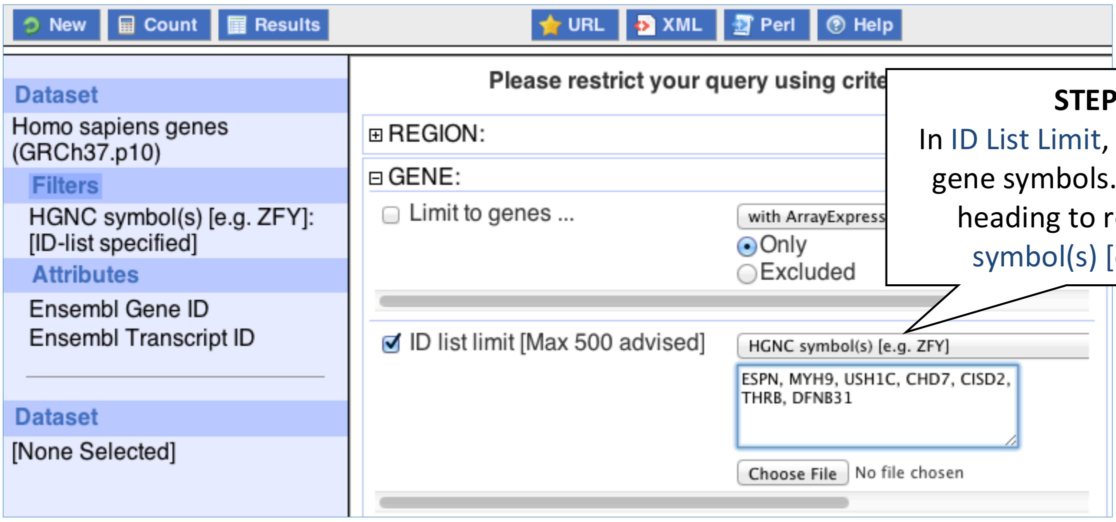
Ensembl Gene ID
Ensembl Transcript ID

Dataset

[None Selected]

⊞ GENE:
⊞ TRANSCRIPT EVENT:
⊞ GENE ONTOLOGY:
⊞ EXPRESSION:
⊞ MULTI SPECIES COMPARISONS:
⊞ PROTEIN DOMAINS:
⊞ VARIATION:

STEP 4:
Click Filters at the left.
Expand the GENE panel.



Please restrict your query using crite...

Dataset

Homo sapiens genes (GRCh37.p10)

Filters

HGNC symbol(s) [e.g. ZFY]: [ID-list specified]

Attributes

Ensembl Gene ID
Ensembl Transcript ID

Dataset

[None Selected]

⊞ REGION:

⊟ GENE:

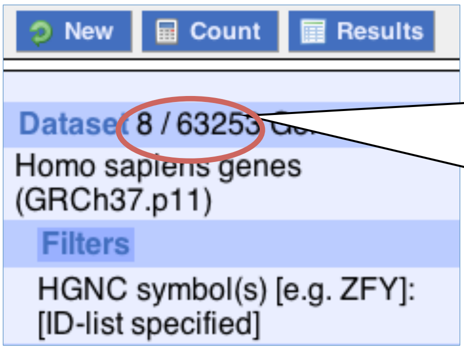
□ Limit to genes ... with ArrayExpress ● Only ○ Excluded

☑ ID list limit [Max 500 advised] HGNC symbol(s) [e.g. ZFY]

ESPN, MYH9, USH1C, CHD7, CISD2, THRB, DFNB31

Choose File No file chosen

STEP 5:
In ID List Limit, paste in your gene symbols. Change the heading to read HGNC symbol(s) [e.g. ZFY].



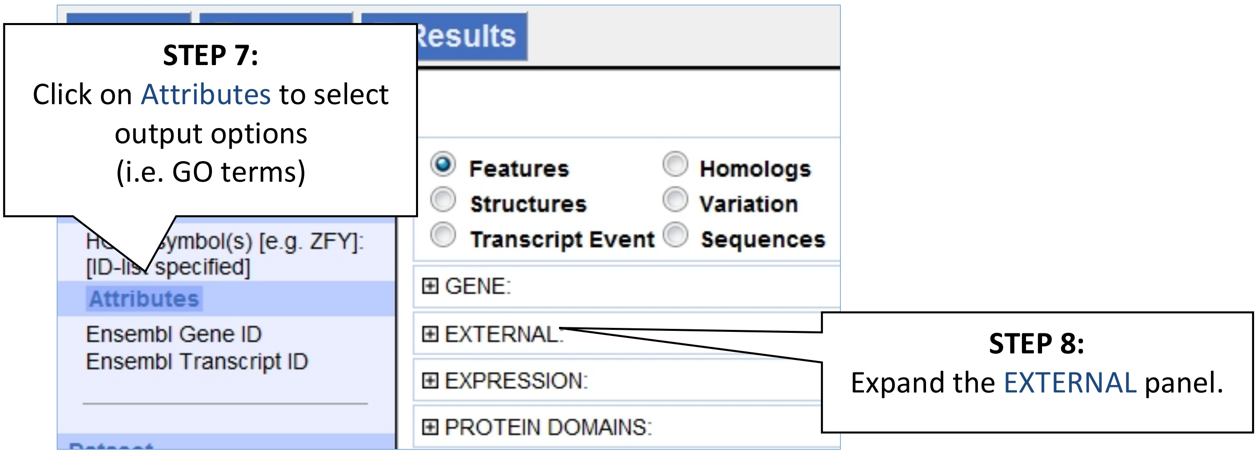
Dataset 8 / 63253 Ge...

Homo sapiens genes (GRCh37.p11)

Filters

HGNC symbol(s) [e.g. ZFY]: [ID-list specified]

STEP 6:
Click Count to see BioMart is reading 8 genes out of 63,253 possible *H. sapiens* genes (this number includes ncRNA genes) (note that *CHD7* has an Ensembl copy and an LRG copy (http://www.ensembl.org/Help/Glossary?id=406), hence 8 counts for 7 HGNC genes).

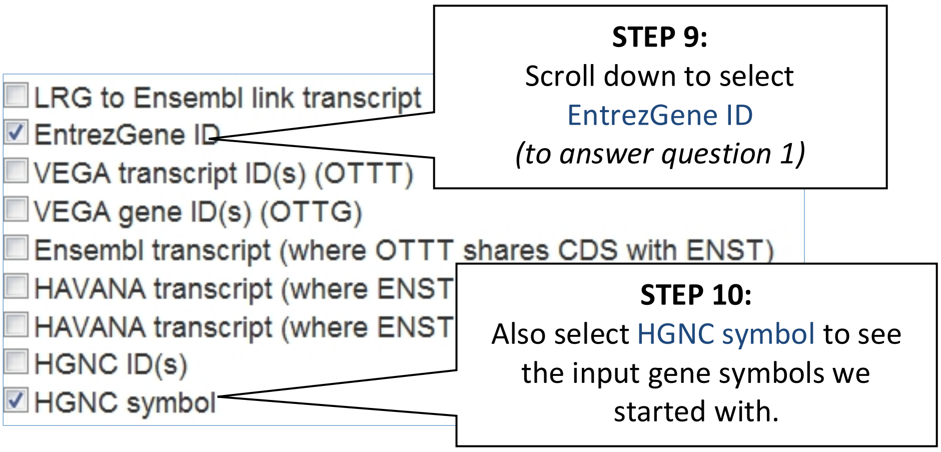


● Features ○ Homologs
○ Structures ○ Variation
○ Transcript Event ○ Sequences

⊞ GENE:
⊞ EXTERNAL:
⊞ EXPRESSION:
⊞ PROTEIN DOMAINS:

STEP 7:
Click on Attributes to select output options (i.e. GO terms)

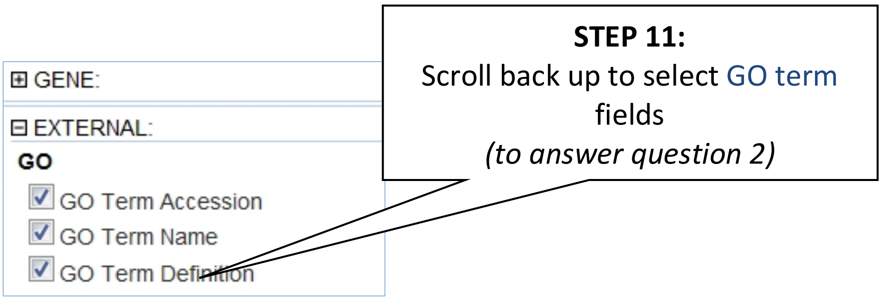
STEP 8:
Expand the EXTERNAL panel.



□ LRG to Ensembl link transcript
☑ EntrezGene ID
□ VEGA transcript ID(s) (OTTT)
□ VEGA gene ID(s) (OTTG)
□ Ensembl transcript (where OTTT shares CDS with ENST)
□ HAVANA transcript (where ENST
□ HAVANA transcript (where ENST
□ HGNC ID(s)
☑ HGNC symbol

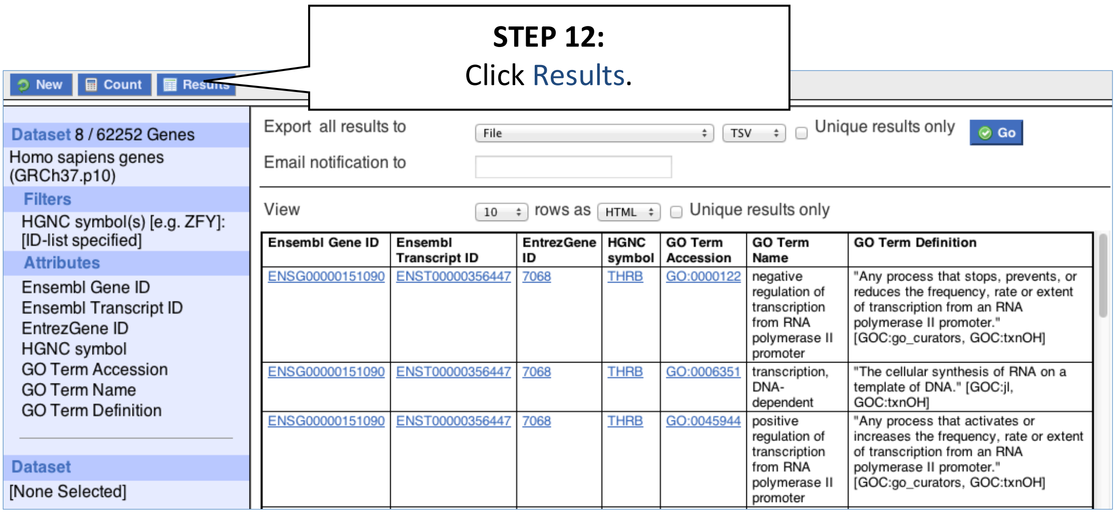
STEP 9:
Scroll down to select EntrezGene ID
(to answer question 1)

STEP 10:
Also select HGNC symbol to see the input gene symbols we started with.



⊞ GENE:
⊟ EXTERNAL:
GO
☑ GO Term Accession
☑ GO Term Name
☑ GO Term Definition

STEP 11:
Scroll back up to select GO term fields
(to answer question 2)



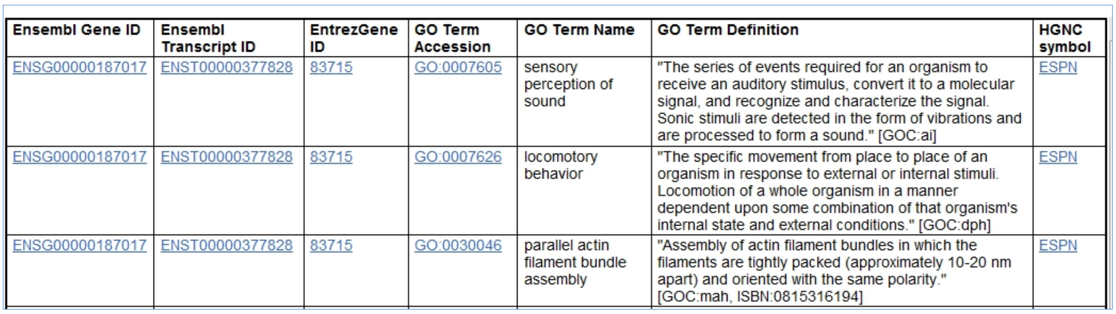
STEP 12:
Click Results.

Dataset 8 / 62252 Genes
Homo sapiens genes (GRCh37.p10)
Filters
HGNC symbol(s) [e.g. ZFY]: [ID-list specified]
Attributes
Ensembl Gene ID
Ensembl Transcript ID
EntrezGene ID
HGNC symbol
GO Term Accession
GO Term Name
GO Term Definition
Dataset
[None Selected]

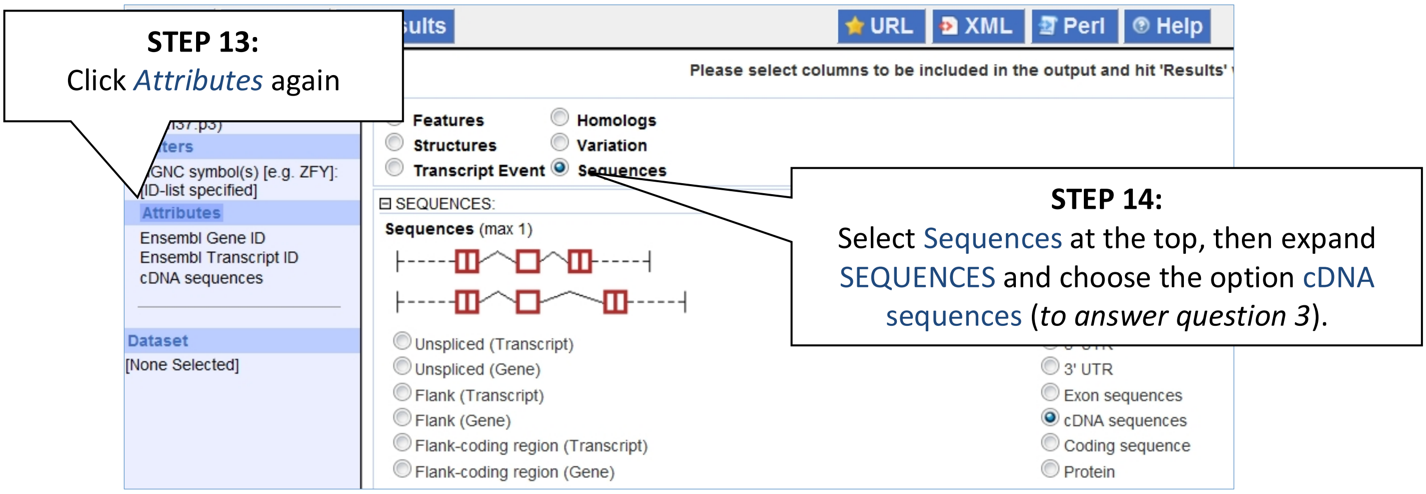
Export all results to File TSV □ Unique results only
Email notification to
View 10 rows as HTML □ Unique results only

Ensembl Gene ID	Ensembl Transcript ID	EntrezGene ID	HGNC symbol	GO Term Accession	GO Term Name	GO Term Definition
ENSG00000151090	ENST00000356447	7068	THRB	GO:0000122	negative regulation of transcription from RNA polymerase II promoter	"Any process that stops, prevents, or reduces the frequency, rate or extent of transcription from an RNA polymerase II promoter." [GOC:go_curators, GOC:txnOH]
ENSG00000151090	ENST00000356447	7068	THRB	GO:0006351	transcription, DNA-dependent	"The cellular synthesis of RNA on a template of DNA." [GOC:jl, GOC:txnOH]
ENSG00000151090	ENST00000356447	7068	THRB	GO:0045944	positive regulation of transcription from RNA polymerase II promoter	"Any process that activates or increases the frequency, rate or extent of transcription from an RNA polymerase II promoter." [GOC:go_curators, GOC:txnOH]

Why are there multiple rows for one gene ID? For example, look at the first few rows.



Ensembl Gene ID	Ensembl Transcript ID	EntrezGene ID	GO Term Accession	GO Term Name	GO Term Definition	HGNC symbol
ENSG00000187017	ENST00000377828	83715	GO:0007605	sensory perception of sound	"The series of events required for an organism to receive an auditory stimulus, convert it to a molecular signal, and recognize and characterize the signal. Sonic stimuli are detected in the form of vibrations and are processed to form a sound." [GOC:ai]	ESPN
ENSG00000187017	ENST00000377828	83715	GO:0007626	locomotory behavior	"The specific movement from place to place of an organism in response to external or internal stimuli. Locomotion of a whole organism in a manner dependent upon some combination of that organism's internal state and external conditions." [GOC:dph]	ESPN
ENSG00000187017	ENST00000377828	83715	GO:0030046	parallel actin filament bundle assembly	"Assembly of actin filament bundles in which the filaments are tightly packed (approximately 10-20 nm apart) and oriented with the same polarity." [GOC:mah, ISBN:0815316194]	ESPN



Please select columns to be included in the output and hit 'Results'

h37.p3)

ters

HGNC symbol(s) [e.g. ZFY]:
ID-list specified]

Attributes

Ensembl Gene ID
Ensembl Transcript ID
cDNA sequences

Dataset

[None Selected]

○ Features ○ Homologs
○ Structures ○ Variation
○ Transcript Event ● Sequences

⊟ SEQUENCES:

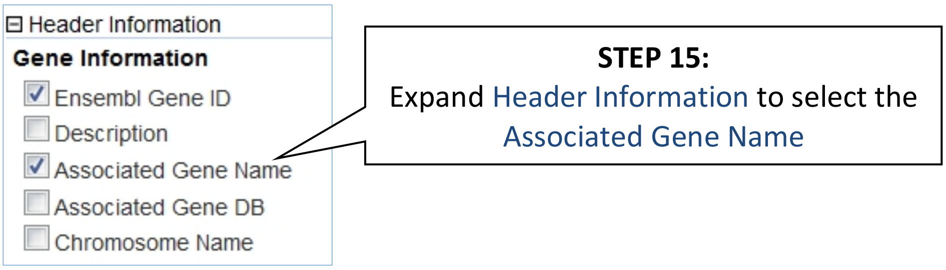
Sequences (max 1)

○ Unspliced (Transcript)
○ Unspliced (Gene)
○ Flank (Transcript)
○ Flank (Gene)
○ Flank-coding region (Transcript)
○ Flank-coding region (Gene)

○ 3' UTR
○ Exon sequences
● cDNA sequences
○ Coding sequence
○ Protein

STEP 13:
Click *Attributes* again

STEP 14:
Select Sequences at the top, then expand SEQUENCES and choose the option cDNA sequences (*to answer question 3*).

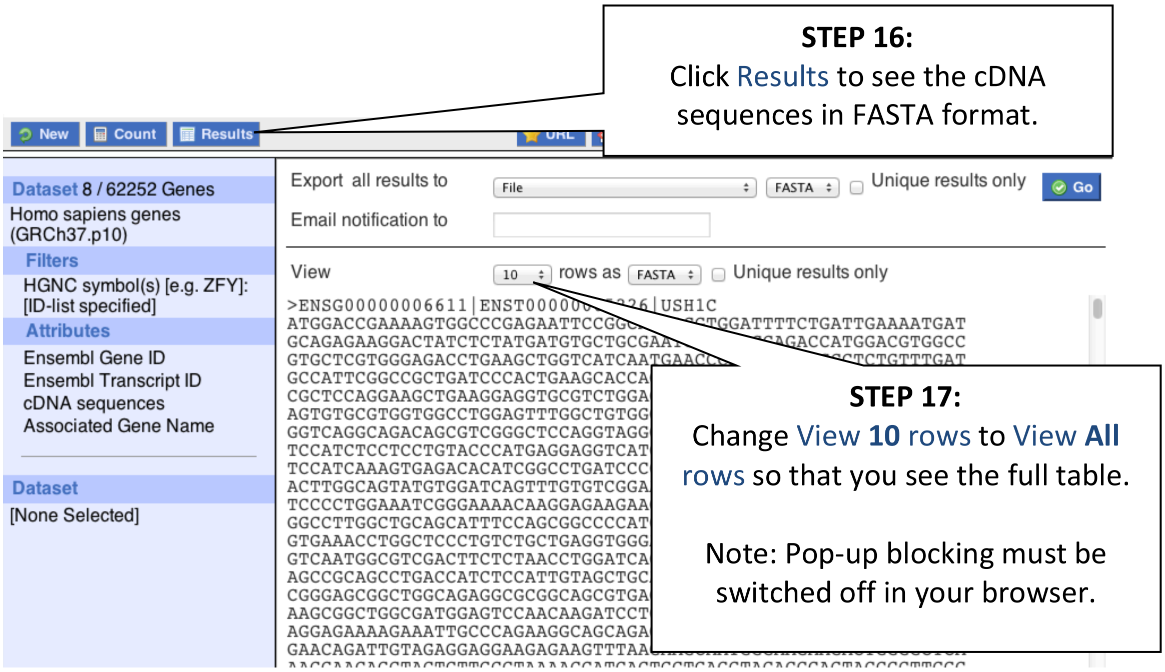


⊟ Header Information

Gene Information

☑ Ensembl Gene ID
□ Description
☑ Associated Gene Name
□ Associated Gene DB
□ Chromosome Name

STEP 15:
Expand Header Information to select the Associated Gene Name



STEP 16:
Click Results to see the cDNA sequences in FASTA format.

Dataset 8 / 62252 Genes
Homo sapiens genes (GRCh37.p10)

Filters
HGNC symbol(s) [e.g. ZFY]: [ID-list specified]

Attributes
Ensembl Gene ID
Ensembl Transcript ID
cDNA sequences
Associated Gene Name

Dataset
[None Selected]

Export all results to File FASTA □ Unique results only Go

Email notification to

View 10 rows as FASTA □ Unique results only

```
>ENSG00000006611|ENST00000     226|USH1C
ATGGACCGAAAAGTGGCCCGAGAATTCCGGC       CCTGGATTTTCTGATTGAAAATGAT
GCAGAGAAGGACTATCTCTATGATGTGCTGCGAAT      CAGACCATGGACGTGGCC
GTGCTCGTGGGAGACCTGAAGCTGGTCATCAATGAACC      CCTCTGTTTGAT
GCCATTCGGCCGCTGATCCCACTGAAGCACCA
CGCTCCAGGAAGCTGAAGGAGGTGCGTCTGGA
AGTGTGCGTGGTGGCCTGGAGTTTGGCTGTGG
GGTCAGGCAGACAGCGTCGGGCTCCAGGTAGG
TCCATCTCCTCCTGTACCCATGAGGAGGTCAT
TCCATCAAAGTGAGACACATCGGCCTGATCCC
ACTTGGCAGTATGTGGATCAGTTTGTGTCGGA
TCCCCTGGAAATCGGGAAAACAAGGAGAAGAA
GGCCTTGGCTGCAGCATTTCCAGCGGCCCCAT
GTGAAACCTGGCTCCCTGTCTGCTGAGGTGGG
GTCAATGGCGTCGACTTCTCTAACCTGGATCA
AGCCGCAGCCTGACCATCTCCATTGTAGCTGC
CGGGAGCGGCTGGCAGAGGCGCGGCAGCGTGA
AAGCGGCTGGCGATGGAGTCCAACAAGATCCT
AGGAGAAAAGAAATTGCCCAGAAGGCAGCAGA
GAACAGATTGTAGAGGAGGAAGAGAAGTTTAA
```

STEP 17:
Change View **10** rows to View **All** rows so that you see the full table.

Note: Pop-up blocking must be switched off in your browser.

Note: you can use the Go button to export a file.

What did you learn about the human genes in this exercise? Could you learn these things from the Ensembl browser? Would it take longer?

For more details on BioMart, have a look at these publications:

Smedley, D. *et al*
BioMart – biological queries made easy
BMC Genomics 2009 Jan 14;10:22

Kinsella, R.J. *et al*
Ensembl BioMarts: a hub for data retrieval across taxonomic space.
Database (Oxford) 2011:bar030

Exercises: BioMart

Exercise 10 – Finding genes by protein domain

Find mouse proteins with transmembrane domains located on chromosome 9.

Exercise 11 – Convert IDs

BioMart is a very handy tool when you want to convert IDs from different databases. The following is a list of 29 IDs of **human proteins** from the NCBI **RefSeq** database (http://www.ncbi.nlm.nih.gov/projects/RefSeq/):

NP_001218, NP_203125, NP_203124, NP_203126, NP_001007233, NP_150636, NP_150635, NP_001214, NP_150637, NP_150634, NP_150649, NP_001216, NP_116787, NP_001217, NP_127463, NP_001220, NP_004338, NP_004337, NP_116786, NP_036246, NP_116756, NP_116759, NP_001221, NP_203519, NP_001073594, NP_001219, NP_001073593, NP_203520, NP_203522

Generate a list that shows to which Ensembl Gene IDs and to which HGNC symbols these RefSeq IDs correspond. Do these 29 proteins correspond to 29 genes?

Hint: For this exercise, it's easier to copy and paste the IDs from the online exercise booklet. One copy is here:
URL

Exercise 12 – Export homologues

For a list of *Ciona savignyi* Ensembl genes, export the human orthologues.

ENSCSAVG00000000002, ENSCSAVG00000000003, ENSCSAVG00000000006, ENSCSAVG00000000007, ENSCSAVG00000000009, ENSCSAVG00000000011

Exercise 13 – Export structural variants

You can use BioMart to query variants, not just genes.

(a) Export the study accession, source name, chromosome, sequence region start and end (in bp) of human structural variations (SV) on chromosome 1, starting at 130,408 and ending at 210,597.

(b) In a new BioMart query, find the alleles, phenotype descriptions, and associated genes for rs1801500 and rs1801368. Can you view this same information in the Ensembl browser?

Exercise 14 – Find genes associated with array probes

Forrest *et al* performed a microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers (Environ Health Perspect. 2005 June; 113(6): 801–807). The microarray used was the human Affymetrix U133A/B (also called U133 plus 2) GeneChip. The top 25 up-regulated probe-sets were:

207630_s_at, 221840_at, 219228_at, 204924_at, 227613_at, 223454_at, 228962_at, 214696_at, 210732_s_at, 212370_at, 225390_s_at, 227645_at, 226652_at, 221641_s_at, 202055_at, 226743_at, 228393_s_at, 225120_at, 218515_at, 202224_at, 200614_at, 212014_x_at, 223461_at, 209835_x_at, 213315_x_at

(a) Retrieve for the genes corresponding to these probe-sets the Ensembl Gene and Transcript IDs as well as their HGNC symbols and descriptions.

(b) In order to analyse these genes for possible promoter/enhancer elements, retrieve the 2000 bp upstream of the transcripts of these genes.

(c) In order to be able to study these human genes in mouse, identify their mouse orthologues. Also retrieve the genomic coordinates of these orthologues.

Exercise 15 – BioMart in Ensembl Fungi

You can use BioMart for the non-vertebrate species hosted at www.ensemblgenomes.org. Export a list of Gene IDs (from the PomBase project only) for *S. pombe* that are protein coding and located on Chromosome III.

(Start at http://fungi.ensembl.org)

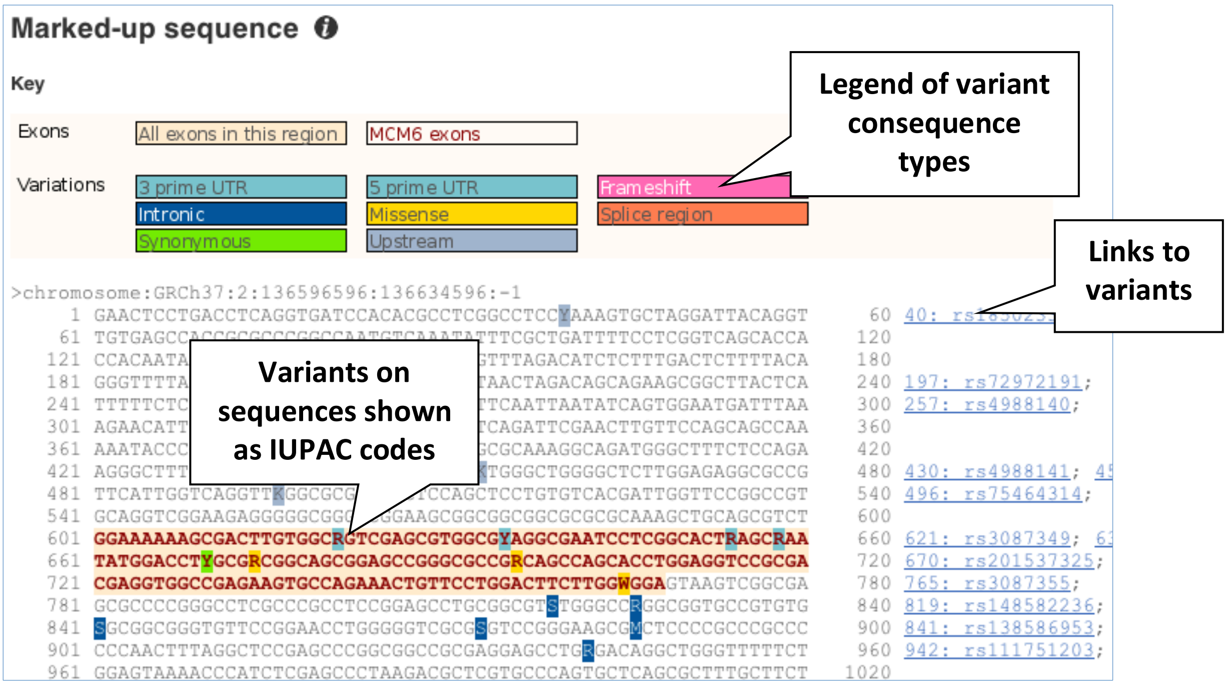
Variation

Demo: Finding variants in Ensembl

In any of the sequence views shown in the Gene and Transcript tabs, you can view variants on the sequence. You can do this by clicking on Configure this page from any of these views.

Let's take a look at the Gene sequence view for *MCM6* in human. Search for *MCM6* and go to the Sequence view.

If you can't see variants marked on this view, click on Configure this page and select Show variations: Yes and show links.



Marked-up sequence

Key

Exons	All exons in this region	MCM6 exons	
Variations	3 prime UTR	5 prime UTR	Frameshift
	Intronic	Missense	Splice region
	Synonymous	Upstream	

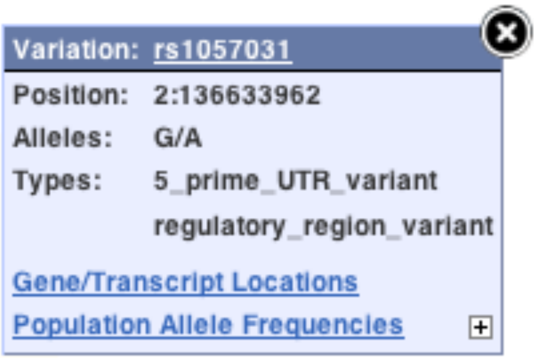
```
>chromosome:GRCh37:2:136596596:136634596:-1
  1 GAACTCCTGACCTCAGGTGATCCACACGCCTCGGCCTCCYAAAGTGCTAGGATTACAGGT     60 40: rs
 61 TGTGAGCCACCGCGCCCGGCCAATGTCAAATATTTCGCTGATTTTCCTCGGTCAGCACCA    120
121 CCACAATA                         GTTTAGACATCTCTTTGACTCTTTTACA    180
181 GGGTTTTA                         TAACTAGACAGCAGAAGCGGCTTACTCA    240 197: rs72972191;
241 TTTTTCTC                         TTCAATTAATATCAGTGGAATGATTTAA    300 257: rs4988140;
301 AGAACATT                         TCAGATTCGAACTTGTTCCAGCAGCCAA    360
361 AAATACCC                         GCGCAAAGGCAGATGGGCTTTCTCCAGA    420
421 AGGGCTTT                         KTGGGCTGGGGCTCTTGGAGAGGCGCCG    480 430: rs4988141; 4
481 TTCATTGGTCAGGTTKGGCGCG       TCCAGCTCCTGTGTCACGATTGGTTCCGGCCGT    540 496: rs75464314;
541 GCAGGTCGGAAGAGGGGGCGGG   GGAAGCGGCGGCGGCGCGCGCAAAGCTGCAGCGTCT    600
601 GGAAAAAAGCGACTTGTGGCRGTCGAGCGTGGCGYAGGCGAATCCTCGGCACTRAGCRAA    660 621: rs3087349; 6
661 TATGGACCTYGCGRCGGCAGCGGAGCCGGGCGCCGRCAGCCAGCACCTGGAGGTCCGCGA    720 670: rs201537325;
721 CGAGGTGGCCGAGAAGTGCCAGAAACTGTTCCTGGACTTCTTGGWGGAGTAAGTCGGCGA    780 765: rs3087355;
781 GCGCCCCGGGCCTCGCCCGCCTCCGGAGCCTGCGGCGTSTGGGCCRGGCGGTGCCGTGTG    840 819: rs148582236;
841 SGCGGCGGGTGTTCCGGAACCTGGGGGTCGCGSGTCCGGGAAGCGMCTCCCCGCCCGCCC    900 841: rs138586953;
901 CCCAACTTTAGGCTCCGAGCCCGGCGGCCGCGAGGAGCCTGRGACAGGCTGGGTTTTTCT    960 942: rs111751203;
961 GGAGTAAAACCCATCTCGAGCCCTAAGACGCTCGTGCCCAGTGCTCAGCGCTTTGCTTCT   1020
```

Legend of variant consequence types

Links to variants

Variants on sequences shown as IUPAC codes

Find out more about a variant by clicking on it.



Variation:	rs1057031
Position:	2:136633962
Alleles:	G/A
Types:	5_prime_UTR_variant regulatory_region_variant

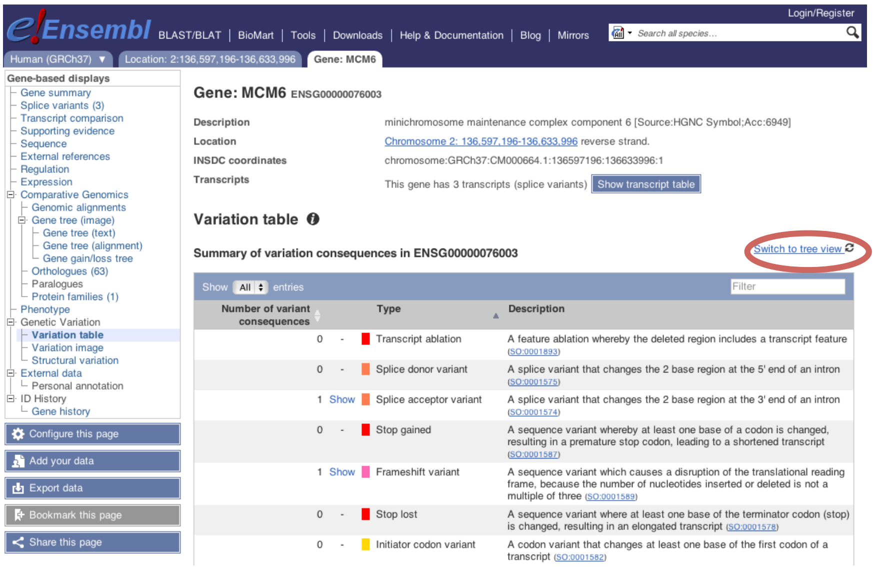
Gene/Transcript Locations

Population Allele Frequencies

You can add variants to all other sequence views in the same way.

You can go to the Variation tab by clicking on the variant ID. For now, we'll explore more ways of finding variants.

To view all the sequence variations in table form, click the Variation table link at the left of the gene tab.



Gene: MCM6 ENSG00000076003

Description	minichromosome maintenance complex component 6 [Source:HGNC Symbol;Acc:6949]
Location	Chromosome 2: 136,597,196-136,633,996 reverse strand.
INSDC coordinates	chromosome:GRCh37:CM000664.1:136597196:136633996:1
Transcripts	This gene has 3 transcripts (splice variants) Show transcript table

Variation table

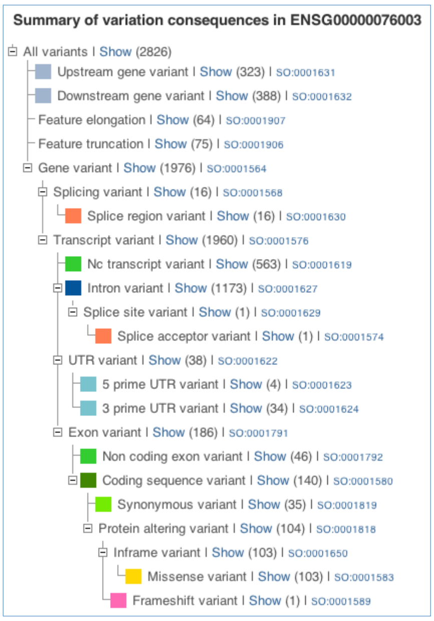
Summary of variation consequences in ENSG00000076003

Switch to tree view

Show All entries Filter

Number of variant consequences		Type	Description
0	-	🟥 Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature (SO:0001893)
0	-	🟧 Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron (SO:0001575)
1	Show	🟧 Splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an intron (SO:0001574)
0	-	🟥 Stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript (SO:0001587)
1	Show	🟪 Frameshift variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three (SO:0001589)
0	-	🟥 Stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript (SO:0001578)
0	-	🟨 Initiator codon variant	A codon variant that changes at least one base of the first codon of a transcript (SO:0001582)

The table is divided into consequence types. You can also view the consequence types as an ontology tree. Click on Switch to tree view at the right hand side to change the view.

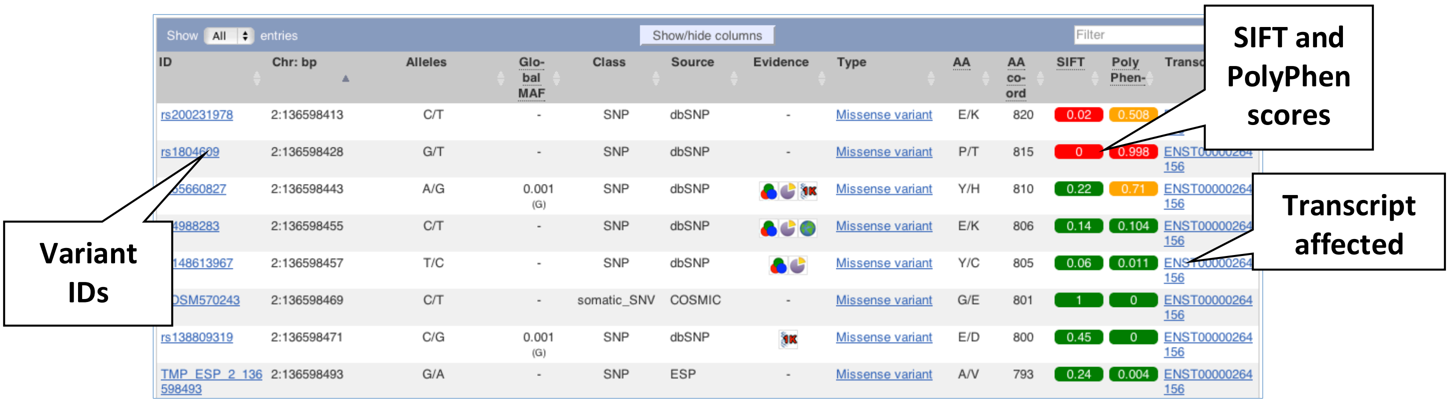


Summary of variation consequences in ENSG00000076003

- All variants | Show (2826)
  - 🟦 Upstream gene variant | Show (323) | SO:0001631
  - 🟦 Downstream gene variant | Show (388) | SO:0001632
  - Feature elongation | Show (64) | SO:0001907
  - Feature truncation | Show (75) | SO:0001906
  - Gene variant | Show (1976) | SO:0001564
    - Splicing variant | Show (16) | SO:0001568
      - 🟧 Splice region variant | Show (16) | SO:0001630
    - Transcript variant | Show (1960) | SO:0001576
      - 🟩 Nc transcript variant | Show (563) | SO:0001619
      - 🟦 Intron variant | Show (1173) | SO:0001627
        - Splice site variant | Show (1) | SO:0001629
          - 🟧 Splice acceptor variant | Show (1) | SO:0001574
      - UTR variant | Show (38) | SO:0001622
        - 🟦 5 prime UTR variant | Show (4) | SO:0001623
        - 🟦 3 prime UTR variant | Show (34) | SO:0001624
      - Exon variant | Show (186) | SO:0001791
        - 🟩 Non coding exon variant | Show (46) | SO:0001792
        - 🟩 Coding sequence variant | Show (140) | SO:0001580
          - 🟩 Synonymous variant | Show (35) | SO:0001819
          - Protein altering variant | Show (104) | SO:0001818
            - Inframe variant | Show (103) | SO:0001650
              - 🟨 Missense variant | Show (103) | SO:0001583
            - 🟪 Frameshift variant | Show (1) | SO:0001589

Click on Show to expand a detailed table for any of the consequence types available.

Let’s expand Missense variants.



Show All entries | Show/hide columns | Filter

ID	Chr: bp	Alleles	Global MAF	Class	Source	Evidence	Type	AA	AA co-ord	SIFT	PolyPhen-	Transc
rs200231978	2:136598413	C/T	-	SNP	dbSNP	-	Missense variant	E/K	820	0.02	0.508	
rs1804609	2:136598428	G/T	-	SNP	dbSNP	-	Missense variant	P/T	815	0	0.998	ENST00000264 156
55660827	2:136598443	A/G	0.001 (G)	SNP	dbSNP	● ● 1K	Missense variant	Y/H	810	0.22	0.71	ENST00000264 156
4988283	2:136598455	C/T	-	SNP	dbSNP	● ● ●	Missense variant	E/K	806	0.14	0.104	ENST00000264 156
148613967	2:136598457	T/C	-	SNP	dbSNP	● ●	Missense variant	Y/C	805	0.06	0.011	ENST00000264 156
OSM570243	2:136598469	C/T	-	somatic_SNV	COSMIC	-	Missense variant	G/E	801	1	0	ENST00000264 156
rs138809319	2:136598471	C/G	0.001 (G)	SNP	dbSNP	1K	Missense variant	E/D	800	0.45	0	ENST00000264 156
TMP_ESP_2_136 598493	2:136598493	G/A	-	SNP	ESP	-	Missense variant	A/V	793	0.24	0.004	ENST00000264 156

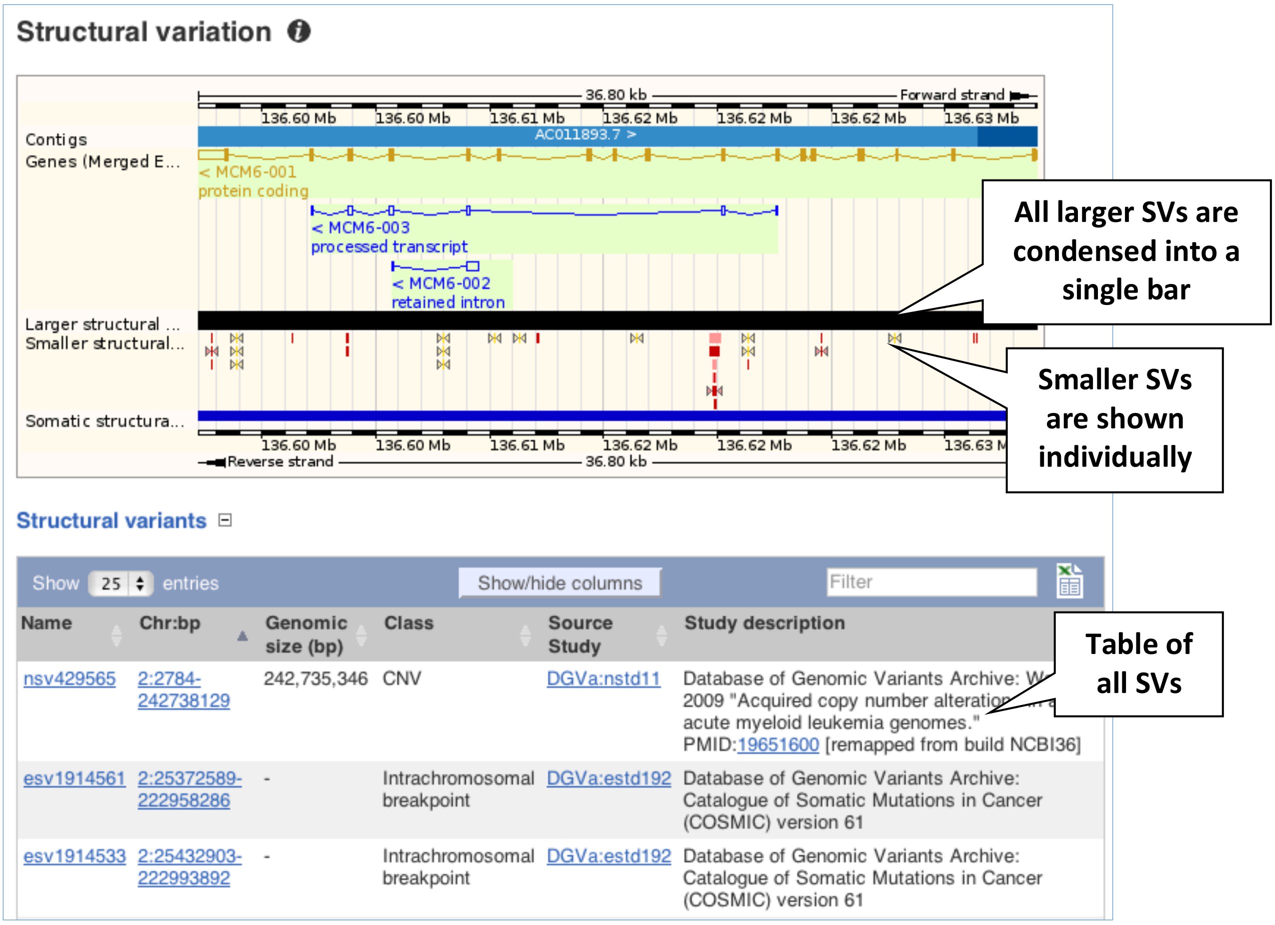
SIFT and PolyPhen scores

Transcript affected

Variant IDs

The table contains lots of information about the variants. You can click on the IDs here to go to the Variation tab too.

Let’s look at Structural Variation in the Gene Tab. You’ll find it in the left-hand menu.



Structural variation

All larger SVs are condensed into a single bar

Smaller SVs are shown individually

Structural variants

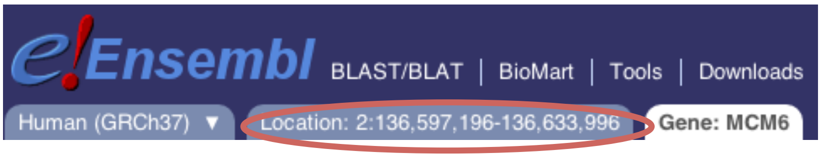
Show 25 entries | Show/hide columns | Filter

Name	Chr:bp	Genomic size (bp)	Class	Source Study	Study description
nsv429565	2:2784-242738129	242,735,346	CNV	DGVa:nstd11	Database of Genomic Variants Archive: W 2009 "Acquired copy number alteration acute myeloid leukemia genomes." PMID:19651600 [remapped from build NCBI36]
esv1914561	2:25372589-222958286	-	Intrachromosomal breakpoint	DGVa:estd192	Database of Genomic Variants Archive: Catalogue of Somatic Mutations in Cancer (COSMIC) version 61
esv1914533	2:25432903-222993892	-	Intrachromosomal breakpoint	DGVa:estd192	Database of Genomic Variants Archive: Catalogue of Somatic Mutations in Cancer (COSMIC) version 61

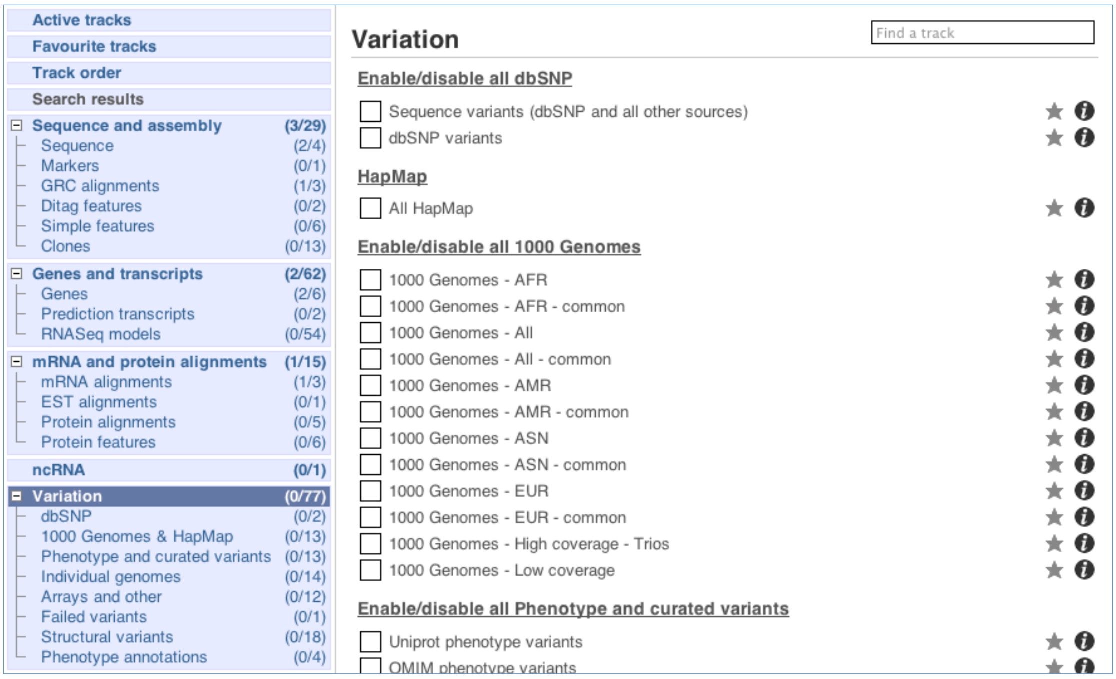
Table of all SVs

You can click on the structural variants (SVs) in the image, or on their IDs in the table to go to the SV tab.

Let's have a look at variants in the Location tab. Click on the Location tab in the top bar.


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads
Human (GRCh37) ▼
Location: 2:136,597,196-136,633,996
Gene: MCM6

Configure this page and open Variation from the left-hand menu.



Variation

Enable/disable all dbSNP

□ Sequence variants (dbSNP and all other sources)
□ dbSNP variants

HapMap

□ All HapMap

Enable/disable all 1000 Genomes

□ 1000 Genomes - AFR
□ 1000 Genomes - AFR - common
□ 1000 Genomes - All
□ 1000 Genomes - All - common
□ 1000 Genomes - AMR
□ 1000 Genomes - AMR - common
□ 1000 Genomes - ASN
□ 1000 Genomes - ASN - common
□ 1000 Genomes - EUR
□ 1000 Genomes - EUR - common
□ 1000 Genomes - High coverage - Trios
□ 1000 Genomes - Low coverage

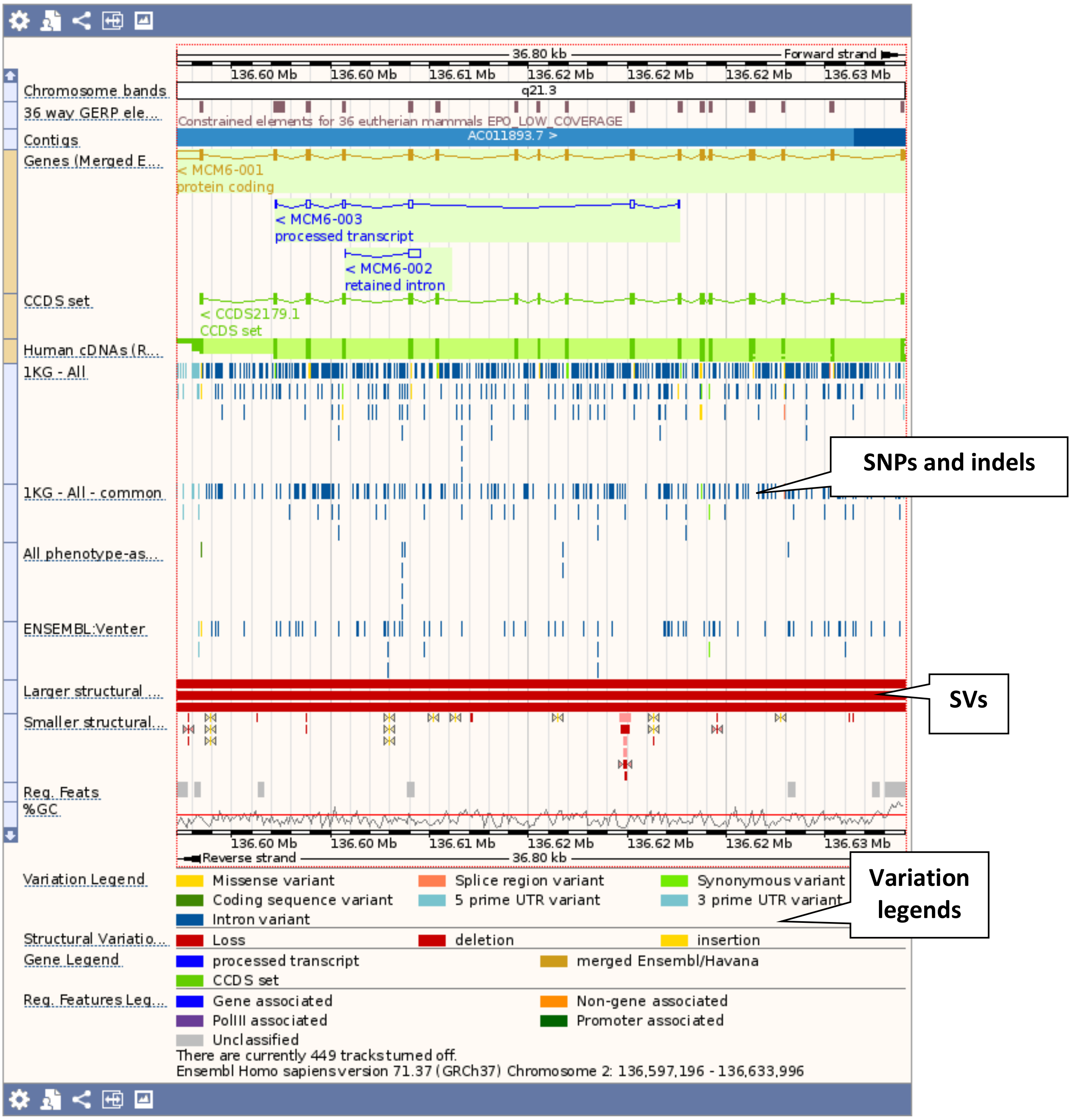
Enable/disable all Phenotype and curated variants

□ Uniprot phenotype variants
□ OMIM phenotype variants

There are various options for turning on variants. You can turn on variants by source, by frequency, presence of a phenotype or by individual genome they were isolated from. Turn on the following sequence variants in Normal.

- 1000 genomes – All
- 1000 genomes – All – common
- All phenotype-associated variants
- ENSEMBL:Venter

Also turn on Larger and Smaller Structural variants (all sources) in Expanded.

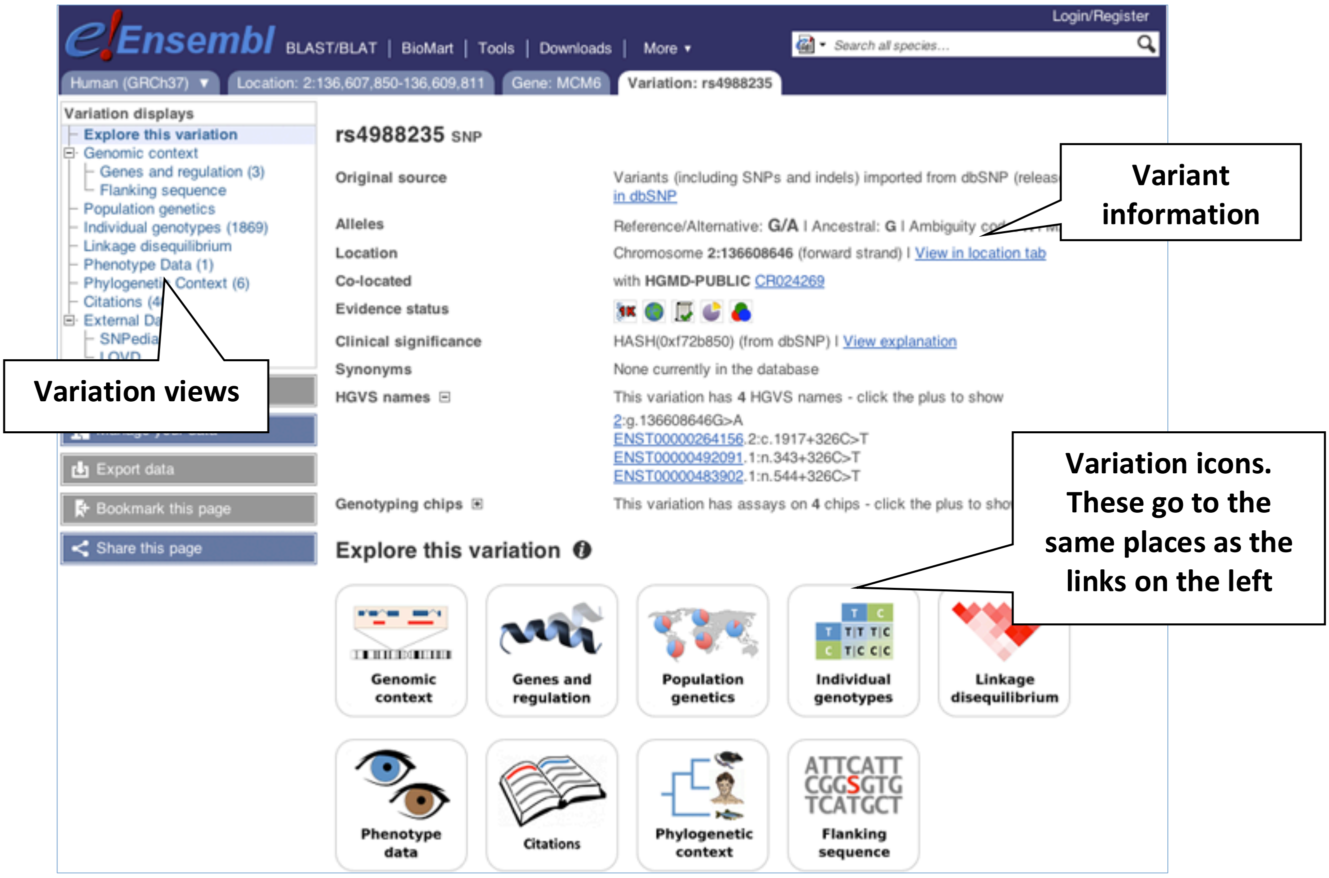


There are currently 449 tracks turned off.
Ensembl Homo sapiens version 71.37 (GRCh37) Chromosome 2: 136,597,196 - 136,633,996

Click on a variant to find out more information. It may be easier to see the individual variants if you zoom in.

Let's zoom in on the region 2:136607850-136609811 by typing it into the Location box.

Change the track style on one of the SNP variant tracks to Expanded with name. Click on the variant rs4988235 to open a pop-up, then click on the ID to open the Variation tab.



rs4988235 SNP

Original source: Variants (including SNPs and indels) imported from dbSNP (releas
in dbSNP

Alleles: Reference/Alternative: **G/A** | Ancestral: **G** | Ambiguity co

Location: Chromosome **2:136608646** (forward strand) | View in location tab

Co-located: with **HGMD-PUBLIC** CR024269

Evidence status

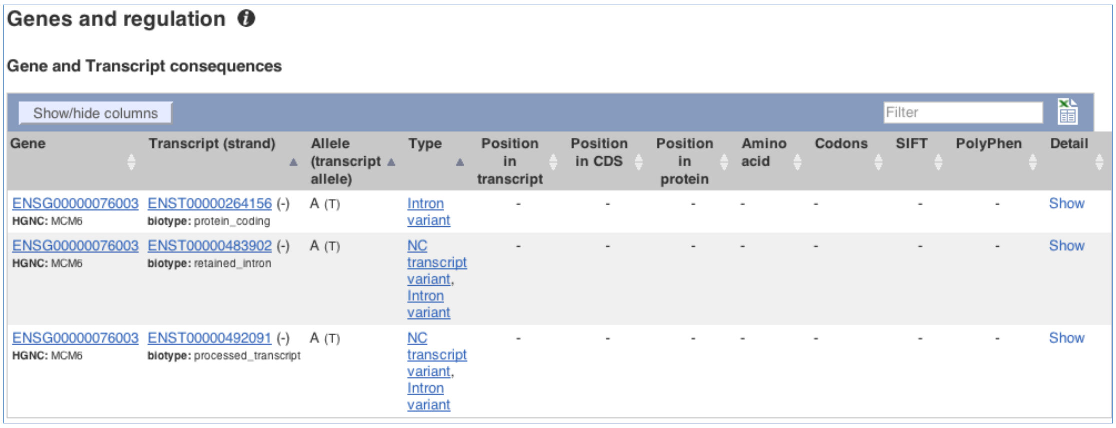
Clinical significance: HASH(0xf72b850) (from dbSNP) | View explanation

Synonyms: None currently in the database

HGVS names ⊟: This variation has **4** HGVS names - click the plus to show
2:g.136608646G>A
ENST00000264156.2:c.1917+326C>T
ENST00000492091.1:n.343+326C>T
ENST00000483902.1:n.544+326C>T

Genotyping chips ⊞: This variation has assays on **4** chips - click the plus to sho

The icons show you what information is available for this variant. Click on Genes and regulation, or follow the link at the left.



Genes and regulation

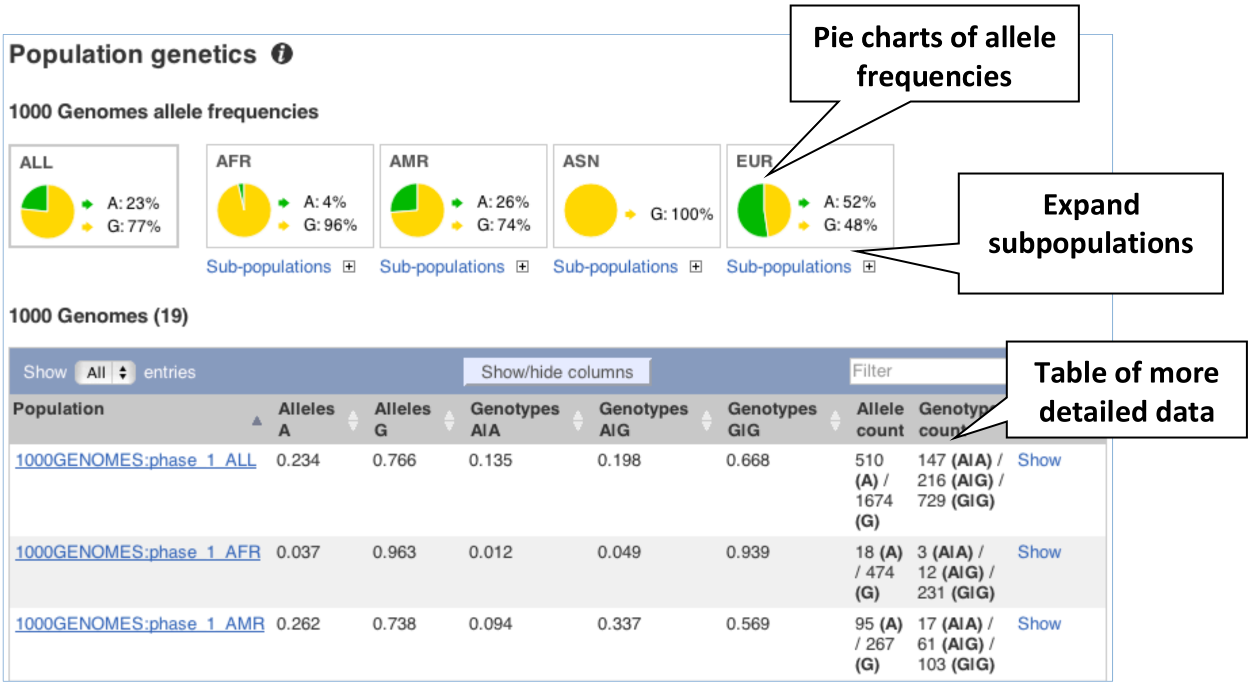
Gene and Transcript consequences

Show/hide columns | Filter

Gene	Transcript (strand)	Allele (transcript allele)	Type	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen	Detail
ENSG00000076003 **HGNC:** MCM6	ENST00000264156 (-) **biotype:** protein_coding	A (T)	Intron variant	-	-	-	-	-	-	-	Show
ENSG00000076003 **HGNC:** MCM6	ENST00000483902 (-) **biotype:** retained_intron	A (T)	NC transcript variant, Intron variant	-	-	-	-	-	-	-	Show
ENSG00000076003 **HGNC:** MCM6	ENST00000492091 (-) **biotype:** processed_transcript	A (T)	NC transcript variant, Intron variant	-	-	-	-	-	-	-	Show

This variant is found in three transcripts of the *MCM6* gene. It has not been associated with any regulatory features or motifs.

Let's look at population genetics. Either click on Explore this variant in the left hand menu then click on the Population genetics icon, or click on Population genetics in the left-hand menu.



Population genetics

1000 Genomes allele frequencies

Pie charts of allele frequencies

Expand subpopulations

Table of more detailed data

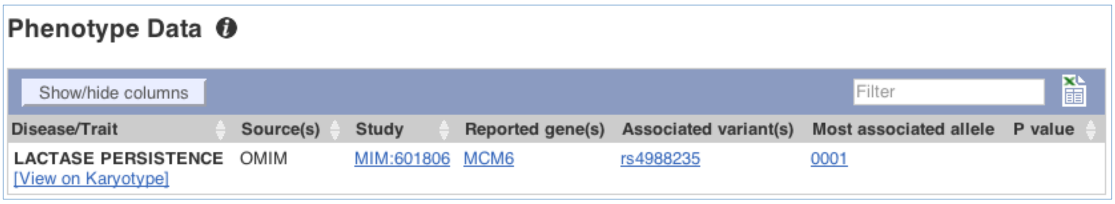
1000 Genomes (19)

Show All entries | Show/hide columns | Filter

Population	Alleles A	Alleles G	Genotypes A|A	Genotypes A|G	Genotypes G|G	Allele count	Genotype count	
1000GENOMES:phase_1_ALL	0.234	0.766	0.135	0.198	0.668	510 (A) / 1674 (G)	147 (A|A) / 216 (A|G) / 729 (G|G)	Show
1000GENOMES:phase_1_AFR	0.037	0.963	0.012	0.049	0.939	18 (A) / 474 (G)	3 (A|A) / 12 (A|G) / 231 (G|G)	Show
1000GENOMES:phase_1_AMR	0.262	0.738	0.094	0.337	0.569	95 (A) / 267 (G)	17 (A|A) / 61 (A|G) / 103 (G|G)	Show

These data are mostly from the **1000 genomes** and **HapMap** projects in human.

There are big differences in allele frequencies between populations. Let’s have a look at the phenotypes associated with this variant to see if they are known to be specific to certain human populations. Either click on Explore this variant in the left hand menu then click on the Phenotype data icon, or click on Phenotype Data in the left-hand menu.



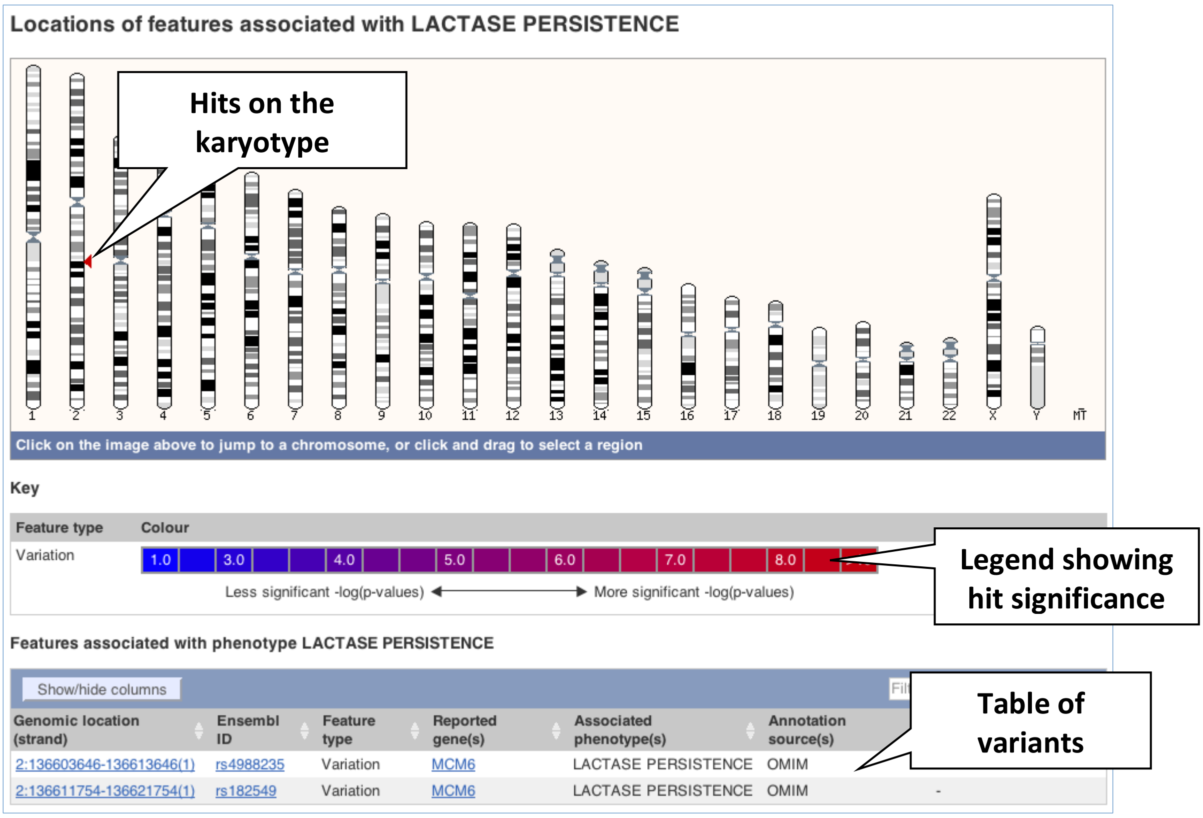
Phenotype Data

Show/hide columns | Filter

Disease/Trait	Source(s)	Study	Reported gene(s)	Associated variant(s)	Most associated allele	P value
LACTASE PERSISTENCE [View on Karyotype]	OMIM	MIM:601806	MCM6	rs4988235	0001	

This variant is associated with lactase persistence, which is known to be common in European populations, and rare in Asian populations, exactly as we saw in the allele frequencies in these populations.

Are there other variants in the genome that also cause lactase persistence? Click on View on Karyotype to find out.



Locations of features associated with LACTASE PERSISTENCE

Hits on the karyotype

Click on the image above to jump to a chromosome, or click and drag to select a region

Key

Feature type	Colour
Variation	1.0 3.0 4.0 5.0 6.0 7.0 8.0

Less significant -log(p-values) ←→ More significant -log(p-values)

Legend showing hit significance

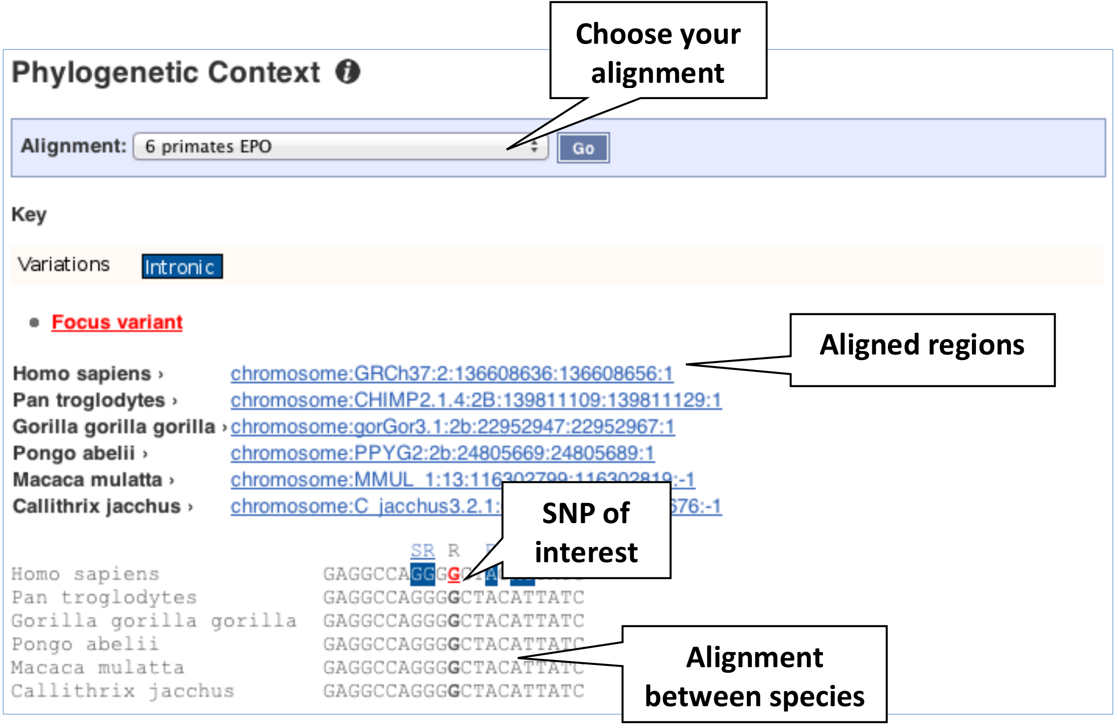
Features associated with phenotype LACTASE PERSISTENCE

Genomic location (strand)	Ensembl ID	Feature type	Reported gene(s)	Associated phenotype(s)	Annotation source(s)	
2:136603646-136613646(1)	rs4988235	Variation	MCM6	LACTASE PERSISTENCE	OMIM	
2:136611754-136621754(1)	rs182549	Variation	MCM6	LACTASE PERSISTENCE	OMIM	-

Table of variants

Two variants are known to be associated with this phenotype. Both are found with the *MCM6* gene.

Click back to the Variation Tab. Click on Phylogenetic context to see the variant in other species.



Phylogenetic Context

Choose your alignment

Alignment: 6 primates EPO

Key

Variations Intronic

- Focus variant

Aligned regions

Homo sapiens › chromosome:GRCh37:2:136608636:136608656:1
Pan troglodytes › chromosome:CHIMP2.1.4:2B:139811109:139811129:1
Gorilla gorilla gorilla › chromosome:gorGor3.1:2b:22952947:22952967:1
Pongo abelii › chromosome:PPYG2:2b:24805669:24805689:1
Macaca mulatta › chromosome:MMUL_1:13:116302799:116302819:-1
Callithrix jacchus › chromosome:C_jacchus3.2.1: 676:-1

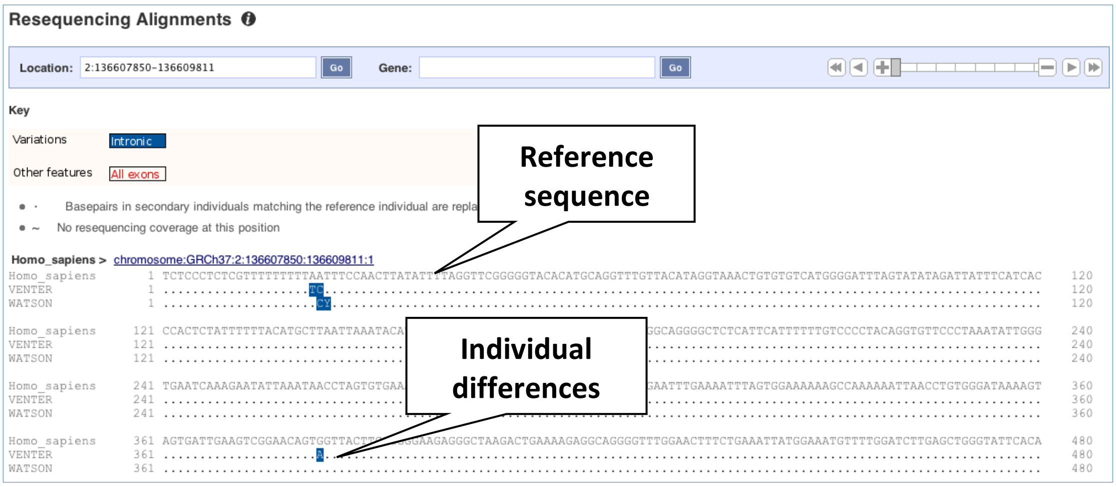
SNP of interest

```
                              SR R
Homo sapiens                  GAGGCCAGGGG
Pan troglodytes               GAGGCCAGGGGCTACATTATC
Gorilla gorilla gorilla       GAGGCCAGGGGCTACATTATC
Pongo abelii                  GAGGCCAGGGGCTACATTATC
Macaca mulatta                GAGGCCAGGGGCTACATTATC
Callithrix jacchus            GAGGCCAGGGGCTACATTATC
```

Alignment between species

The variant is not marked in the other species. This means that the variant arose in humans.

Another way to look at variation is using the **Resequencing** view. Click on the Location tab, then choose Resequencing from the left-hand menu.



Resequencing Alignments

Location: 2:136607850-136609811 Go Gene: Go

Key

Variations	Intronic
Other features	All exons

- · Basepairs in secondary individuals matching the reference individual are repla
- ~ No resequencing coverage at this position

Homo_sapiens > chromosome:GRCh37:2:136607850:136609811:1

```
Homo_sapiens      1 TCTCCCTCTCGTTTTTTTTTAATTTCCAACTTATATTTTAGGTTCGGGGGTACACATGCAGGTTTGTTACATAGGTAAACTGTGTGTCATGGGGATTTAGTATATAGATTATTTCATCAC  120
VENTER            1 ....................TC..................................................................................................  120
WATSON            1 .....................CY.................................................................................................  120

Homo_sapiens    121 CCACTCTATTTTTTACATGCTTAATTAAATACA                    GGCAGGGGCTCTCATTCATTTTTTGTCCCCTACAGGTGTTCCCTAAATATTGGG  240
VENTER          121 .................................                    ......................................................  240
WATSON          121 .................................                    ......................................................  240

Homo_sapiens    241 TGAATCAAAGAATATTAAATAACCTAGTGTGAA                    GAATTTGAAAATTTAGTGGAAAAAAGCCAAAAAATTAACCTGTGGGATAAAAGT  360
VENTER          241 .................................                    ......................................................  360
WATSON          241 .................................                    ......................................................  360

Homo_sapiens    361 AGTGATTGAAGTCGGAACAGTGGTTACTT    GGGAAGAGGGCTAAGACTGAAAAGAGGCAGGGGTTTGGAACTTTCTGAAATTATGGAAATGTTTTGGATCTTGAGCTGGGTATTCACA  480
VENTER          361 .....................A..................................................................................................  480
WATSON          361 ........................................................................................................................  480
```

This view is used to look at sequence between individuals. Craig Venter and James Watson are shown by default. You can change the individuals shown by clicking on Configure this page.

It can also be used to look at different mouse strains or dog breeds. Can you find our variant rs4988235 in this view? Is the alternate allele present in James Watson or Craig Venter?

Exercises: Finding variants in Ensembl

Exercise 16 – Human population genetics and phenotype data

The SNP rs1738074 in the 5’ UTR of the human *TAGAP* gene has been identified as a genetic risk factor for a few diseases.

(a) In which transcripts is this SNP found?

(b) What is the least frequent genotype for this SNP in the Yoruba (YRI) population from the HapMap set?

(c) What is the ancestral allele? Is it conserved in the 36 eutherian mammals?

(d) With which diseases is this SNP associated? Are there any known risk (or associated) alleles?

Exercise 17 – Exploring a SNP in human

The missense variation rs1801133 in the human *MTHFR* gene has been linked to elevated levels of homocysteine, an amino acid whose plasma concentration seems to be associated with the risk of cardiovascular diseases, neural tube defects, and loss of cognitive function. This SNP is also referred to as 'A222V', 'Ala222Val' as well as other HGVS names.

(a) Find the page with information for rs1801133.

(Note: a bug in the current release means that these alleles are erroneously reported as G/A/CT/CT. Please ignore these extra CT alleles.)

(b) Is rs1801133 a Missense variation in all transcripts of the *MTHFR* gene?

(c) Why are the alleles for this variation in Ensembl given as G/A and not as C/T, as in dbSNP and literature? (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=1801133)

(d) What is the major allele in rs1801133?

(e) In which paper is the association between rs1801133 and homocysteine levels described?

(f) According to the data imported from dbSNP, the ancestral allele for rs1801133 is G. Ancestral alleles in dbSNP are based on a comparison between human and chimp. Does the sequence at this same position in four other primates, i.e. gorilla, orangutan, macaque and marmoset, confirm that the ancestral allele is G?

(g) Were both alleles of rs1801133 already present in Neanderthal? To answer this question, have a look at the individual reads at its genomic position in the Neanderthal Genome Browser (http://neandertal.ensemblgenomes.org/).

Exercise 18 – Structural variation in human

In the paper 'The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility' (Gonzalez *et al* Science. 2005 Mar 4; 307(5704):1434-40) it is shown that a higher copy number of the *CCL3L1* (Chemokine (C-C motif) ligand 3-like 1) gene is associated with lower susceptibility to HIV infection.

(a) Find the human *CCL3L1* gene.

(b) Have any CNVs been annotated for this gene? Note: In Ensembl, CNVs are classified as structural variants**.**

Exercise 19 – Exploring a SNP in mouse

Madsen *et al* in the paper 'Altered metabolic signature in pre-diabetic NOD mice' (PloS One. 2012; 7(4): e35445) have described several regulatory and coding SNPs, some of them in genes residing within the previously defined *insulin dependent diabetes (IDD)* regions. The authors describe that one of the identified SNPs in the murine *Xdh* gene (rs29522348) would lead to an amino acid substitution and could be damaging as predicted as by SIFT (http://sift.jcvi.org/).

(a) Which chromosome and coordinates in the SNP located?

(b) What is the HGVS recommendation nomenclature for this SNP?

(c) Why does Ensembl put the C allele first (C/T)?

(d) Are there differences between the genotypes reported in NOD/LTJ and BALB/cByJ?

Demo: The Variant Effect Predictor (VEP)

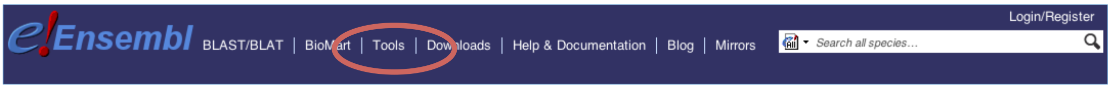
We have analysed a samples from a patient with a genetic disorder. The patient presents with facial and limb deformities, mental

retardation and gastrointestinal reflux. Our genotyping has identified a mutation that may be responsible for the phenotype:
An A->G mutation on chromosome 5 at 37,017,205 on the + strand.

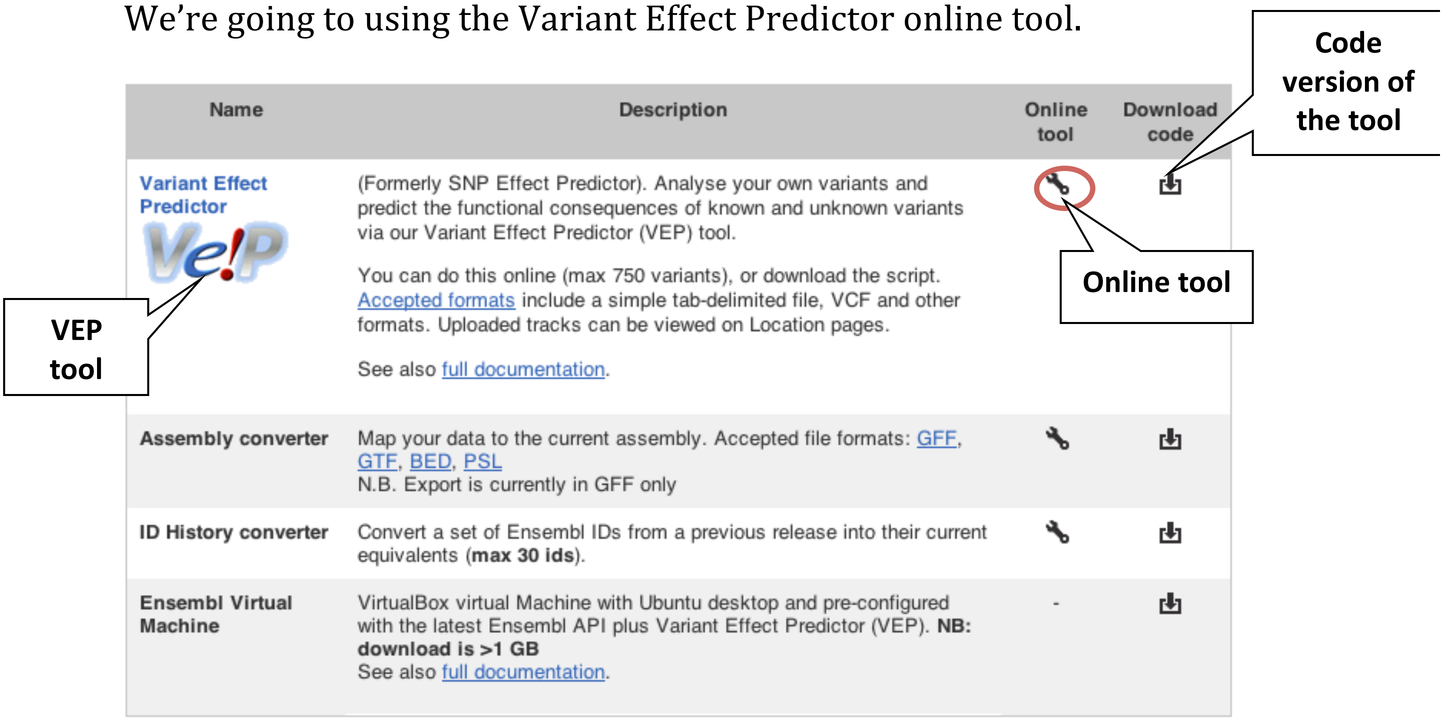
We will use the **Ensembl VEP** to determine:

- Has my variant already been annotated in Ensembl?
- What genes are affected by my variant?
- Does my variant result in a protein change?

From any page in Ensembl click on Tools in the top bar.



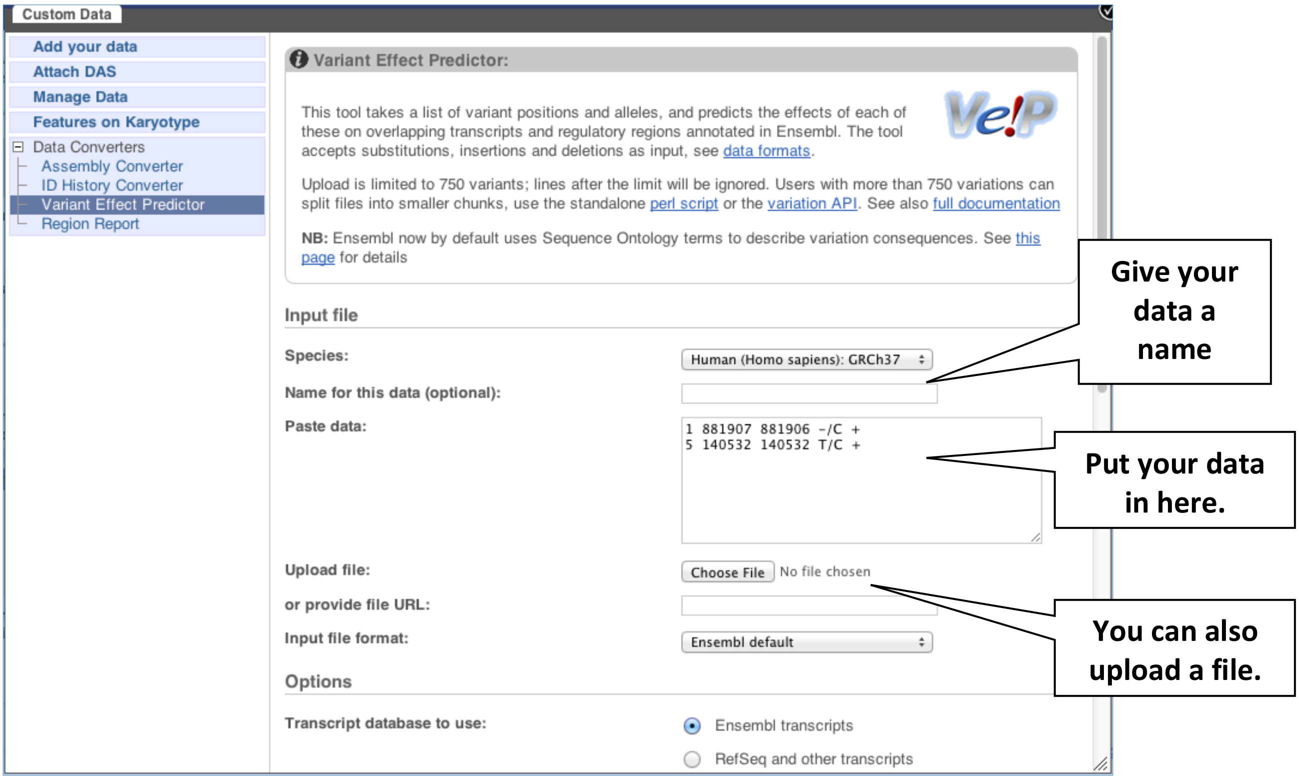
This provides a table showing the tools that are available in Ensembl. There's a short description of what each of the tools do. All are available as both an online tool, and code that you can download. We're going to using the Variant Effect Predictor online tool.



Name	Description	Online tool	Download code
Variant Effect Predictor	(Formerly SNP Effect Predictor). Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool. You can do this online (max 750 variants), or download the script. Accepted formats include a simple tab-delimited file, VCF and other formats. Uploaded tracks can be viewed on Location pages. See also full documentation.	🔧	⤓
Assembly converter	Map your data to the current assembly. Accepted file formats: GFF, GTF, BED, PSL N.B. Export is currently in GFF only	🔧	⤓
ID History converter	Convert a set of Ensembl IDs from a previous release into their current equivalents (**max 30 ids**).	🔧	⤓
Ensembl Virtual Machine	VirtualBox virtual Machine with Ubuntu desktop and pre-configured with the latest Ensembl API plus Variant Effect Predictor (VEP). **NB: download is >1 GB** See also full documentation.	-	⤓

Click on the spanner 🔧 beside the Variant Effect Predictor (VEP).

This will open up a dialogue box. This allows us to input data on our variant.



Variant Effect Predictor:

This tool takes a list of variant positions and alleles, and predicts the effects of each of these on overlapping transcripts and regulatory regions annotated in Ensembl. The tool accepts substitutions, insertions and deletions as input, see data formats.

Upload is limited to 750 variants; lines after the limit will be ignored. Users with more than 750 variations can split files into smaller chunks, use the standalone perl script or the variation API. See also full documentation

NB: Ensembl now by default uses Sequence Ontology terms to describe variation consequences. See this page for details

Input file

Species: Human (Homo sapiens): GRCh37

Name for this data (optional):

Paste data:

```
1 881907 881906 -/C +
5 140532 140532 T/C +
```

Upload file:

or provide file URL:

Input file format: Ensembl default

Options

Transcript database to use: ● Ensembl transcripts ○ RefSeq and other transcripts

Give your data a name

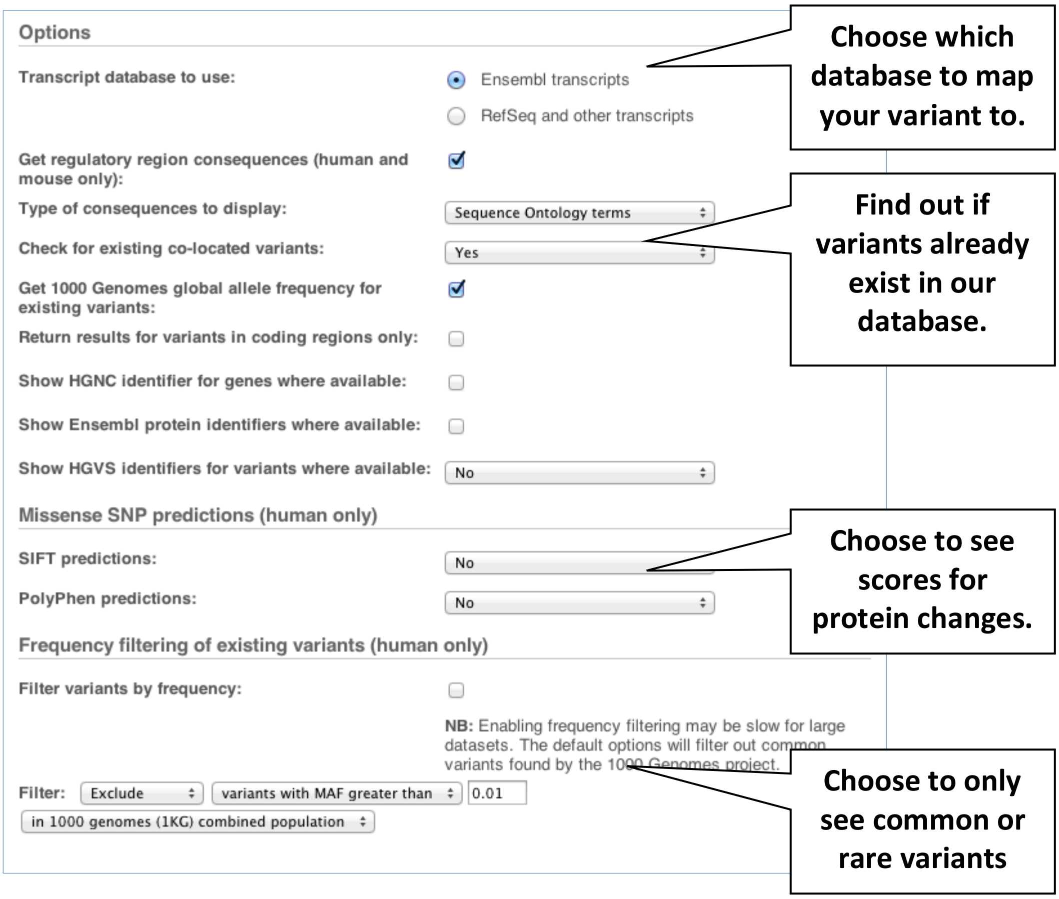
Put your data in here.

You can also upload a file.

The data is in the format:
Chromosome Start End alleles (reference/mutation) strand

Delete the writing already in the Paste data box and type in:
5 37017205 37017205 A/G +

Scroll down to see some of the options we can also choose.



Options

Transcript database to use: ● Ensembl transcripts ○ RefSeq and other transcripts

Get regulatory region consequences (human and mouse only): ☑

Type of consequences to display: Sequence Ontology terms

Check for existing co-located variants: Yes

Get 1000 Genomes global allele frequency for existing variants: ☑

Return results for variants in coding regions only: □

Show HGNC identifier for genes where available: □

Show Ensembl protein identifiers where available: □

Show HGVS identifiers for variants where available: No

Missense SNP predictions (human only)

SIFT predictions: No

PolyPhen predictions: No

Frequency filtering of existing variants (human only)

Filter variants by frequency: □

NB: Enabling frequency filtering may be slow for large datasets. The default options will filter out common variants found by the 1000 Genomes project.

Filter: Exclude | variants with MAF greater than | 0.01 | in 1000 genomes (1KG) combined population

Choose which database to map your variant to.

Find out if variants already exist in our database.

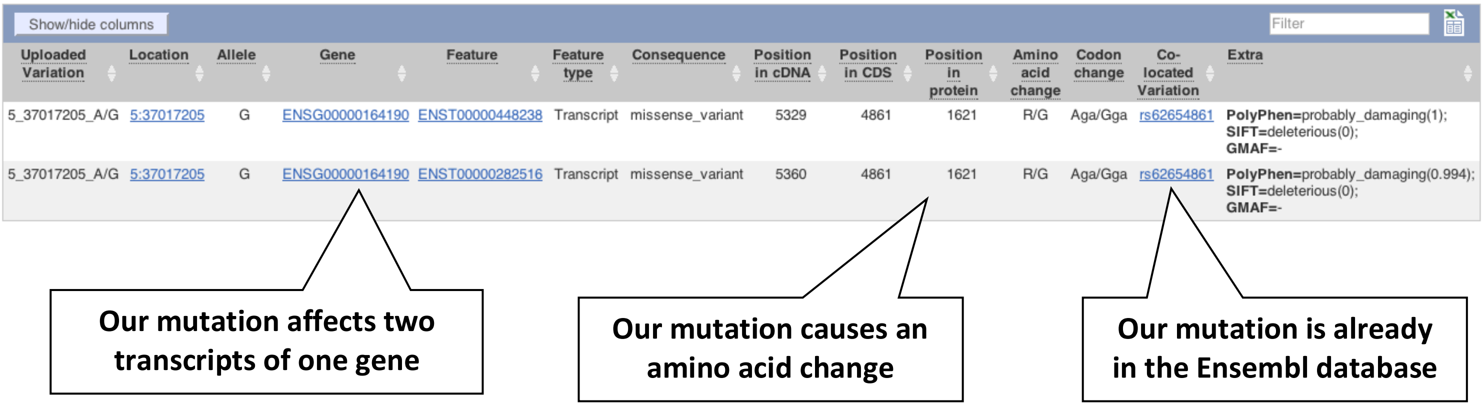
Choose to see scores for protein changes.

Choose to only see common or rare variants

Select Prediction and Score for SIFT predictions and PolyPhen predictions. These are algorithms that predict how deleterious a mutation will be on a protein.

When you've selected everything you need, scroll right to the bottom and click Next.

Click HTML to view your results with clickable links.



Uploaded Variation	Location	Allele	Gene	Feature	Feature type	Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
5_37017205_A/G	5:37017205	G	ENSG00000164190	ENST00000448238	Transcript	missense_variant	5329	4861	1621	R/G	Aga/Gga	rs62654861	**PolyPhen**=probably_damaging(1); **SIFT**=deleterious(0); **GMAF**=-
5_37017205_A/G	5:37017205	G	ENSG00000164190	ENST00000282516	Transcript	missense_variant	5360	4861	1621	R/G	Aga/Gga	rs62654861	**PolyPhen**=probably_damaging(0.994); **SIFT**=deleterious(0); **GMAF**=-

Our mutation affects two transcripts of one gene

Our mutation causes an amino acid change

Our mutation is already in the Ensembl database

Exercise: The Variant Effect Predictor (VEP)

Exercise 20 – VEP

Resequencing of the genomic region of the human *CFTR* (cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) gene (ENSG00000001626) has revealed the following variants (alleles defined in the forward strand):

- G/A at 7:117,171,039
- T/C at 7:117,171,092
- T/C at 7:117,171,122

(a) Use the VEP tool in Ensembl and choose the options to see SIFT and PolyPhen predictions. Do these variants result in a change in the proteins encoded by any of the Ensembl genes? Which gene? Have the variants already been found?

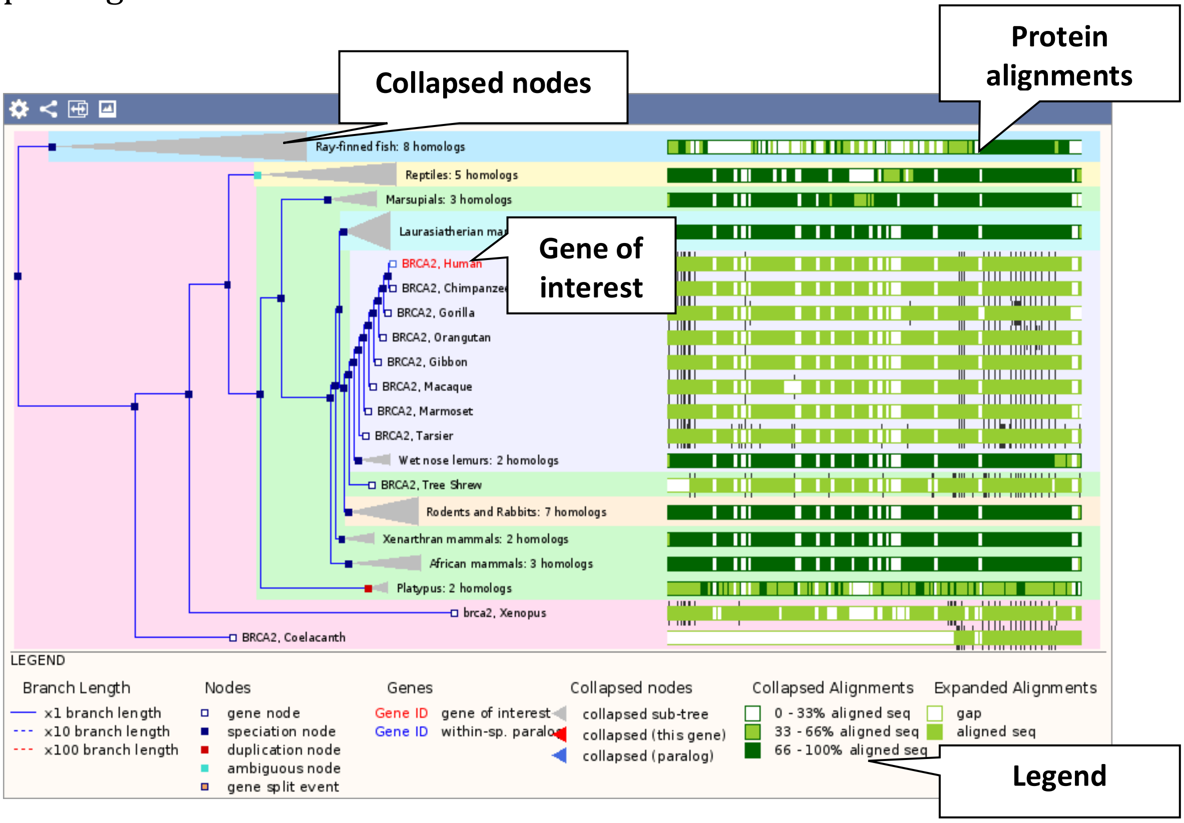
(b) Go to Region in detail for *CFTR*. Do you see the VEP track?

Comparative genomics

Demo: Gene trees and homologues

Let's look at the homologues of human *BRCA2*. Search for the gene and go to the **Gene tab**.

Click on Gene tree (image), which will display the current gene in the context of a phylogenetic tree used to determine orthologues and paralogues.



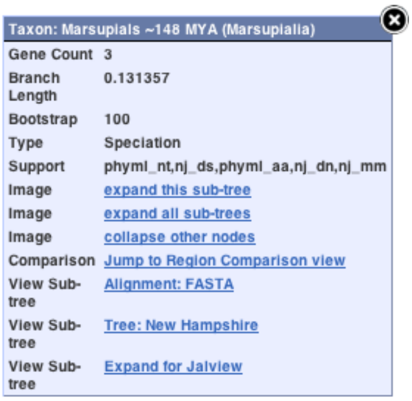
Collapsed nodes

Protein alignments

Gene of interest

Legend

Funnels indicate collapsed nodes. We can expand them by clicking on the node and selecting Expand this sub-tree from the pop-up menu.

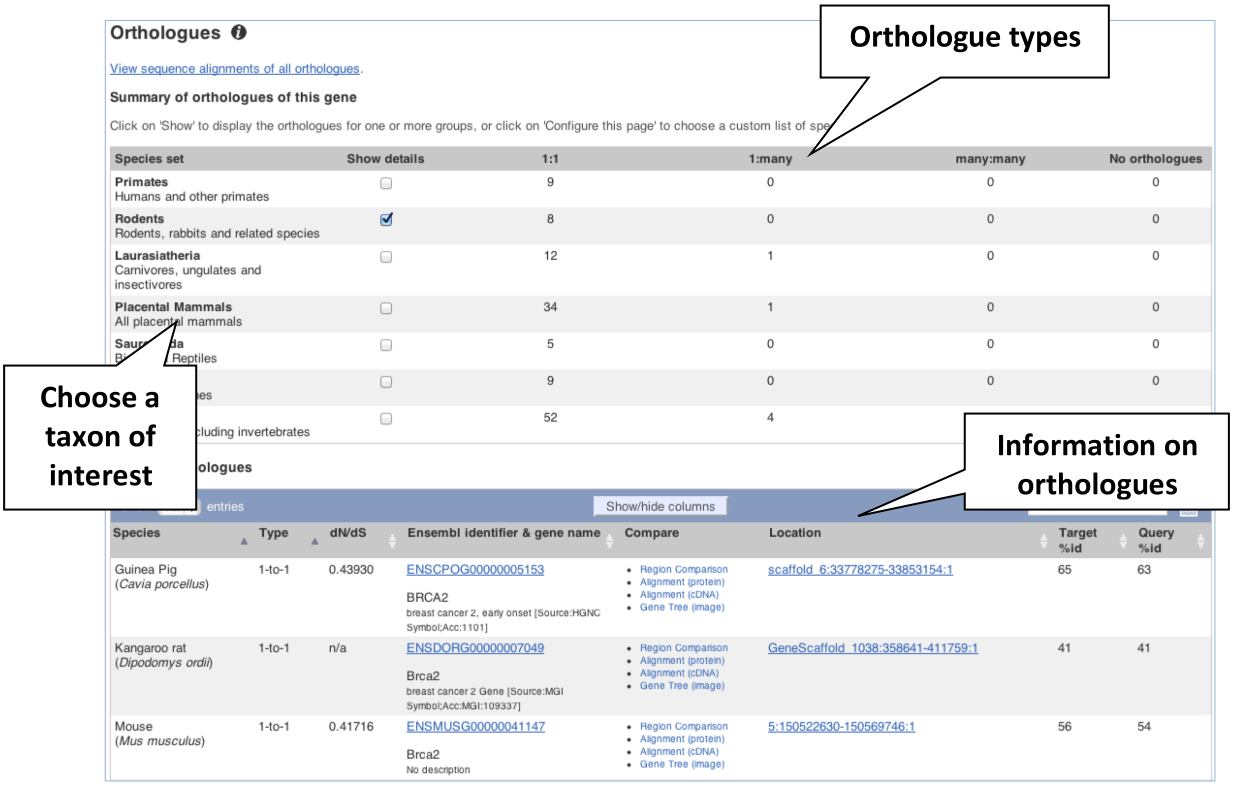


Taxon: Marsupials ~148 MYA (Marsupialia)

Gene Count	3
Branch Length	0.131357
Bootstrap	100
Type	Speciation
Support	phyml_nt,nj_ds,phyml_aa,nj_dn,nj_mm
Image	expand this sub-tree
Image	expand all sub-trees
Image	collapse other nodes
Comparison	Jump to Region Comparison view
View Sub-tree	Alignment: FASTA
View Sub-tree	Tree: New Hampshire
View Sub-tree	Expand for Jalview

We can look at homologues in the Orthologues and Paralogues pages, which can be accessed from the left-hand menu. The numbers of orthologues or paralogues available are indicated in brackets alongside the name. If there are none, then the name will be greyed out. Paralogues is greyed out for *BRCA2* indicating that there are no paralogues available.

Click on Orthologues to see the 56 orthologues available.



Orthologue types

Choose a taxon of interest

Information on orthologues

Orthologues

View 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 spe

Species set	Show details	1:1	1:many	many:many	No orthologues
Primates Humans and other primates	□	9	0	0	0
Rodents Rodents, rabbits and related species	☑	8	0	0	0
Laurasiatheria Carnivores, ungulates and insectivores	□	12	1	0	0
Placental Mammals All placental mammals	□	34	1	0	0
Saur da Bi Reptiles	□	5	0	0	0
es	□	9	0	0	0
cluding invertebrates	□	52	4		

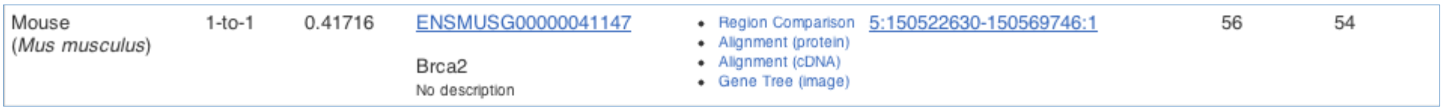
ologues

entries

Show/hide columns

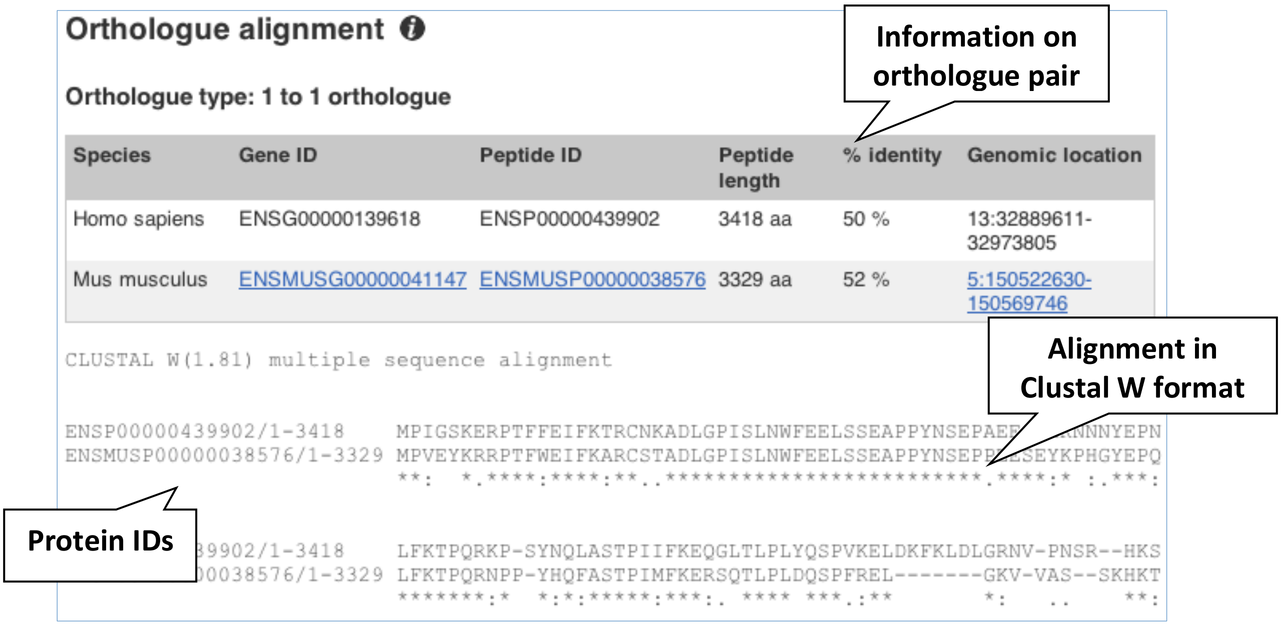
Species	Type	dN/dS	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Guinea Pig (*Cavia porcellus*)	1-to-1	0.43930	ENSCPOG00000005153 BRCA2 breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	scaffold_6:33778275-33853154:1	65	63
Kangaroo rat (*Dipodomys ordii*)	1-to-1	n/a	ENSDORG00000007049 Brca2 breast cancer 2 Gene [Source:MGI Symbol;Acc:MGI:109337]	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	GeneScaffold_1038:358641-411759:1	41	41
Mouse (*Mus musculus*)	1-to-1	0.41716	ENSMUSG00000041147 Brca2 No description	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	5:150522630-150569746:1	56	54

Choose to see only **Rodent** orthologues by selecting the box. The table below will now only show details of rodent orthologues. Let's look at mouse.



Mouse (*Mus musculus*)	1-to-1	0.41716	ENSMUSG00000041147 Brca2 No description	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	5:150522630-150569746:1	56	54

Links from the orthologue allow you to go to alignments of the orthologous proteins and cDNAs. Click on Alignment (protein) for the mouse orthologue.



Orthologue alignment

Orthologue type: 1 to 1 orthologue

Species	Gene ID	Peptide ID	Peptide length	% identity	Genomic location
Homo sapiens	ENSG00000139618	ENSP00000439902	3418 aa	50 %	13:32889611-32973805
Mus musculus	ENSMUSG00000041147	ENSMUSP00000038576	3329 aa	52 %	5:150522630-150569746

```
CLUSTAL W(1.81) multiple sequence alignment


ENSP00000439902/1-3418     MPIGSKERPTFFEIFKTRCNKADLGPISLNWFEELSSEAPPYNSEPAEE    NNNYEPN
ENSMUSP00000038576/1-3329 MPVEYKRRPTFWEIFKARCSTADLGPISLNWFEELSSEAPPYNSEPP  ESEYKPHGYEPQ
                           **:  *.****:****:**..************************.****:*  :.***:


         39902/1-3418     LFKTPQRKP-SYNQLASTPIIFKEQGLTLPLYQSPVKELDKFKLDLGRNV-PNSR--HKS
         00038576/1-3329 LFKTPQRNPP-YHQFASTPIMFKERSQTLPLDQSPFREL-------GKV-VAS--SKHKT
                           *******:*  *:*:*****:***:. **** ***.:**       *:   ..   **:
```

Exercises: Gene trees and homologues

Exercise 21 - Orthologues, paralogues and genetrees for the human *BRAF* gene.

(a) How many orthologues are predicted for this gene in primates? Note the Target %id and Query %id.
How much sequence identity does the *Tarsius syrichta* protein have to the human one? Click on the Alignment link next to the Ensembl identifier column to view a protein alignment in Clustal format.

(b) Go to the orthologue in marmoset. Is there a genomic alignment between marmoset and human? Is there a gene for both species in this region?

Exercise 22 – Zebrafish orthologues

Go to www.ensembl.org to find the *dbh* gene on the zebrafish genome.
(a) Go to the Location page for this gene. View the Alignments (image) and Alignments (text) for the 5 teleost fish. Which fish genomes are represented in the alignment? Do all the fish show a gene in these alignments?

(b) Export the alignments (as Clustal).

(c) Click on the Region in detail link at the left and turn on the tracks for multiple alignments and conservation score for the 5 teleost fish EPO by configuring the page.

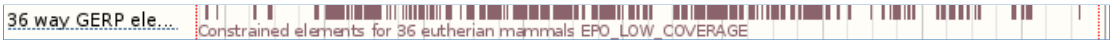
What is the difference between the 5 teleost fish EPO multiple alignment track and the Constrained elements already turned on by default? Which regions of the gene, do most of the constrained element blocks match up to?

Can you find more information on how the constrained elements track was generated?

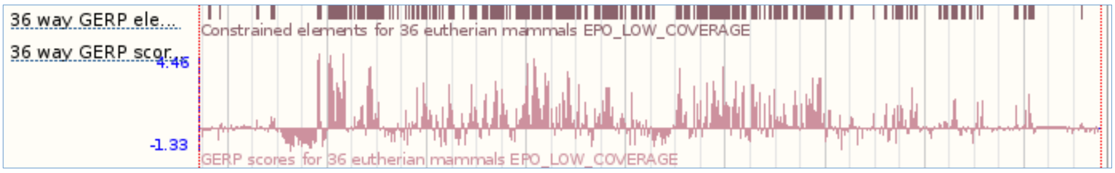
Demo: Whole genome alignments

Let's look at some of the comparative genomics views in the Location tab. Go to the region 2:176914144-177094980 in human, which contains the *HoxD* cluster which is involved in limb development and is highly conserved between species.

In the **Region in detail** view, we can already see the Constrained elements for 36 eutherian mammals EPO_LOW_COVERAGE track by default. This track indicates regions of high conservation between species, considered to be "constrained" by evolution.


36 way GERP ele...
Constrained elements for 36 eutherian mammals EPO_LOW_COVERAGE

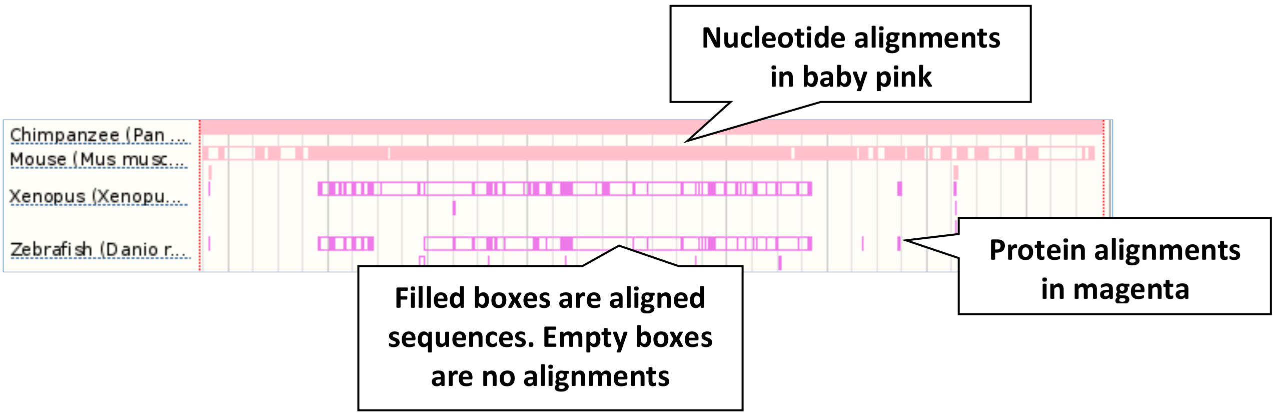
This track has a matching conservation score track. Click on Configure this page, then Comparative genomics and turn on the track for Conservation score for 36 eutherian mammals EPO_LOW_COVERAGE. Save and close the menu.


36 way GERP ele...
Constrained elements for 36 eutherian mammals EPO_LOW_COVERAGE
36 way GERP scor...
4.46
-1.33
GERP scores for 36 eutherian mammals EPO_LOW_COVERAGE

You can now see the conservation scores that were used to determine the peaks indicated in the constrained elements track.

We can also look at individual species comparative genomics tracks in this view by clicking on Configure this page.

Select BLASTz/LASTz alignments from the left-hand menu to choose alignments between closely related species. Turn on the alignments for Mouse and Chimpanzee in Normal. Go to Translated blat alignments and turn on alignments with Zebrafish and Xenopus in Normal. Save and close the menu.



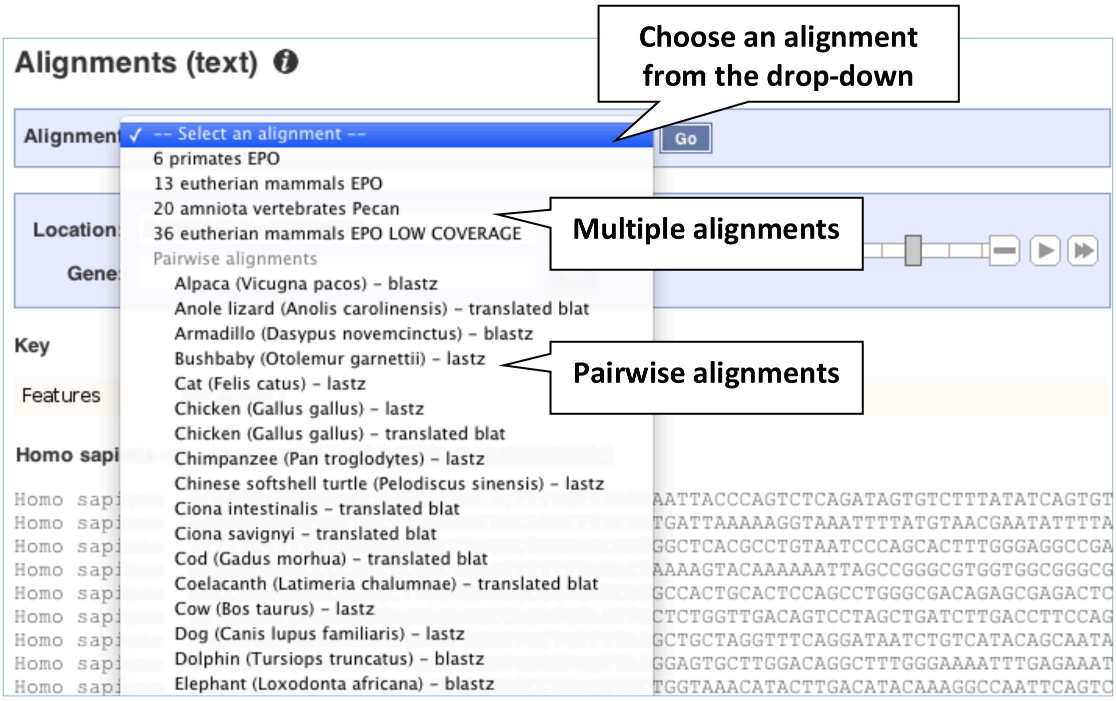
Nucleotide alignments in baby pink

Protein alignments in magenta

Filled boxes are aligned sequences. Empty boxes are no alignments

The alignment is greatest between closely related species.

We can also look at the alignment between species or groups of species as text. Click on Alignments (text) in the left hand menu.


Alignments (text)
Alignment
✓ -- Select an alignment --
6 primates EPO
13 eutherian mammals EPO
20 amniota vertebrates Pecan
36 eutherian mammals EPO LOW COVERAGE
Pairwise alignments
Alpaca (Vicugna pacos) – blastz
Anole lizard (Anolis carolinensis) – translated blat
Armadillo (Dasypus novemcinctus) – blastz
Bushbaby (Otolemur garnettii) – lastz
Cat (Felis catus) – lastz
Chicken (Gallus gallus) – lastz
Chicken (Gallus gallus) – translated blat
Chimpanzee (Pan troglodytes) – lastz
Chinese softshell turtle (Pelodiscus sinensis) – lastz
Ciona intestinalis – translated blat
Ciona savignyi – translated blat
Cod (Gadus morhua) – translated blat
Coelacanth (Latimeria chalumnae) – translated blat
Cow (Bos taurus) – lastz
Dog (Canis lupus familiaris) – lastz
Dolphin (Tursiops truncatus) – blastz
Elephant (Loxodonta africana) – blastz
Go
Location
Gene
Key
Features
Homo sapi
Homo sapi AATTACCCAGTCTCAGATAGTGTCTTTATATCAGTGT
Homo sapi TGATTAAAAAGGTAAATTTTATGTAACGAATATTTTA
Homo sapi GGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGA
Homo sapi AAAAGTACAAAAAATTAGCCGGGCGTGGTGGCGGGCG
Homo sapi GCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTC
Homo sapi CTCTGGTTGACAGTCCTAGCTGATCTTGACCTTCCAG
Homo sapi GCTGCTAGGTTTCAGGATAATCTGTCATACAGCAATA
Homo sapi GGAGTGCTTGGACAGGCTTTGGGAAAATTTGAGAAAT
Homo sapi TGGTAAACATACTTGACATACAAAGGCCAATTCAGTC

Choose an alignment from the drop-down

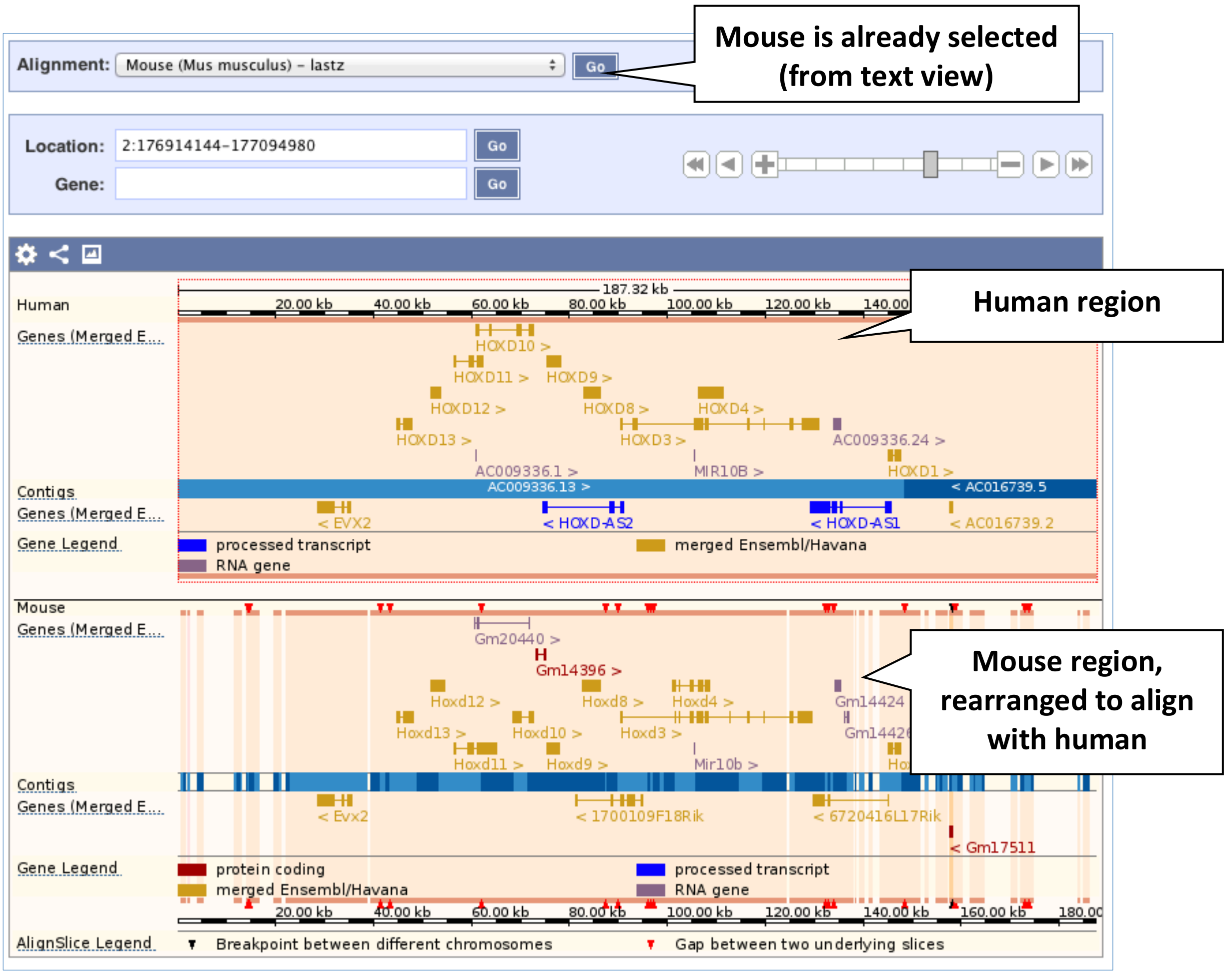
Multiple alignments

Pairwise alignments

Select Mouse from the alignments list then click Go.

You will see a list of the regions aligned, followed by the sequence alignment. Exons are shown in red.

This can also be viewed graphically. Click on Alignments (image) in the left-hand menu.



Alignment: Mouse (Mus musculus) – lastz

Location: 2:176914144-177094980

Gene:

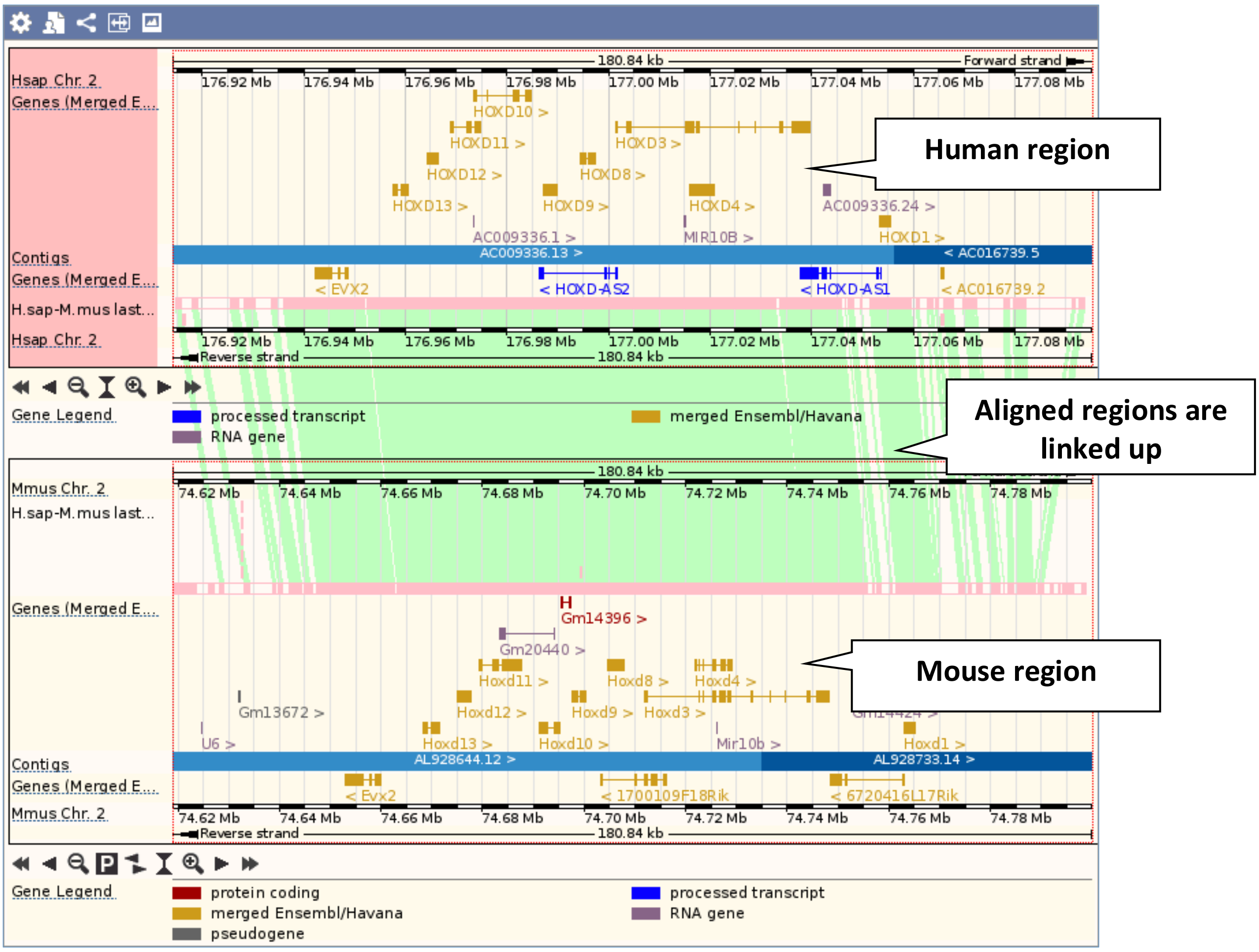
Mouse is already selected (from text view)

Human region

Mouse region, rearranged to align with human

In both alignment views the contig is the compared species is rearranged to align to the species of interest. To compare with both contigs in their natural order, go to Region comparison.

To add species to this view, click on the blue Select species or regions button. Choose Mouse from the list then close the menu.



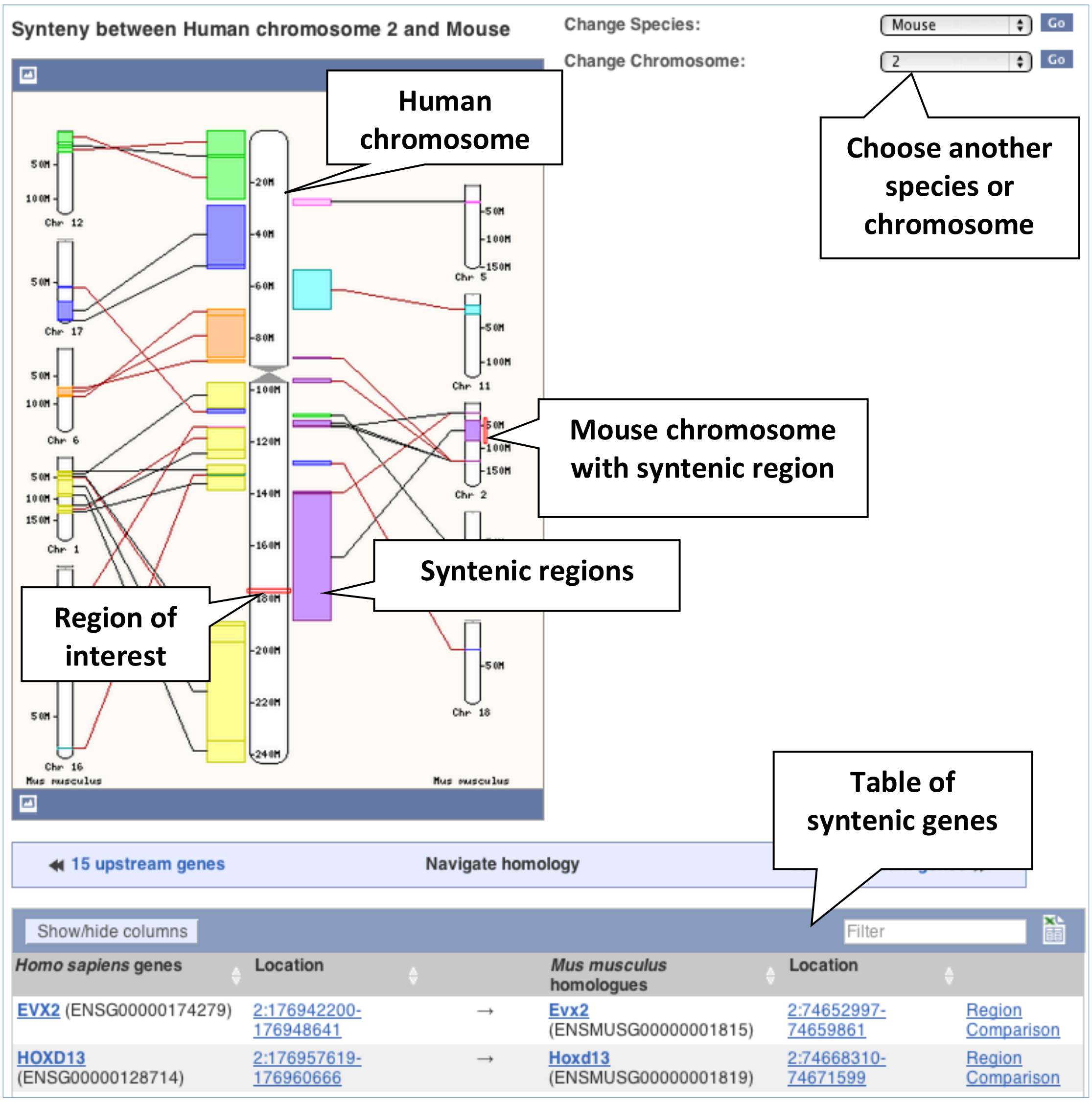
Human region

Aligned regions are linked up

Mouse region

We can view large scale syntenic regions from our chromosome of interest. Click on Synteny in the left hand menu.

]



Synteny between Human chromosome 2 and Mouse

Change Species: Mouse Go

Change Chromosome: 2 Go

Human chromosome

Choose another species or chromosome

Mouse chromosome with syntenic region

Syntenic regions

Region of interest

Table of syntenic genes

15 upstream genes — Navigate homology

Show/hide columns — Filter

Homo sapiens genes	Location		*Mus musculus* homologues	Location	
EVX2 (ENSG00000174279)	2:176942200-176948641	→	Evx2 (ENSMUSG00000001815)	2:74652997-74659861	Region Comparison
HOXD13 (ENSG00000128714)	2:176957619-176960666	→	Hoxd13 (ENSMUSG00000001819)	2:74668310-74671599	Region Comparison

Exercises: Whole genome alignments

Exercise 23 – Synteny

Go to www.ensembl.org
Find the Rhodopsin (*RHO*) gene for Human. Go to the Location tab.

(a) Click Synteny at the left. Are there any syntenic regions in dog? If so, which chromosomes are shown in this view?

(b) Stay in the Synteny view. Is there a homologue in dog for human *RHO*? Are there more genes in this syntenic block with homologues?

Exercise 24 – Whole genome alignments

(a) Find the Ensembl *BRCA2* (Breast cancer type 2 susceptibility protein) gene for human and go to the Region in detail page.

(b) Turn on the BLASTZ alignment tracks for chicken, chimp, mouse and platypus and the Translated BLAT alignment tracks for anole lizard and zebrafish. Does the degree of conservation between human and the various other species reflect their evolutionary relationship? Which parts of the *BRCA2* gene seem to be the most conserved? Did you expect this?

(c) Have a look at the Conservation score and Constrained elements tracks for the set of 36 mammals and the set of 19 vertebrates. Do these tracks confirm what you already saw in the tracks with pairwise alignment data?

(d) Retrieve the genomic alignment for a constrained element. Highlight the bases that match in >50% of the species in the alignment.

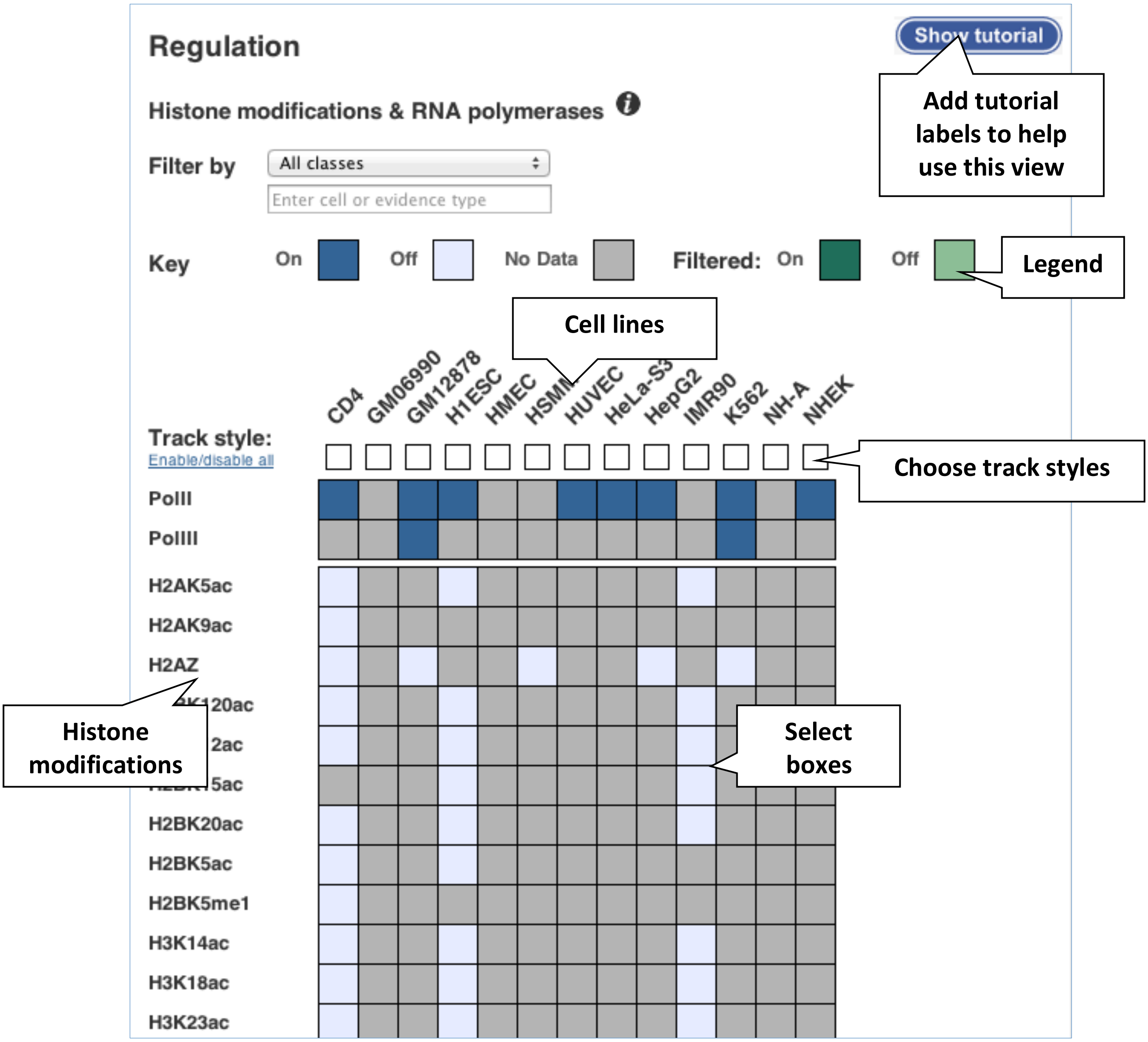
(e) Retrieve the genomic alignment for the *BRCA2* gene for primates. Highlight the bases that match in >50% of the species in the alignment.

Regulation

Demo: Raw ChIPSeq data

We're going to add some regulation data to the **Region in detail** view. We'll start at the human region 11:2012486-2030153, which contains the imprinted *H19* gene.

Add regulation tracks using Configure this page. First, we're going to add ChIP-seq data for histone modifications and polymerase binding. Click on Histones & polymerases under Regulation in the left-hand menu.



Regulation

Show tutorial

Histone modifications & RNA polymerases

Filter by: All classes

Enter cell or evidence type

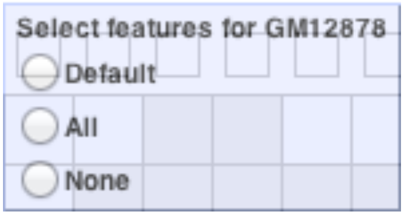
Key: On 🟦 Off ⬜ No Data ▨ Filtered: On 🟩 Off ▩

Track style: Enable/disable all

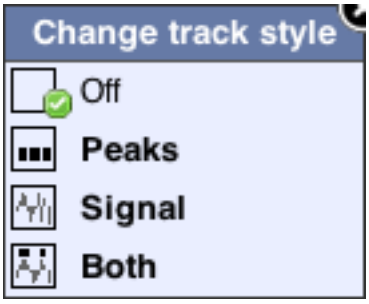
	CD4	GM06990	GM12878	H1ESC	HMEC	HSMM	HUVEC	HeLa-S3	HepG2	IMR90	K562	NH-A	NHEK
	□	□	□	□	□	□	□	□	□	□	□	□	□
PolII	🟦	▨	🟦	🟦	▨	▨	🟦	🟦	🟦	▨	🟦	▨	🟦
PolIII	▨	▨	🟦	▨	▨	▨	▨	▨	▨	▨	🟦	▨	▨
H2AK5ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	▨	▨	▨
H2AK9ac	⬜	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
H2AZ	⬜	▨	⬜	▨	▨	⬜	▨	▨	⬜	▨	⬜	▨	▨
K120ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	▨	▨	▨
2ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	[illegible]	[illegible]	[illegible]
5ac	▨	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	▨	▨	▨
H2BK20ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	▨	▨	▨
H2BK5ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	▨	▨	▨	▨
H2BK5me1	⬜	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
H3K14ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	▨	▨	▨
H3K18ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	▨	▨	▨
H3K23ac	⬜	▨	▨	⬜	▨	▨	▨	▨	▨	⬜	▨	▨	▨

You can turn on a single track by clicking on the box in the matrix. Note that certain tracks are selected for all cell lines by default (PolII, PolIII, H3K27me3, H3K36me3, H3K4me3, H3K9me3). These will appear in the Region in detail view only if you specify a track style for the cell lines.

Turn on all the tracks for GM12878. Hover over the cell line name then select All.

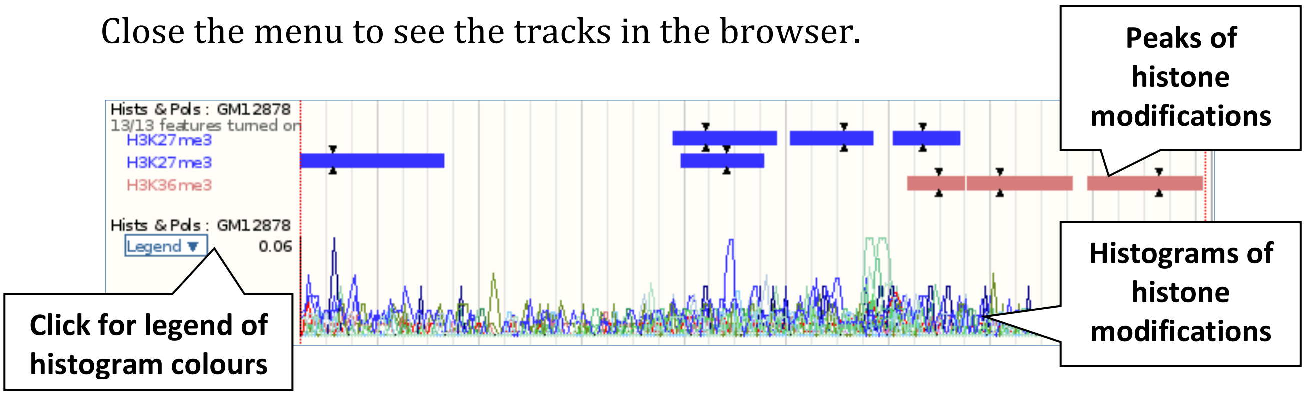


Now choose the track style for the tracks you've switched on. Click on the track style box for GM12878 and select Both.



There is a similar matrix for Open chromatin &TFBS. Use this to turn on all tracks for GM12878 in Both.

Close the menu to see the tracks in the browser.



Peaks of histone modifications

Histograms of histone modifications

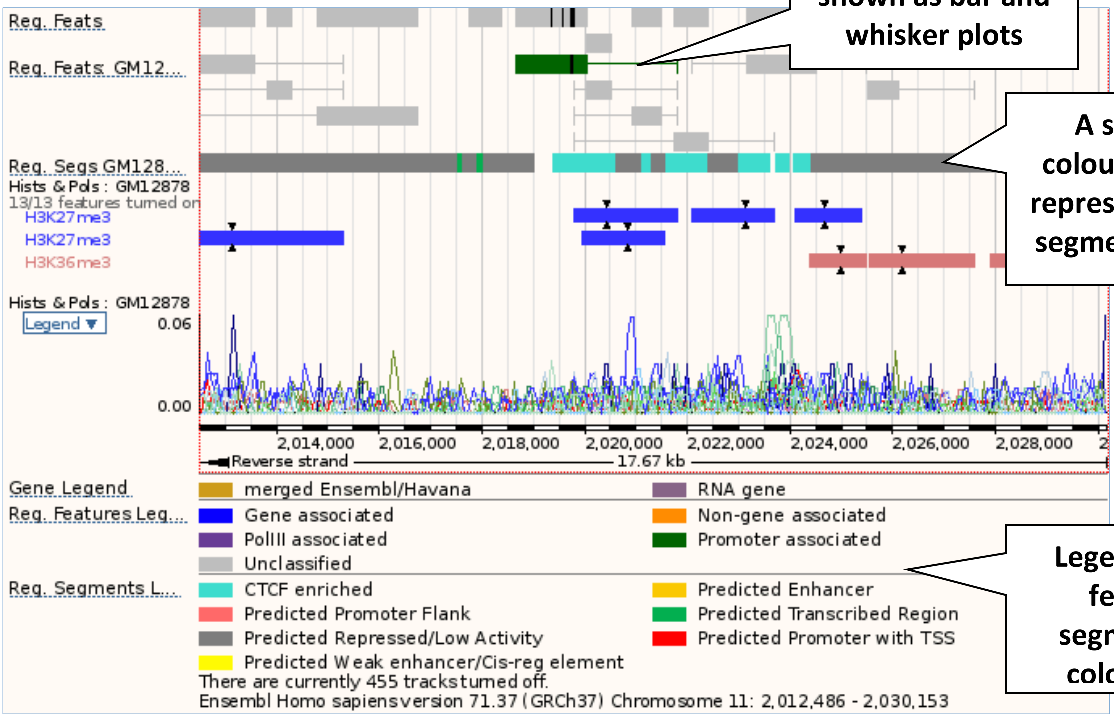
Click for legend of histogram colours

Demo: Regulatory features and segmentation

These data are used to construct the **Reg-feats** and **Segmentation features**. The merged Reg-feats are switched on in the Region in detail view by default.

Click on Configure this page. Then select Regulatory features. Turn on the Reg. Feats: GM12878 and Reg. Segs: GM12878 tracks.

Save and close the menu.



There are currently 455 tracks turned off.

Ensembl Homo sapiens version 71.37 (GRCh37) Chromosome 11: 2,012,486 - 2,030,153

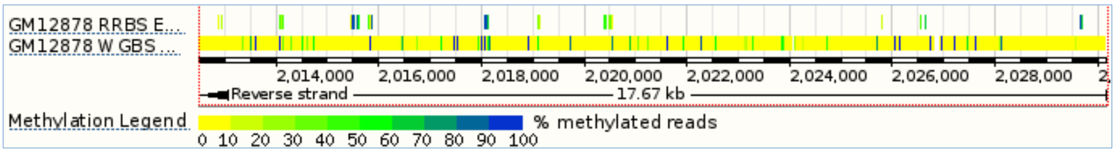
Reg feats are shown as bar and whisker plots

A single coloured bar represents the segmentation

Legends of reg feats and segmentation colour codes

Can you see correlations between the different kinds of regulatory data representation?

You can also add methylation data using Configure this page. Find it under DNA methylation and turn on GM12878 RRBS ENCODE and GM12878 WGBS ENCODE.



Our regulatory data incorporates the ENCODE data. To see the raw ENCODE data and the ENCODE segmentation, you need to add the ENCODE hub.

From ensembl.org, click on the ENCODE icon.

This page contains information about the ENCODE data and how it is incorporated into Ensembl.

Add the ENCODE hub by clicking on the Link to add the ENCODE track hub.

This will take you directly to the matrices for adding ENCODE data to the Region in detail view. The ENCODE matrices work in the same way as the Open chromatin &TFBS and Histones & polymerases matrices, except that some have multiple options (indicated by numbers within the boxes).

Exercises: Regulation

Exercise 25 – Gene regulation: Human *STX7*

(a) Find the Location tab (Region in detail page) for the *STX7* gene. Are there regulatory features in this gene region? If so, where in the gene do they appear?

(b) Click Configure this page and on the Regulatory features menu in the left hand side. Turn on Segmentation features for HUVEC, HeLa-S3, and HepG2 cell types. Do any of these cells show predicted enhancer regions in the *STX7* region?

(c) Use Configure this page to add supporting data indicating open chromatin for HeLa-S3 cells. Are there sites enriched for marks of open chromatin (DNase1 and FAIRE) in HeLa cells at the 5' end of *STX7*?

(d) Configure this page once again to add histone modification supporting data for the same cell type as above (e.g.HeLa-S3). Which ones are present at the 5' end of *STX7*?

(e) Is there any data to support methylated CpG sites in this region (5' end) of *STX7* in B-cells?

(f) Create a Share link for this display. Email it to yourself then open the link.

Exercise 26 – Regulatory features in human

The *HLA-DRB1* and *HLA-DQA1* genes are part of the human major histocompatibility complex class II (MHC-II) region and are located about 44 kb from each other on chromosome 6. In the paper 'The human major histocompatibility complex class II *HLA-DRB1* and *HLA-DQA1* genes are separated by a CTCF-binding enhancer-blocking element' (Majumder *et al* J Biol Chem. 2006 Jul 7;281(27):18435-43) a region of high acetylation located in the intergenic sequences between *HLA-DRB1* and *HLA-DQA1* is described. This region, termed XL9, coincided with sequences that bound the insulator protein CCCTC-binding factor (CTCF). Majumder *et al* hypothesise that the XL9 region may have evolved to separate the transcriptional units of the *HLA-DR* and *HLA-DQ* genes.

(a) Go to the region from 32,540,000 to 32,620,000 bp on human chromosome 6

(b) Is there a regulatory feature annotated in the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes that has CTCF binding supporting data as (part of) its core evidence?

(c) Has the CTCF binding detected at this position been observed in all cell/tissue types analysed?

(d) Have a look at the Regulatory supporting evidence - Histones & Polymerases configuration matrix. For which cell/tissue type are the most histone acetylation data sets available? In this cell/tissue type, is the region that shows CTCF binding also a region of high acetylation, as found by Majumder *et al*?

Quick Guide to Databases and Projects

Here is a list of databases and projects you will come across in these exercises. Google any of these to learn more. Projects include many species, unless otherwise noted.

Other help:
The Ensembl Glossary: http://www.ensembl.org/Help/Glossary
Ensembl FAQs:
http://www.ensembl.org/Help/Faq

SEQUENCES

EMBL-Bank, NCBI GenBank, DDBJ – Contain nucleic acid sequences deposited by submitters such as wet-lab biologists and gene sequencing projects. These three databases are synchronised with each other every day, so the same sequences should be found in each.

CCDS – coding sequences that are agreed upon by Ensembl, VEGA-Havana, UCSC, and NCBI. *(human and mouse).*

NCBI Entrez Gene – NCBI's gene collection
`

NCBI RefSeq – NCBI's collection of 'reference sequences', includes genomic DNA, transcripts and proteins. NM stands for 'Known mRNA' (eg NM_005476) and NP (eg NP_005467) are 'Known proteins'.

UniProtKB – the "Protein knowledgebase", a comprehensive set of protein sequences. Divided into two parts: Swiss-Prot and TrEMBL

UniProt Swiss-Prot – the manually annotated, reviewed protein sequences in the UniProtKB. High quality.

UniProt TrEMBL – the automatically annotated, unreviewed set of proteins (EMBL-Bank translated). Varying quality.

VEGA – Vertebrate Genome Annotation, a selection of manually-curated genes, transcripts, and proteins. *(human, mouse, zebrafish, gorilla, wallaby, pig, and dog).*

VEGA-HAVANA – The main contributor to the VEGA project, located at the Wellcome Trust Sanger Institute, Hinxton, UK.

GENE NAMES

HGNC – HUGO Gene Nomenclature Committee, a project assigning a unique and meaningful name and symbol to every human gene. *(Human).*

ZFIN – The Zebrafish Model Organism Database. Gene names are only one part of this project. *(Z-fish).*

PROTEIN SIGNATURES

InterPro – A collection of domains, motifs, and other protein signatures. Protein signature records are extensive, and combine information from individual projects such as UniProt, along with other databases such as SMART, PFAM and PROSITE (explained below).

PFAM – A collection of protein families

PROSITE – A collection of protein domains, families, and functional sites.

SMART – A collection of evolutionarily conserved protein domains.

OTHER PROJECTS

NCBI dbSNP – A collection of sequence polymorphisms; mainly single nucleotide polymorphisms, along with insertion-deletions.

NCBI OMIM – Online Mendelian Inheritance in Man – a resource showing phenotypes and diseases related to genes *(human).*