# Browsing Genes and Genomes with Ensembl

## Ensembl Browser Workshop

## Academia Sinica

## Taiwan
## 22nd February 2016

Dr Denise Carvalho-Silva

# Notes:

This workshop is based on Ensembl release 83 (December 2015) and Ensembl Genomes release 30 (December 2015)

Some useful links:

1) Ensembl Browser website
www.ensembl.org

2) Ensembl Genomes Browser website
www.ensemblgenomes.org

3) Workshop materials (in pdf)
http://www.ebi.ac.uk/~denise/taiwan/sinica

Feel free to tackle questions relative to your own research instead of following the ones provided in our course booklet. The answers for the latter can be found here:

http://www.ebi.ac.uk/~denise/taiwan/answers

More exercises? http://tinyurl.com/e-exercises

Questions or comments?

helpdesk@ensembl.org
helpdesk@ensemblgenomes.org

# TABLE OF CONTENTS

# OVERVIEW

Ensembl provides annotation of genes and other genomic features such as sequence variants, conserved regions across species, and regulatory regions. The Ensembl gene set of the human, mouse and zebrafish genomes is based on protein and nucleotide evidence annotated by both automatic and manual means (the latter carried out by the **Havana** group).

All the data are freely available and can be accessed via the web browser or programmatically via our APIs (PERL or otherwise). 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.

Points covered:

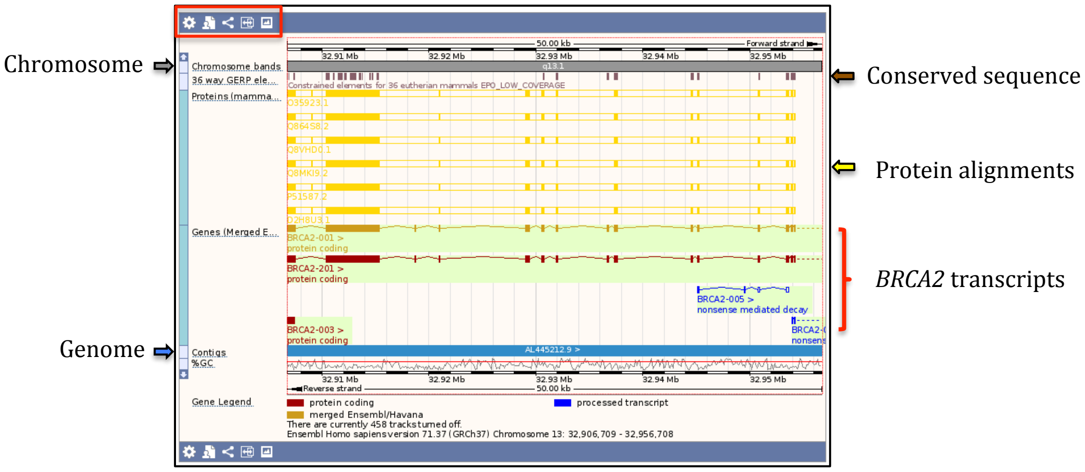
- The need for genome browsers
- An introduction to the Ensembl browser
- Accessing genomic data in Ensembl
- An overview of Ensembl tools

e! Check our video tutorial!
http://www.youtube.com/user/EnsemblHelpdesk

The Ensembl Genome browser
Introduction to BioMart

# INTRODUCTION TO ENSEMBL

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 (e.g. yeast, fruitfly and worm) are also imported for comparative analysis by the Ensembl Comparative Genomics team. Most annotations are updated every two months, leading to increasing Ensembl versions (such as version 83), however the gene sets are determined less frequently.

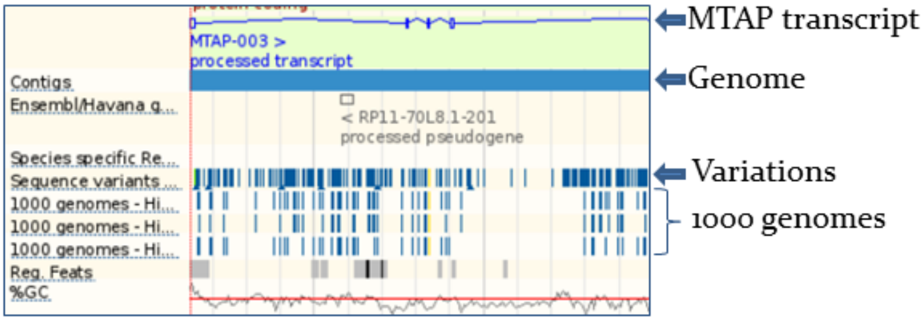


? Click on the cog wheel icon ⚙ to add more data tracks to Ensembl views. Alternatively, you may want to click on the *Configure this page* button ⚙ Configure this page instead at the left hand side.

The vast amount of information associated with genomic sequences demands a way to organise and access that information. This is where genome browsers come in. Ensembl strives to display many layers of genome annotation into a simplified view for the ease of the user. Figure above shows the *Region in detail* page for the *BRCA2* gene in human. The example shows conserved sequence reflecting conservation on a base pair level across dozens of vertebrate species. Conserved regions are displayed as dark blocks that represent local regions of alignment.

In the above figure, proteins alignments from the UniProtKB have been added to the genomic location of the human *BRCA2* gene. Filled yellow blocks show where these UniProtKB proteins align to the genome, and gaps in the alignment are shown as empty yellow blocks. Note that the UniProtKB proteins support most of the exons shown in the Ensembl BRCA2-001 and BRCA2-201 transcripts.

Both **Ensembl** and **Havana** transcripts are displayed as exons (boxes) and introns (connecting lines). Filled boxes show coding sequence and empty boxes reflect **U**n**T**ranslated **R**egions (**UTRs**). This *Region in detail* view is useful for comparing Ensembl gene models with current proteins, mRNAs and ESTs from other databases, such as NCBI RefSeq, ENA, Unigene and UniProtKB. Everything in this view is aligned to the genome.



MTAP transcript

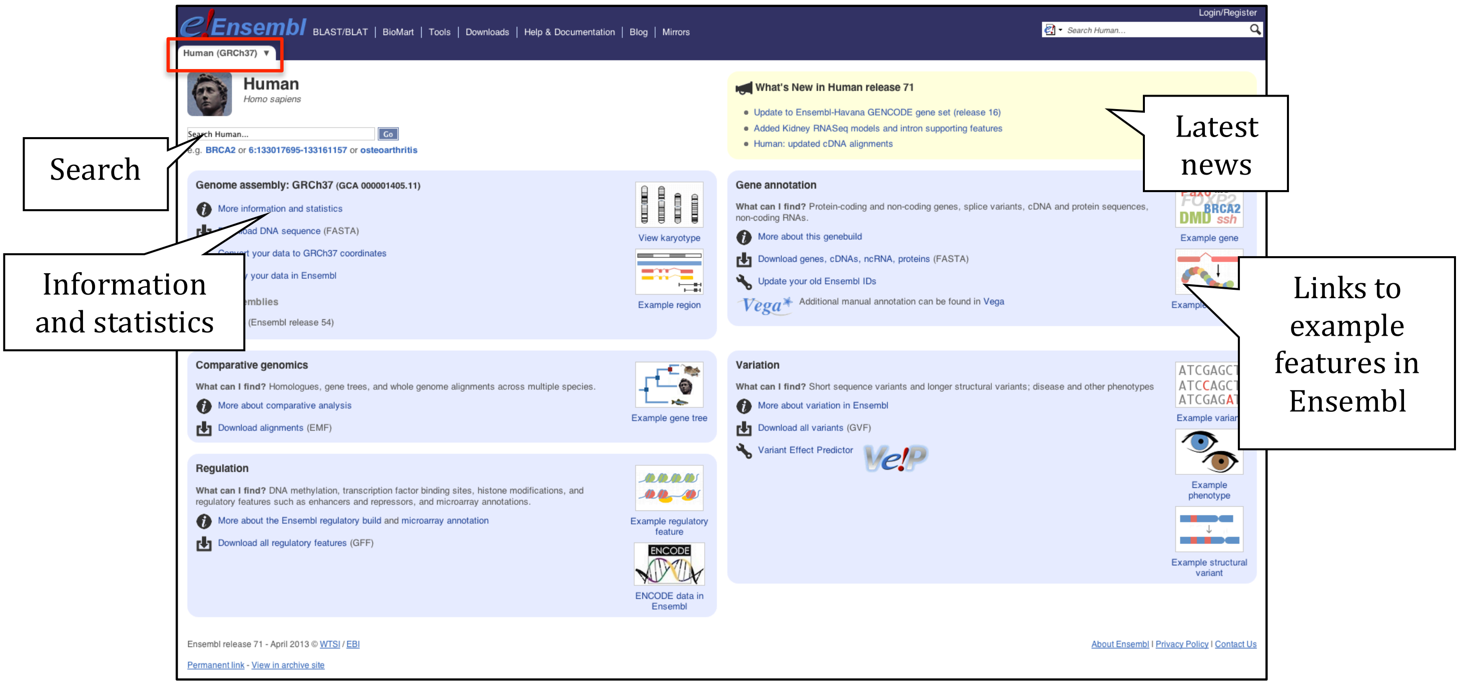
Genome

Variations

1000 genomes

The *Region in detail* view can be configured (using the *Configure this page* button) to show regulatory features, sequence variation, and more! For example, click on dbSNP under the *Variation* menu and turn on the sequence variants (e.g. dbSNP) at the right side of this page. Save and close. Back to the Region in detail view, click on the sequence variation of interest. A pop-up box will show you a few variation properties, such as rs number, alleles, type, and others. Click on the *rs properties* link to take you to an information page for the genetic variation, including links to population frequencies, if available. You can do the same for regulatory features as well.

An index page is provided for each species with information about the source of the genomic sequence assembly, a karyotype (if available), and a link to our archives, which contain previous versions of the Ensembl Browser. The picture below shows the Ensembl homepage for human. Links to the human karyotype, to the previous human assembly and a summary of gene and genome information are found in this index page.


Search
Information and statistics
Latest news
Links to example features in Ensembl

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

Login/Register

Search Human...

Human (GRCh37) ▼

Human
*Homo sapiens*

Search Human... Go
e.g. 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

Genome assembly: GRCh37 (GCA 000001405.11)

More information and statistics
oad DNA sequence (FASTA)
Convert your data to GRCh37 coordinates
y your data in Ensembl
emblies
(Ensembl release 54)

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
Vega Additional manual annotation can be found in Vega

Example gene
Example

Comparative genomics

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

More about comparative analysis
Download alignments (EMF)

Example gene tree

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 VE!P

Example varian
Example phenotype
Example structural variant

Regulation

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

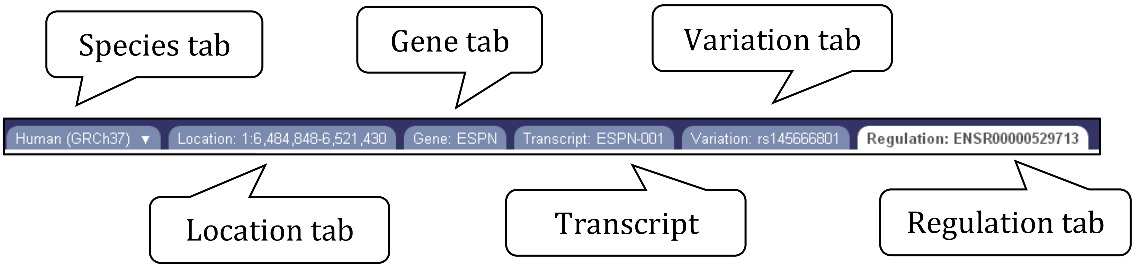
More about the Ensembl regulatory build and microarray annotation
Download all regulatory features (GFF)

Example regulatory feature
ENCODE data in Ensembl

Ensembl release 71 - April 2013 © WTSI / EBI
Permanent link - View in archive site

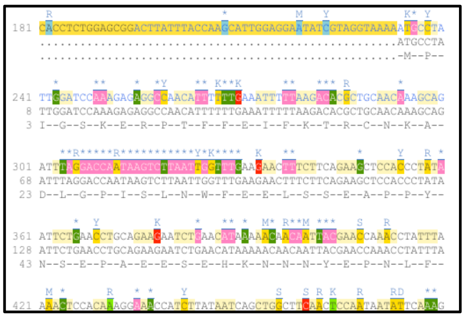
About Ensembl | Privacy Policy | Contact Us

Ensembl uses a tabbed structure for separate pages and views in the browser to display different types of information. The three main entry points in Ensembl are the *Location*, *Gene* and *Transcript* tabs. We have also the *Species*, *Variation* and *Regulation* tabs.

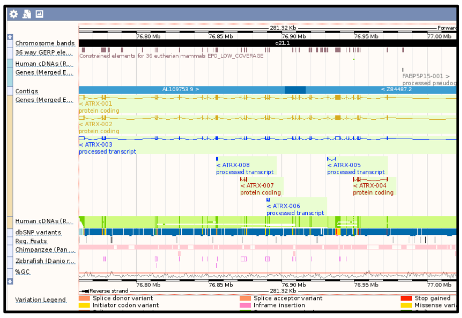


You can for example change the species of interest in the *Species* tab, view a chromosomal region in the *Location* tab (also known as *Region in detail*), visualise gene trees in the *Gene* tab, browse the cDNA sequence alongside the protein translation in the *Transcript* tab, look for genotype information in the *Variation* tab and get regulatory information in the *Regulation* tab. You can also perform a similarity search against any species in Ensembl by using our BLAST/BLAT tools.

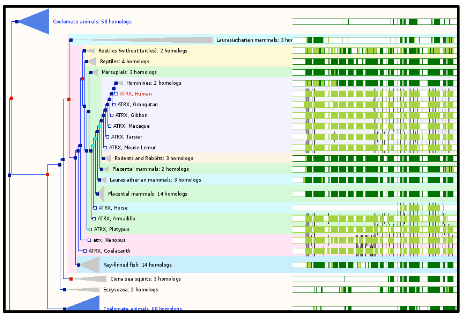
Transcript with variations



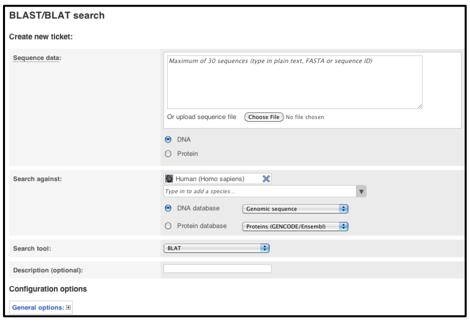
Genes and conserved regions



Homologues in Gene Trees



BLAST and BLAT searches



**BLAST/BLAT search**

**Create new ticket:**

Sequence data: *Maximum of 30 sequences (type in plain text, FASTA or sequence ID)*

Or upload sequence file [Choose File] No file chosen

● DNA
○ Protein

Search against: Human (Homo sapiens) ✕
*Type in to add a species...*

● DNA database: Genomic sequence
○ Protein database: Proteins (GENCODE/Ensembl)

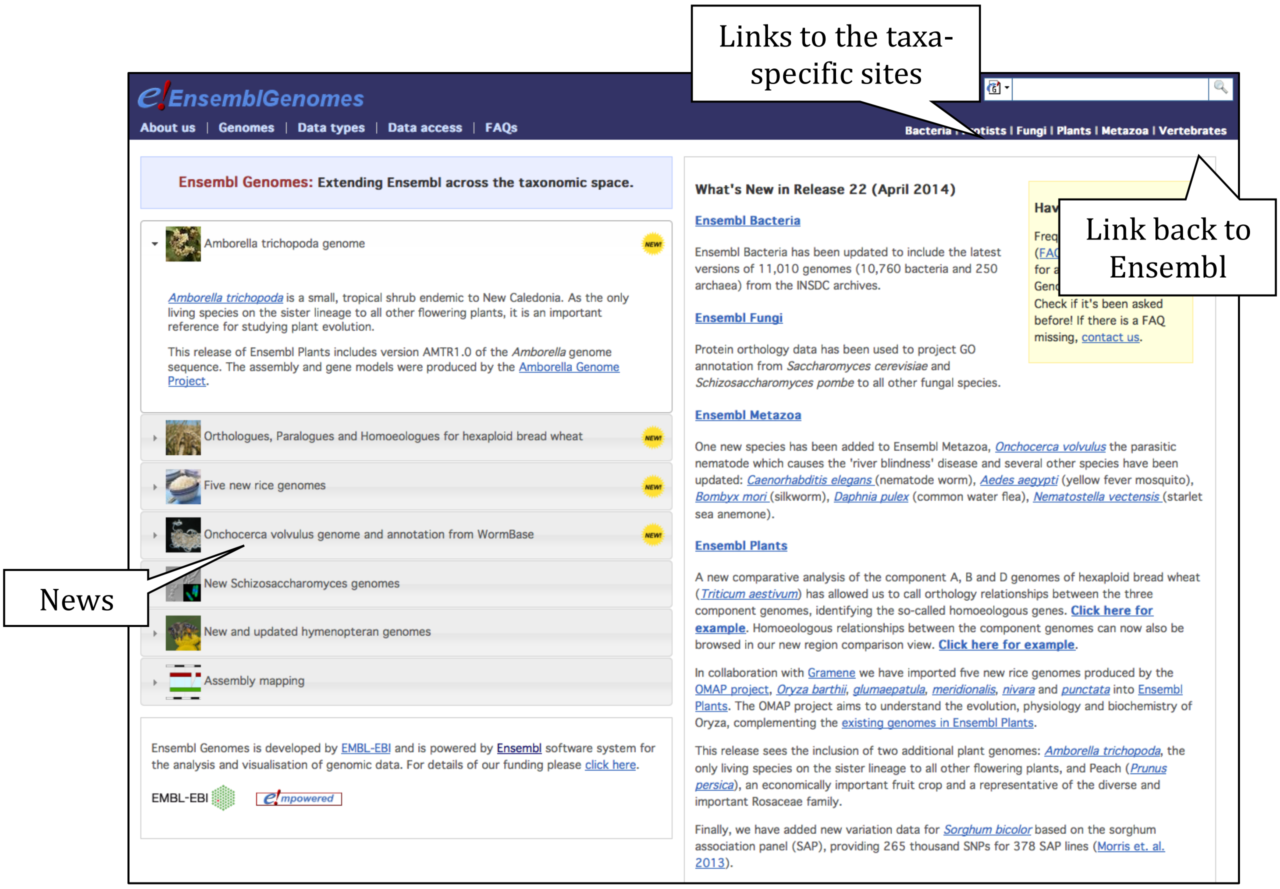
Search tool: BLAT

Description (optional):

**Configuration options**

General options: ⊞

Let’s now take a look at the Ensembl Genomes browser: [www.ensemblgenomes.org](http://www.ensemblgenomes.org)



Links to the taxa-specific sites

Link back to Ensembl

News

EnsemblGenomes
About us | Genomes | Data types | Data access | FAQs
Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates

**Ensembl Genomes: Extending Ensembl across the taxonomic space.**

Amborella trichopoda genome NEW!

*Amborella trichopoda* is a small, tropical shrub endemic to New Caledonia. As the only living species on the sister lineage to all other flowering plants, it is an important reference for studying plant evolution.

This release of Ensembl Plants includes version AMTR1.0 of the *Amborella* genome sequence. The assembly and gene models were produced by the Amborella Genome Project.

- Orthologues, Paralogues and Homoeologues for hexaploid bread wheat NEW!
- Five new rice genomes NEW!
- Onchocerca volvulus genome and annotation from WormBase NEW!
- New Schizosaccharomyces genomes
- New and updated hymenopteran genomes
- Assembly mapping

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.

EMBL-EBI e!mpowered

**What's New in Release 22 (April 2014)**

**Ensembl Bacteria**

Ensembl Bacteria has been updated to include the latest versions of 11,010 genomes (10,760 bacteria and 250 archaea) from the INSDC archives.

**Ensembl Fungi**

Protein orthology data has been used to project GO annotation from *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* to all other fungal species.

**Ensembl Metazoa**

One new species has been added to Ensembl Metazoa, *Onchocerca volvulus* the parasitic nematode which causes the 'river blindness' disease and several other species have been updated: *Caenorhabditis elegans* (nematode worm), *Aedes aegypti* (yellow fever mosquito), *Bombyx mori* (silkworm), *Daphnia pulex* (common water flea), *Nematostella vectensis* (starlet sea anemone).

**Ensembl Plants**

A new comparative analysis of the component A, B and D genomes of hexaploid bread wheat (*Triticum aestivum*) has allowed us to call orthology relationships between the three component genomes, identifying the so-called homoeologous genes. **Click here for example**. Homoeologous relationships between the component genomes can now also be browsed in our new region comparison view. **Click here for example**.

In collaboration with Gramene we have imported five new rice genomes produced by the OMAP project, *Oryza barthii*, *glumaepatula*, *meridionalis*, *nivara* and *punctata* into Ensembl Plants. The OMAP project aims to understand the evolution, physiology and biochemistry of Oryza, complementing the existing genomes in Ensembl Plants.

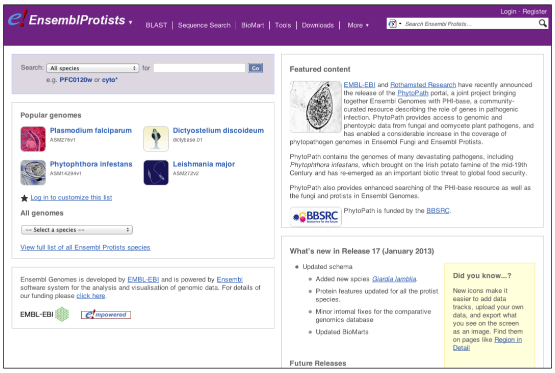
This release sees the inclusion of two additional plant genomes: *Amborella trichopoda*, the only living species on the sister lineage to all other flowering plants, and Peach (*Prunus persica*), an economically important fruit crop and a representative of the diverse and important Rosaceae family.

Finally, we have added new variation data for *Sorghum bicolor* based on the sorghum association panel (SAP), providing 265 thousand SNPs for 378 SAP lines (Morris et. al. 2013).

Hav
Freq
(FAQ
for a
Gen
Check if it's been asked before! If there is a FAQ missing, contact us.

Click on the different taxa to see their homepages. Each of them is colour-coded as it follows:

**Protists** **Fungi** **Metazoa** **Plants** **Bacteria**



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

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.

EMBL-EBI e!mpowered

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.

BBSRC PhytoPath is funded by the BBSRC.

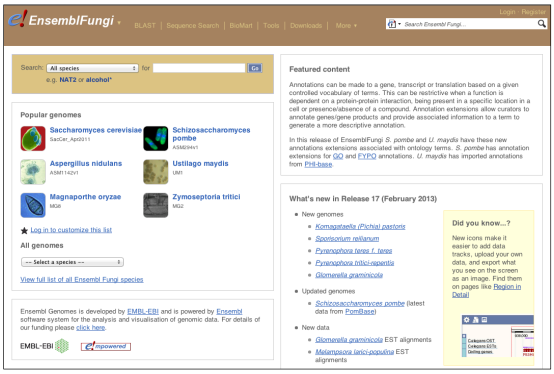
What's new in Release 17 (January 2013)

- 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

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

Future Releases



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

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.

EMBL-EBI e!mpowered

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.

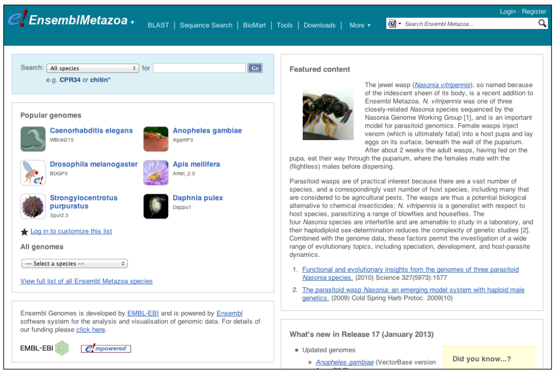
In this release of EnsemblFungi *S. pombe* and *U. maydis* have these new annotations extensions associated with ontology terms. *S. pombe* has annotation extensions for GO and FYPO annotations. *U. maydis* has imported annotations from PHI-base.

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

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



Search: All species for
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

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.

EMBL-EBI e!mpowered

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.

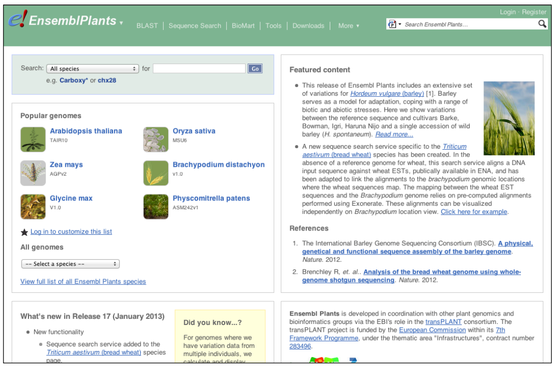
Parasitoid wasps are of practical interest because there are a vast number of species, and a correspondingly vast number of host species, including many that are considered to be agricultural pests. The wasps are thus a potential biological alternative to chemical insecticides; *N. vitripennis* is a 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...?



Search: All species for
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

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

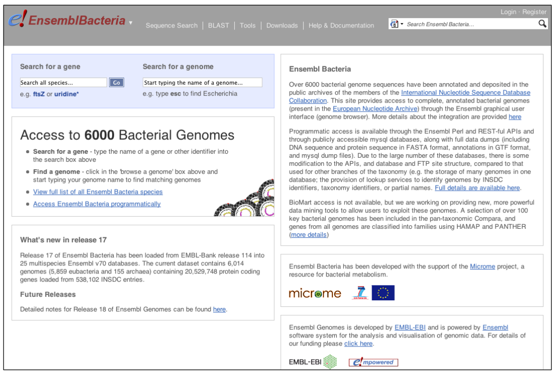
Featured content

- 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.


EnsemblBacteria
Sequence Search | BLAST | Tools | Downloads | Help & Documentation
Login · Register
Search Ensembl Bacteria…
Search all species…
Go
Start typing the name of a genome…

Search for a gene
e.g. ftsZ or uridine*

Search for 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)

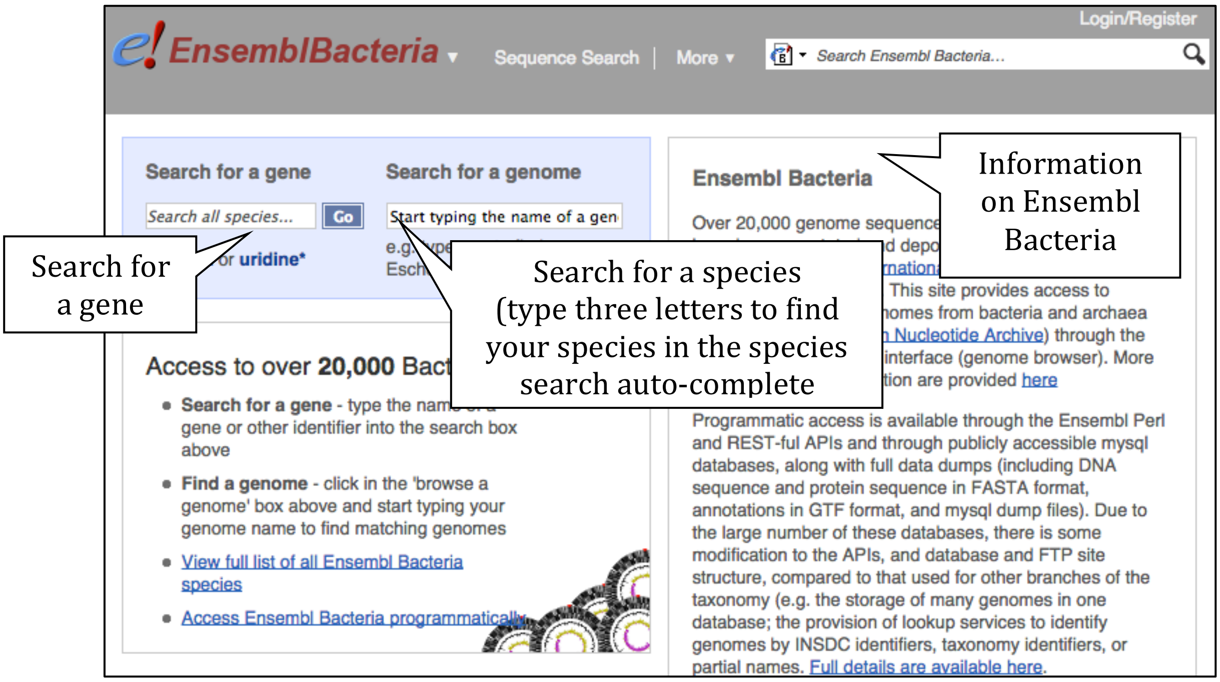
Ensembl Bacteria has been developed with the support of the Microme project, a resource for bacterial metabolism.

microme

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.

EMBL-EBI e!mpowered

You can navigate most of the taxa in the same way as you do with Ensembl, but since Ensembl Bacteria has a large number of genomes (>20,000), it needs slightly different methods for browser navigation. Let's look at this in more detail.



Access to over 20,000 Bact

- **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

**Ensembl Bacteria**

Over 20,000 genome sequence
d depo
rnationa
This site provides access to
nomes from bacteria and archaea
n Nucleotide Archive) through the
interface (genome browser). More
tion 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.

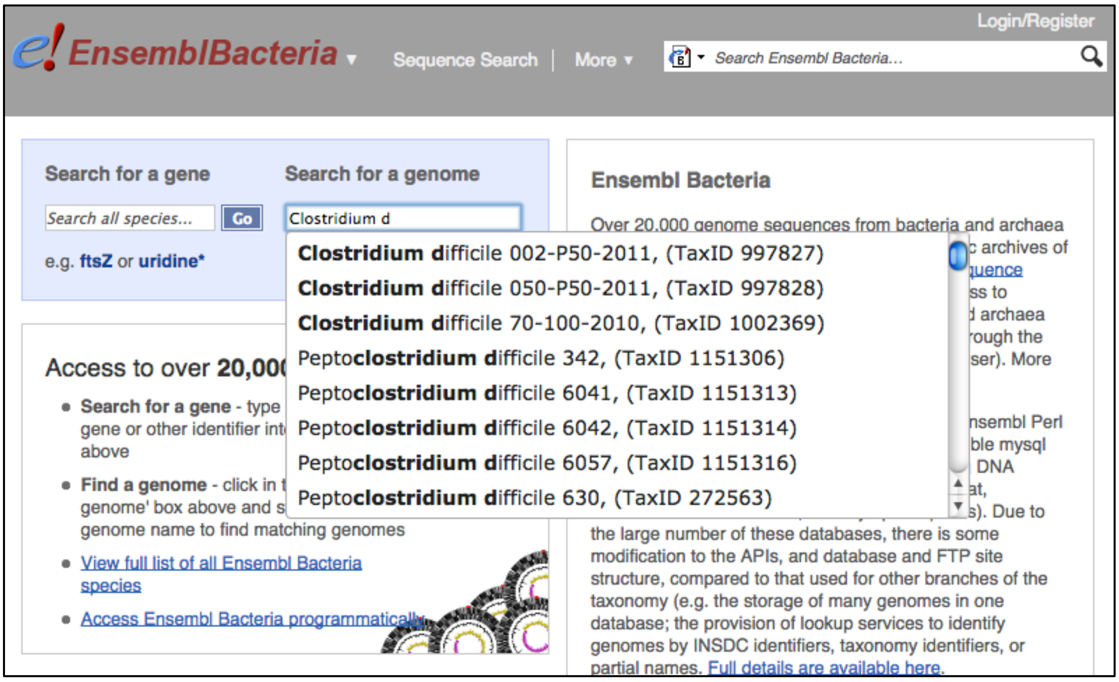
Search for a gene

Search for a species (type three letters to find your species in the species search auto-complete

Information on Ensembl Bacteria

There’s no full species list for bacteria as it would be hard to navigate with the number of species available in Bacteria. To find a species, start to type the species name (or any three letter code) into the species search box. A drop down list will appear with possible species

For example, to find substrains of *Clostridium difficile* type in Clostridium d to find the possible options.



**Clostridium d**ifficile 002-P50-2011, (TaxID 997827)
**Clostridium d**ifficile 050-P50-2011, (TaxID 997828)
**Clostridium d**ifficile 70-100-2010, (TaxID 1002369)
Pepto**clostridium d**ifficile 342, (TaxID 1151306)
Pepto**clostridium d**ifficile 6041, (TaxID 1151313)
Pepto**clostridium d**ifficile 6042, (TaxID 1151314)
Pepto**clostridium d**ifficile 6057, (TaxID 1151316)
Pepto**clostridium d**ifficile 630, (TaxID 272563)

Access to over 20,00

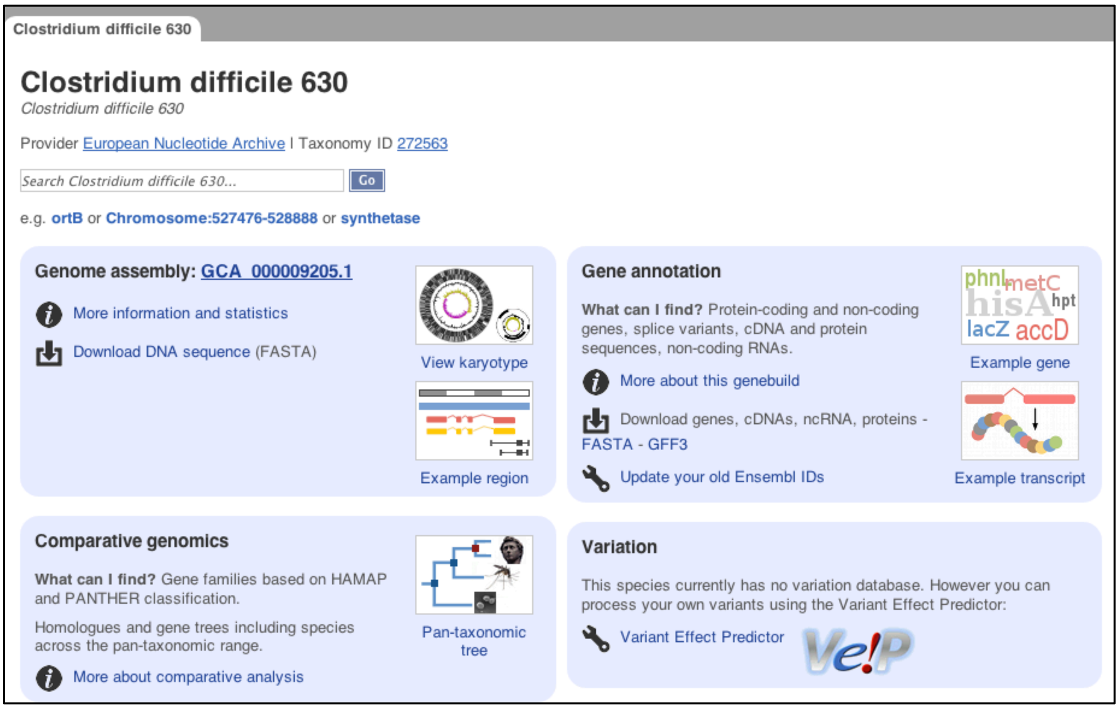
- **Search for a gene** - type gene or other identifier int above
- **Find a genome** - click in t genome' box above and s genome name to find matching genomes
- View full list of all Ensembl Bacteria species
- Access Ensembl Bacteria programmatically

**Ensembl Bacteria**

Over 20,000 genome sequences from bacteria and archaea
c archives of
quence
ss to
d archaea
rough the
ser). More

nsembl Perl
ble mysql
DNA
at,
s). 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.

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



Clostridium difficile 630

## Clostridium difficile 630

*Clostridium difficile 630*

Provider European Nucleotide Archive | Taxonomy ID 272563

Search Clostridium difficile 630... Go

e.g. **ortB** or **Chromosome:527476-528888** 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 - GFF3

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

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

## Retrieving Data from Ensembl

BioMart is a web-interface that can extract information from the Ensembl databases and present the user with a table of information without the need for programming. It can be used to output sequences or tables of genes along with gene positions (chromosome and base pair locations), single nucleotide polymorphisms (SNPs), homologues, and other annotation in HTML, text, or Microsoft Excel format. BioMart can also translate one type of ID to another, identify genes associated with **InterPro** domains or gene ontology (**GO**) terms, export gene expression data and much more.

Ensembl uses MySQL relational databases to store its information. A comprehensive set of Application Program Interfaces (APIs) serve as a middle-layer between underlying database schemes and more specific application programs. The API aims to encapsulate the database layout by providing efficient high-level access to data tables and isolate applications from data layout changes.

## Synopsis: what can I do with Ensembl?

- View genes with different annotations along the chromosome;
- View alternative transcripts (i.e. splice variants) for a given gene;
- Explore homologues and phylogenetic trees across more than 70 chordate 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) and human populations;
- View positions and sequence of mRNAs and proteins that align with an Ensembl genes;
- Display your own data on the Ensembl browser;
- 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;
- Retrieve our data using the Perl or REST APIs.

## 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!

❖ Comments/questions http://www.ensembl.org/Help/Contact

❖ Read our Ensembl blog

❖ Follow us on Twitter @ensembl @ensemblgenomes

❖ Sign up to our mailing lists

## Further reading

Yates, A. *et al.*
**Ensembl 2016**
Nucleic Acids Res (Database Issue)
http://nar.oxfordjournals.org/content/early/2015/12/19/nar.gkv1157.full

Kersey, PJ. *et al*
Nucleic Acids Res (Database Issue)

**For a complete list of publications, see below**
http://www.ensembl.org/info/about/publications.html
http://ensemblgenomes.org/info/publications

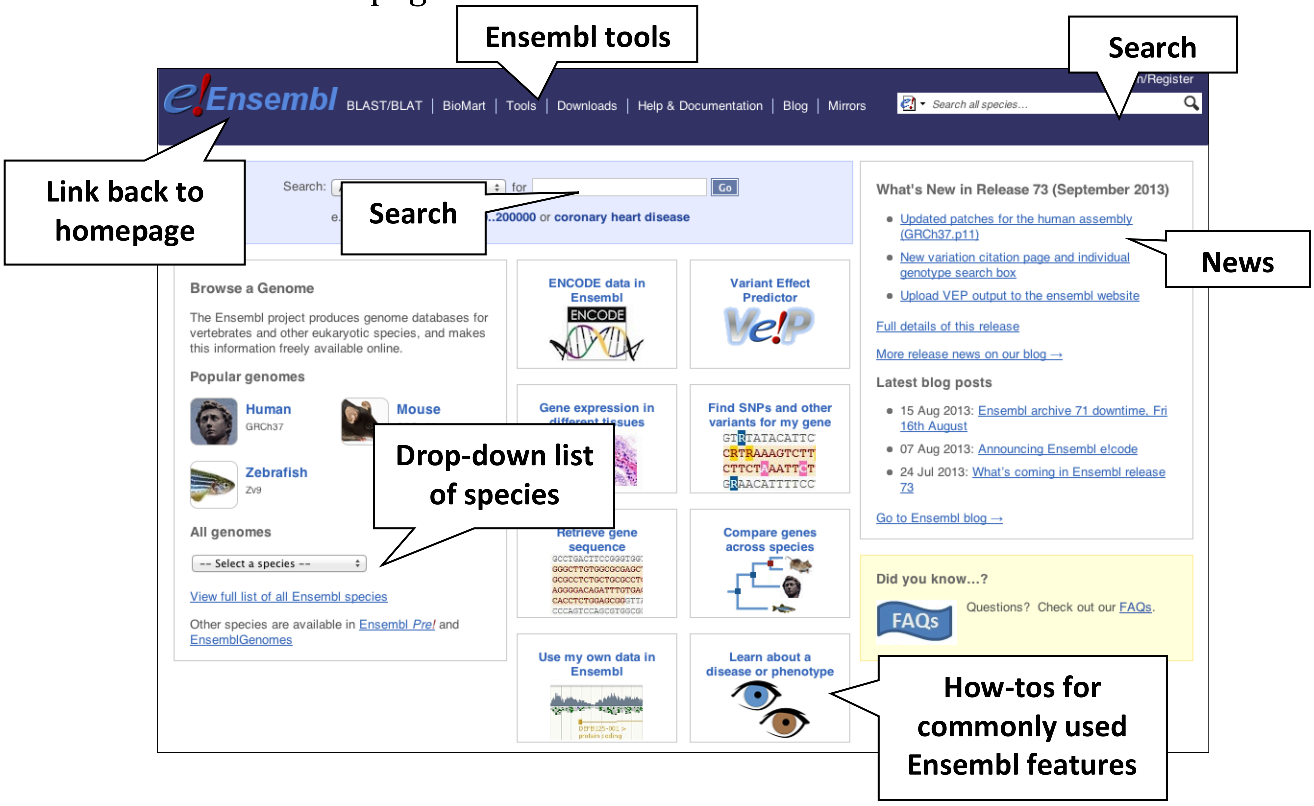
# BROWSER WALKTHROUGH

We will guide you through the website using 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 (www.uniprot.org/uniprot/B1AK53).

The following points will be addressed during the walkthrough:

- **The Location tab and genomic location related links:**
  - How do I zoom out to change the gene focus?
  - How to add a track (e.g. protein alignments, variation data)?
- **The Gene tab and gene related links:**
  - Can I view the genomic sequence of my gene with its variations?
  - How to find orthologues and paralogues?
- **The Transcript tab and transcript related links:**
  - What is the protein sequence?
  - What proteins and mRNAs are found in other databases?
- **Exporting a sequence and running BLAT**

Go to www.ensembl.org and click on the human icon to open the human home page.


Ensembl tools
Search
Link back to homepage
Search
News
Drop-down list of species
How-tos for commonly used Ensembl features
e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
n/Register
Search all species…
Search:
for
Go
ENCODE data in Ensembl
ENCODE
Variant Effect Predictor
Ve!P
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
Compare genes across species
Use my own data in Ensembl
Learn about a disease or phenotype
Human
GRCh37
Mouse
Zebrafish
Zv9
-- Select a species --
FAQs

e. …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

All genomes

View full list of all Ensembl species

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

What's New in Release 73 (September 2013)

- Updated patches for the human assembly (GRCh37.p11)
- New variation citation page and individual genotype search box
- Upload VEP output to the ensembl website

Full details of this release

More release news on our blog →

Latest blog posts

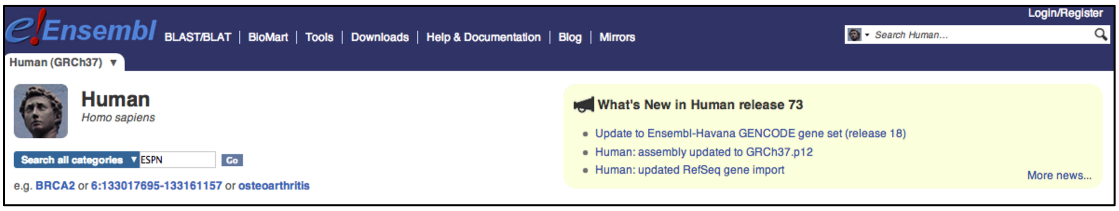
- 15 Aug 2013: Ensembl archive 71 downtime, Fri 16th August
- 07 Aug 2013: Announcing Ensembl e!code
- 24 Jul 2013: What's coming in Ensembl release 73

Go to Ensembl blog →

Did you know…?

Questions? Check out our FAQs.

Type *ESPN* into the search bar and click the *Go* button.



ESPN

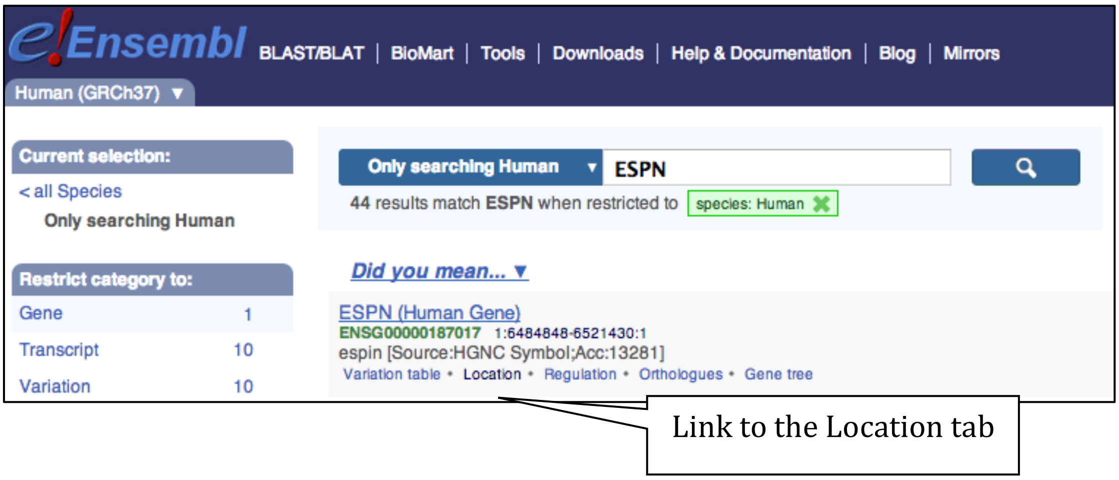
e.g. BRCA2 or 6:133017695-133161157 or osteoarthritis

What's New in Human release 73

- Update to Ensembl-Havana GENCODE gene set (release 18)
- Human: assembly updated to GRCh37.p12
- Human: updated RefSeq gene import

More news...

One gene matches the query in human. Links to the *Gene* tab, *Variation* Table, *Location* tab, among others are provided.



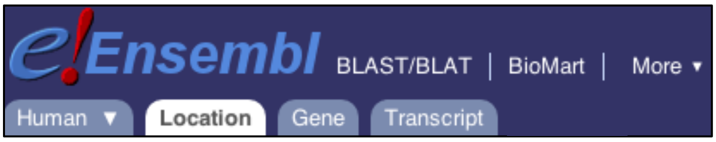
ESPN

44 results match ESPN when restricted to species: Human ✗

ESPN (Human Gene)
ENSG00000187017 1:6484848-6521430:1
espin [Source:HGNC Symbol;Acc:13281]

Link to the Location tab

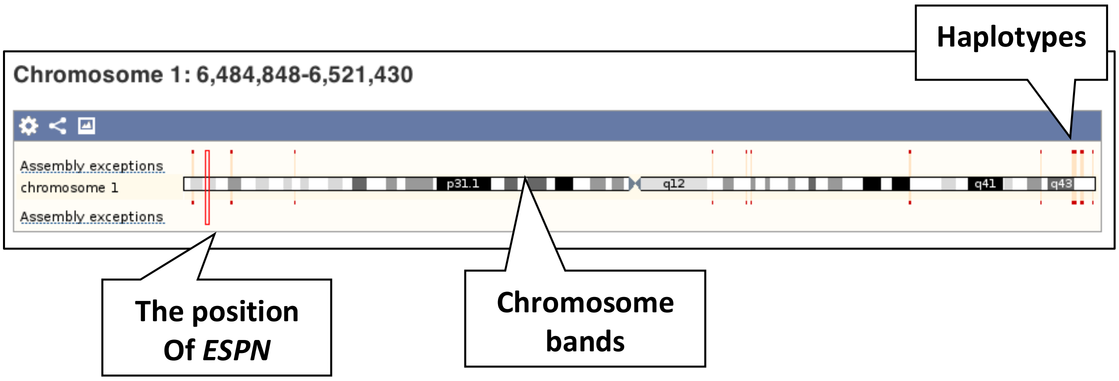
Let's view the genomic region in which this gene is located by clicking on the Location link. The Location tab should open.



The Location tab in Ensembl is also known as *Region in detail* view. There is a help video on this page at http://youtu.be/tTKEvgPUq94.

The *Location* tab contains three images.

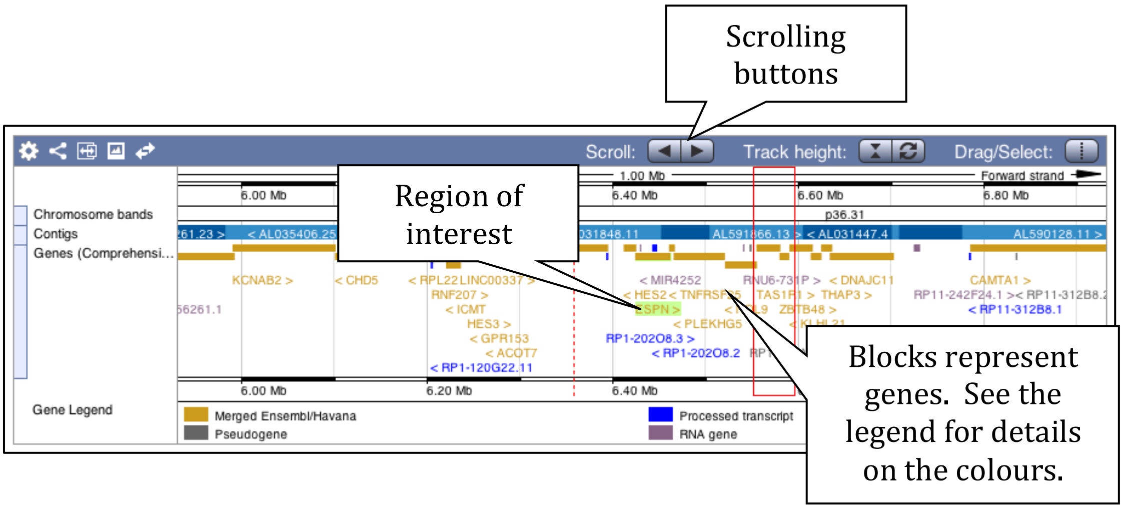
The first image shows an overview of chromosome where the human *ESPN* gene has been mapped. In this image, you can see G banding pattern of the chromosome as well as the regions that correspond to the patches and haplotypes of the human chromosome 1.



More details on haplotypes (e.g. MHC) can be found in the link below:

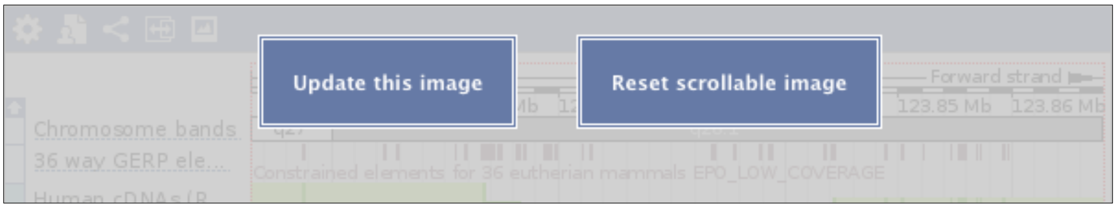
http://www.ensembl.org/Multi/Help/Movie?db=core;id=372

The second image shows a 1Mb region around the *ESPN* gene. This view allows scrolling back and forth along the chromosome. This view can also be configurable/customised.



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.

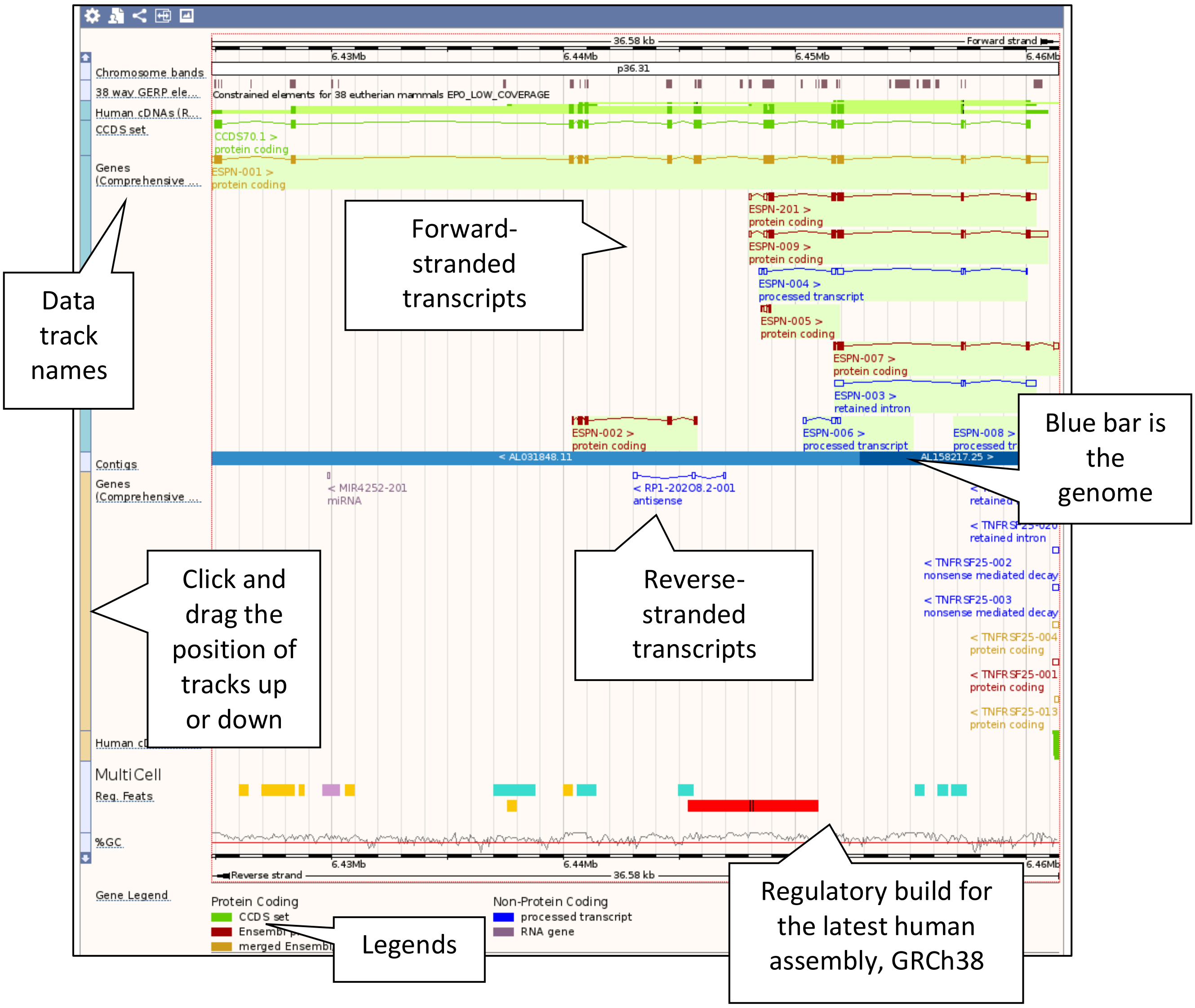
Scroll along the chromosome by clicking and dragging within the image. As you do this, you'll see the image below grey out and two blue buttons appear. Clicking on *Update this image* will jump the lower image to the region central to the scrollable image. If you want to go back to where you started,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



The third and final image is a detailed and highly configurable view of the region.

You can edit what you see on this page by clicking on the blue Configure this page menu at the left or click on the *cog wheel* in the bottom image.

This will open a menu that allows you to change the image and looks like the image below:



Data track names

Forward-stranded transcripts

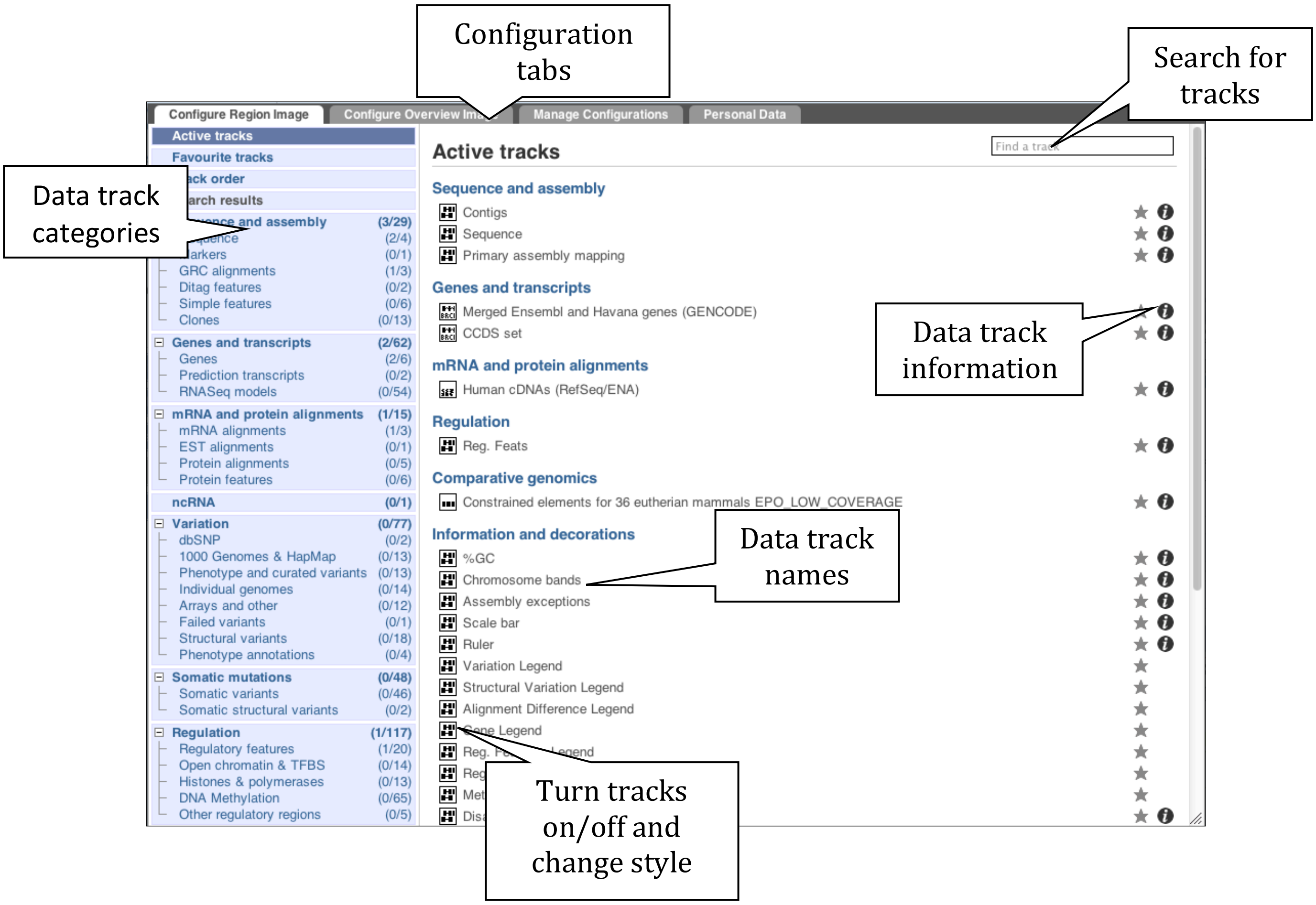
Blue bar is the genome

Click and drag the position of tracks up or down

Reverse-stranded transcripts

Regulatory build for the latest human assembly, GRCh38

Legends



Configure Region Image | Configure Overview Ima | Manage Configurations | Personal Data

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

Find a track

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 egend
Reg
Met
Dis

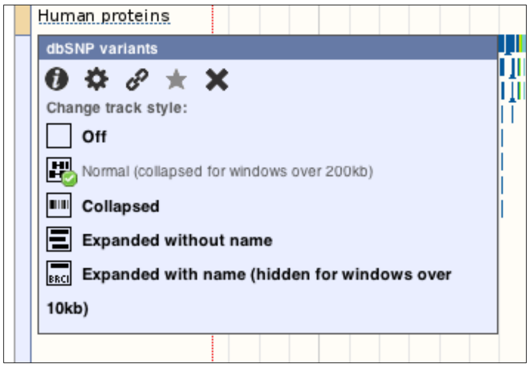
Let's add some tracks to this image, such as:

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

The data tracks can be displayed on different styles. For more details, have a look at our FAQ: http://www.ensembl.org/Help/Faq?id=335.

Now click on the tick in the top left hand to SAVE and close the menu with the newly added configuration. Alternatively, click anywhere outside of the menu.

We can also change the way the tracks appear by hovering over the track name, then the cog wheel to open a menu.



Human proteins

dbSNP variants

Change track style:

□ Off

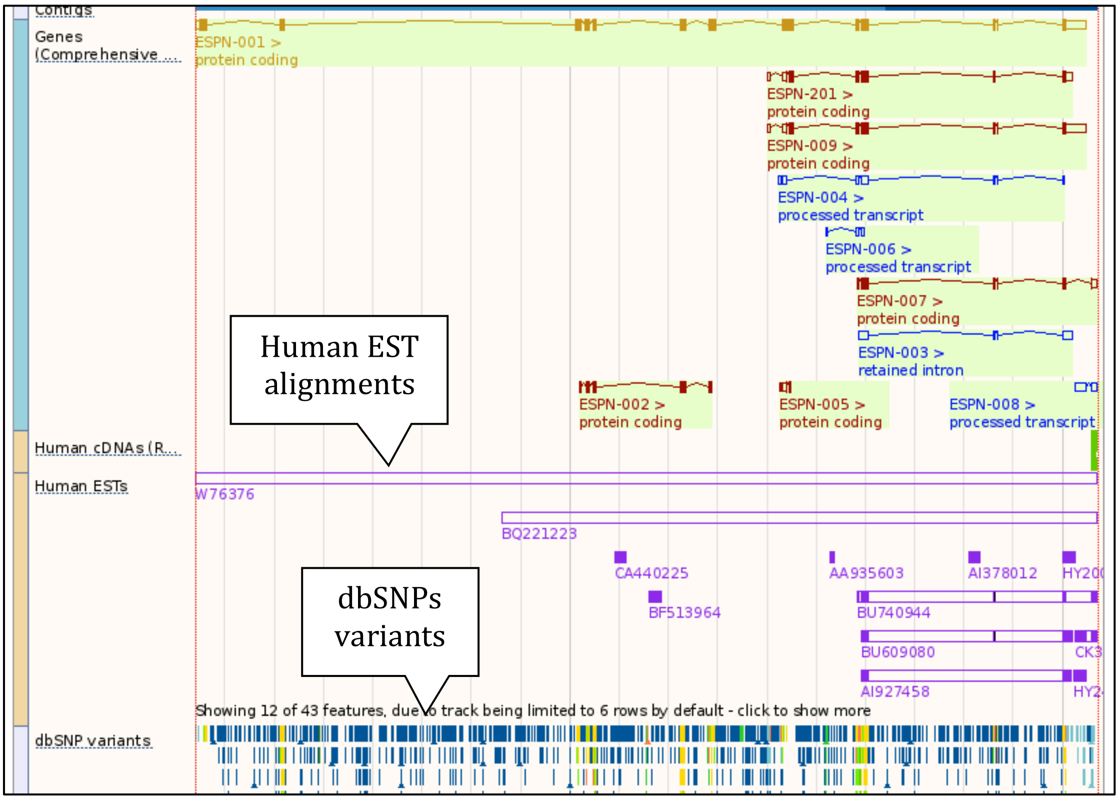
Normal (collapsed for windows over 200kb)

Collapsed

Expanded without name

Expanded with name (hidden for windows over 10kb)

Have a look at the changes in the Location tab (sometimes referred to as *Region in detail* view). Click and drag tracks to reorder them, if it helps with comparing the data.


Contigs
Genes (Comprehensive ...
ESPN-001 > protein coding
ESPN-201 > protein coding
ESPN-009 > protein coding
ESPN-004 > processed transcript
ESPN-006 > processed transcript
ESPN-007 > protein coding
ESPN-003 > retained intron
ESPN-002 > protein coding
ESPN-005 > protein coding
ESPN-008 > processed transcript
Human cDNAs (R...
Human ESTs
W76376
BQ221223
CA440225
AA935603
AI378012
HY20
BF513964
BU740944
BU609080
CK3
AI927458
HY2
dbSNP variants

Human EST alignments

dbSNPs variants

Showing 12 of 43 features, due to track being limited to 6 rows by default - click to show more

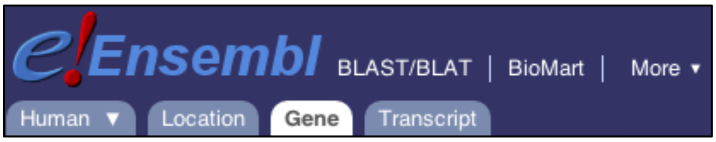
Now that you've got the view how you want it, you might like to show it to a colleague or collaborator. Click on the *Share this page* button to generate a link so that you can email it to someone.

They will see exactly the same view as you, including all the tracks you have added and moved up and/or down. These links contain the

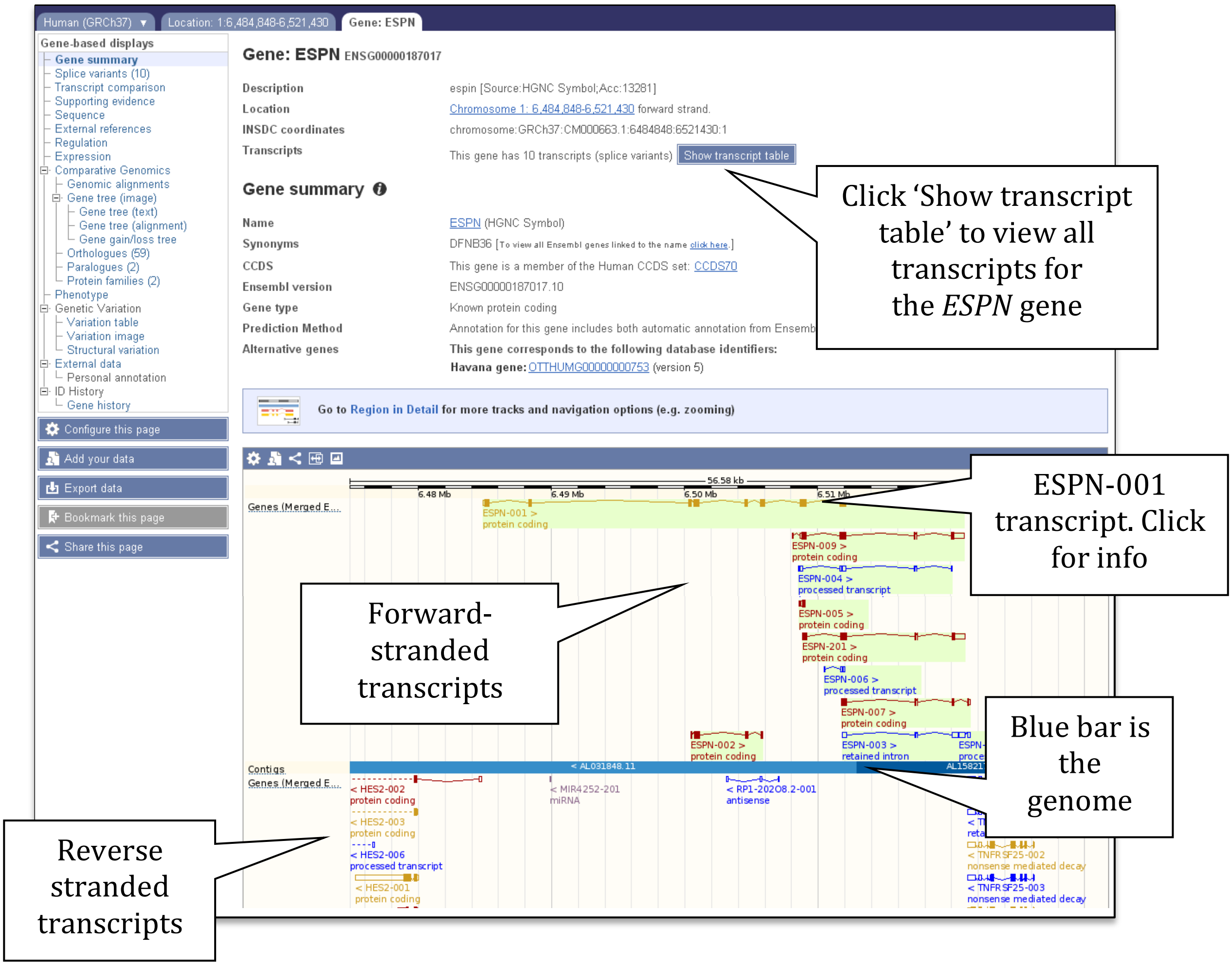
Ensembl release number, so if a new release or even assembly comes out, your link will just take you to the archive site for the release it was made on.

To return to the default view, click on *Configure this page* and select Reset configuration at the bottom of the menu. You can also reset the track order.

Let's now explore the Gene tab.



We will walk you through some of the links in the left hand navigation column. Note that the left hand side menu in the Gene tab differs from the one we saw previously in the Location tab.



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 Ensemb |
| Alternative genes | **This gene corresponds to the following database identifiers:**<br>**Havana gene:** OTTHUMG00000000753 (version 5) |

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

Click 'Show transcript table' to view all transcripts for the *ESPN* gene

ESPN-001 transcript. Click for info

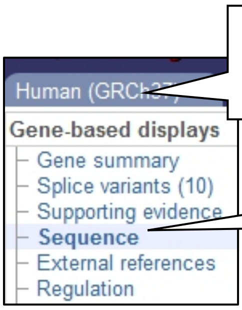
Forward-stranded transcripts

Blue bar is the genome

Reverse stranded transcripts

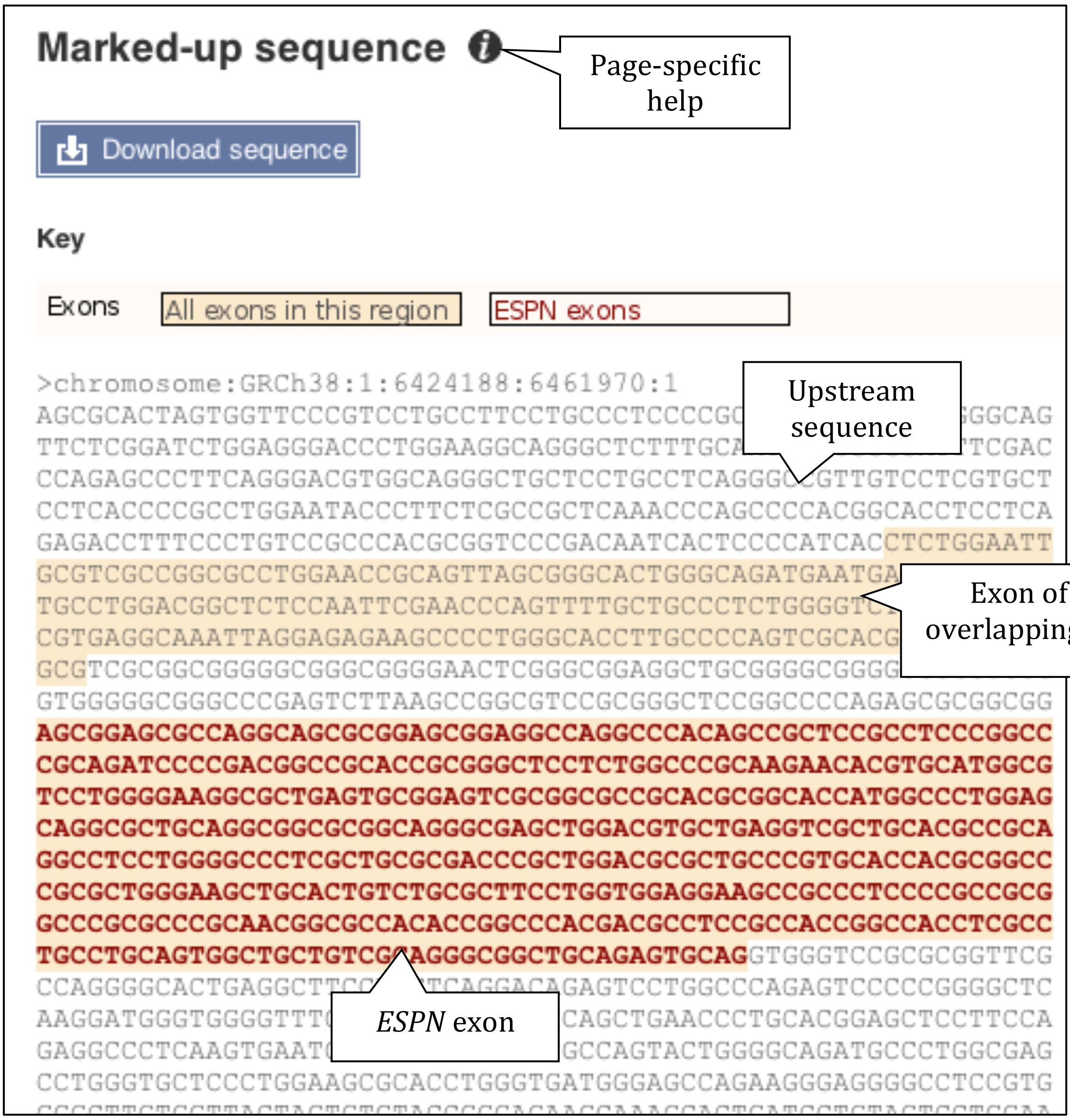
How can we view the genomic sequence of the *ESPN* gene?

Click Sequence at the left of the page.



Most recent human genome assembly, GRCh38

Click Sequence



```
>chromosome:GRCh38:1:6424188:6461970:1
AGCGCACTAGTGGTTCCCGTCCTGCCTTCCTGCCCTCCCCGC          GGCAG
TTCTCGGATCTGGAGGGACCCTGGAAGGCAGGGCTCTTTGCA          TCGAC
CCAGAGCCCTTCAGGGACGTGGCAGGGCTGCTCCTGCCTCAGGGC GTTGTCCTCGTGCT
CCTCACCCCGCCTGGAATACCCTTCTCGCCGCTCAAACCCAGCCCCACGGCACCTCCTCA
GAGACCTTTCCCTGTCCGCCCACGCGGTCCCGACAATCACTCCCCATCACCTCTGGAATT
GCGTCGCCGGCGCCTGGAACCGCAGTTAGCGGGCACTGGGCAGATGAATGA
TGCCTGGACGGCTCTCCAATTCGAACCCAGTTTTGCTGCCCTCTGGGGTC
CGTGAGGCAAATTAGGAGAGAAGCCCCTGGGCACCTTGCCCCAGTCGCACG
GCGTCGCGGCGGGGGCGGGCGGGGAACTCGGGCGGAGGCTGCGGGGCGGGG
GTGGGGGCGGGCCCGAGTCTTAAGCCGGCGTCCGCGGGCTCCGGCCCCAGAGCGCGGCGG
AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCC
CGCAGATCCCCGACGGCCGCACCGCGGGCTCCTCTGGCCCGCAAGAACACGTGCATGGCG
TCCTGGGGAAGGCGCTGAGTGCGGAGTCGCGGCGCCGCACGCGGCACCATGGCCCTGGAG
CAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCTGGACGTGCTGAGGTCGCTGCACGCCGCA
GGCCTCCTGGGGCCCTCGCTGCGCGACCCGCTGGACGCGCTGCCCGTGCACCACGCGGCC
CGCGCTGGGAAGCTGCACTGTCTGCGCTTCCTGGTGGAGGAAGCCGCCCTCCCCGCCGCG
GCCCGCGCCCGCAACGGCGCCACACCGGCCCACGACGCCTCCGCCACCGGCCACCTCGCC
TGCCTGCAGTGGCTGCTGTCGG AGGGCGGCTGCAGAGTGCAGGTGGGTCCGCGCGGTTCG
CCAGGGGCACTGAGGCTTCC   TCAGGACAGAGTCCTGGCCCAGAGTCCCCCGGGGCTC
AAGGATGGGTGGGGTTTC               CAGCTGAACCCTGCACGGAGCTCCTTCCA
GAGGCCCTCAAGTGAATC               GCCAGTACTGGGGCAGATGCCCTGGCGAG
CCTGGGTGCTCCCTGGAAGCGCACCTGGGTGATGGGAGCCAGAAGGGAGGGGCCTCCGTG
```

Page-specific help

Upstream sequence

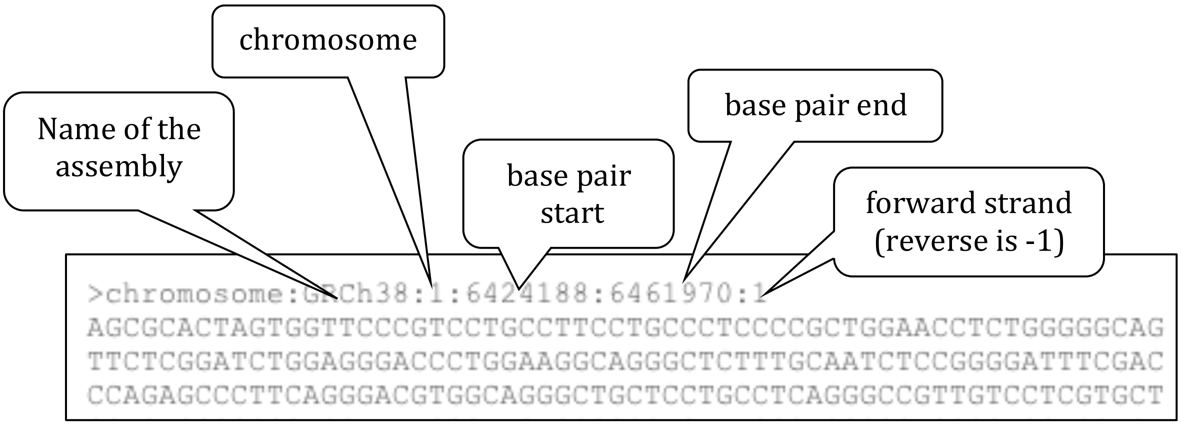
Exon of an overlapping gene

*ESPN* exon

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

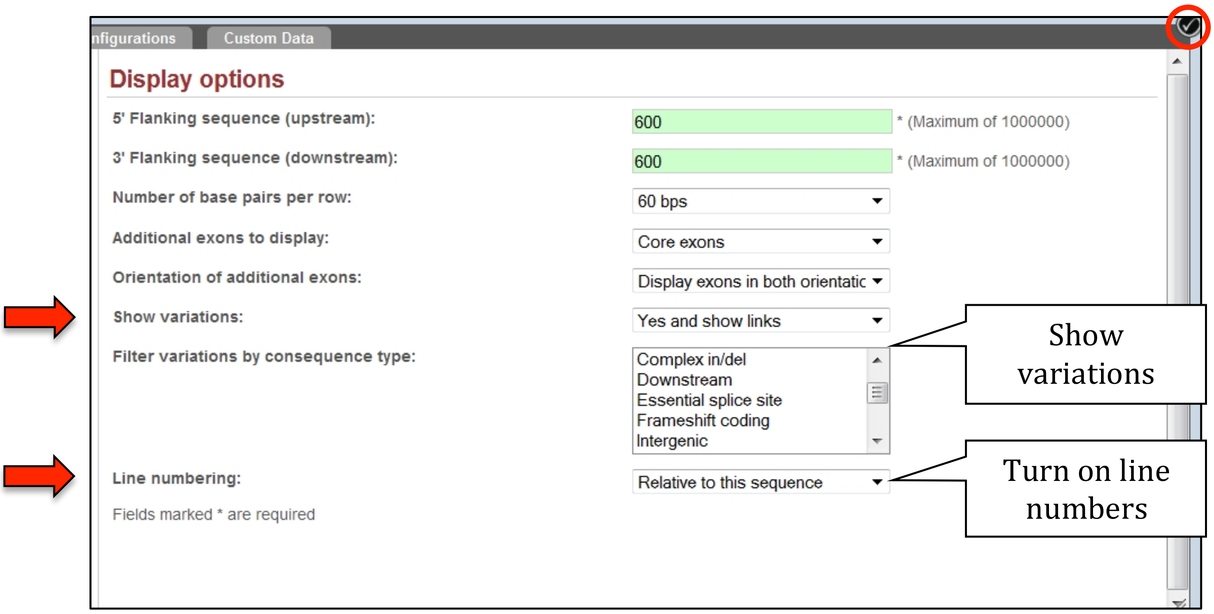
The help pages also provide links to Frequently Asked Questions, a Glossary, Video Tutorials, and a form to Contact HelpDesk.

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



```
>chromosome:GRCh38:1:6424188:6461970:1
AGCGCACTAGTGGTTCCCGTCCTGCCTTCCTGCCCTCCCCGCTGGAACCTCTGGGGGCAG
TTCTCGGATCTGGAGGGACCCTGGAAGGCAGGGCTCTTTGCAATCTCCGGGGATTTCGAC
CCAGAGCCCTTCAGGGACGTGGCAGGGCTGCTCCTGCCTCAGGGCCGTTGTCCTCGTGCT
```

Exons are highlighted within the genomic sequence by default. Variations and other features 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 orientatic

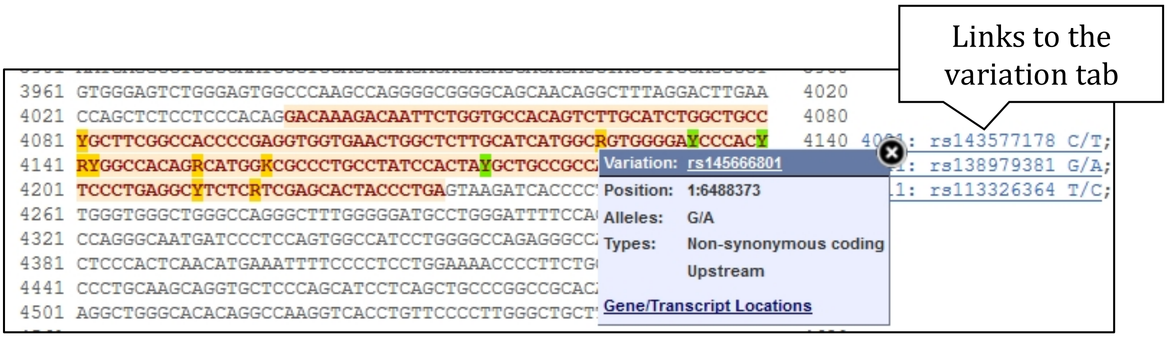
Show variations: Yes and show links

Filter variations by consequence type: Complex in/del, Downstream, Essential splice site, Frameshift coding, Intergenic

Line numbering: Relative to this sequence

Fields marked * are required

Once you have selected the new display options (in this example, Show variations and Line numbering) click at the top right of the image (circled in red above).

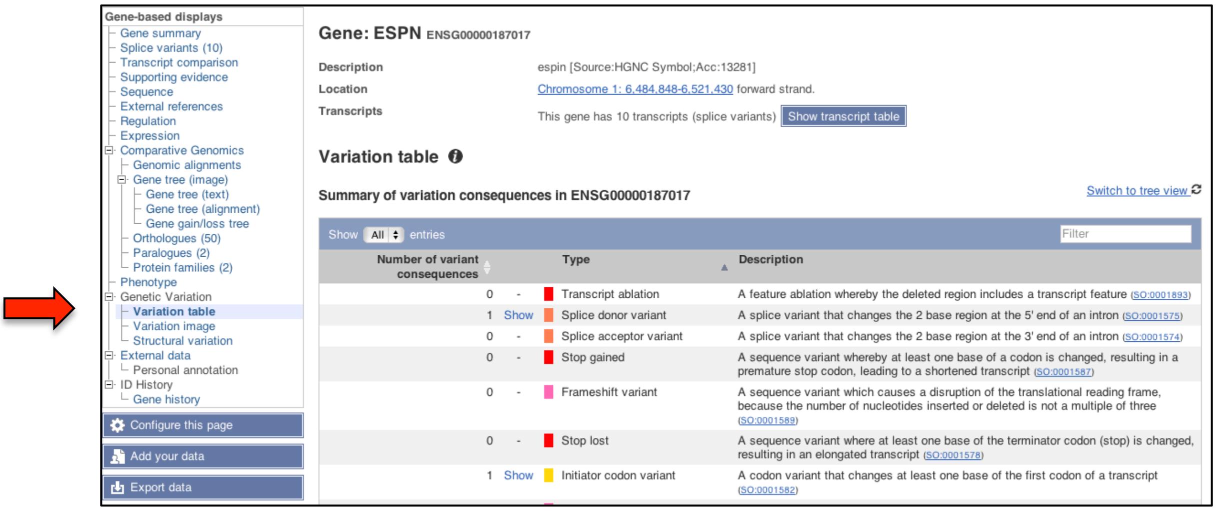


Links to the variation tab

```
3961 GTGGGAGTCTGGGAGTGGCCCAAGCCAGGGGCGGGGCAGCAACAGGCTTTAGGACTTGAA  4020
4021 CCAGCTCTCCTCCCACAGGACAAAGACAATTCTGGTGCCACAGTCTTGCATCTGGCTGCC  4080
4081 YGCTTCGGCCACCCCGAGGTGGTGAACTGGCTCTTGCATCATGGCRGTGGGGAYCCCACY  4140 4081: rs143577178 C/T;
4141 RYGGCCACAGRCATGGKCGCCCTGCCTATCCACTAYGCTGCCGCC                  1: rs138979381 G/A;
4201 TCCCTGAGGCYTCTCRTCGAGCACTACCCTGAGTAAGATCACCCC                  1: rs113326364 T/C;
4261 TGGGTGGGCTGGCCAGGGCTTTGGGGGATGCCTGGGATTTTCCA
4321 CCAGGGCAATGATCCCTCCAGTGGCCATCCTGGGGCCAGAGGGCC
4381 CTCCCACTCAACATGAAATTTTCCCCTCCTGGAAAACCCCTTCTG
4441 CCCTGCAAGCAGGTGCTCCCAGCATCCTCAGCTGCCCGGCCGCAC
4501 AGGCTGGGCACACAGGCCAAGGTCACCTGTTCCCCTTGGGCTGCT
```

Variation: rs145666801
Position: 1:6488373
Alleles: G/A
Types: Non-synonymous coding
Upstream
Gene/Transcript Locations

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



Gene: ESPN ENSG00000187017

Description: espin [Source:HGNC Symbol;Acc:13281]

Location: Chromosome 1: 6,484,848-6,521,430 forward strand.

Transcripts: This gene has 10 transcripts (splice variants) Show transcript table

Variation table

Summary of variation consequences in ENSG00000187017

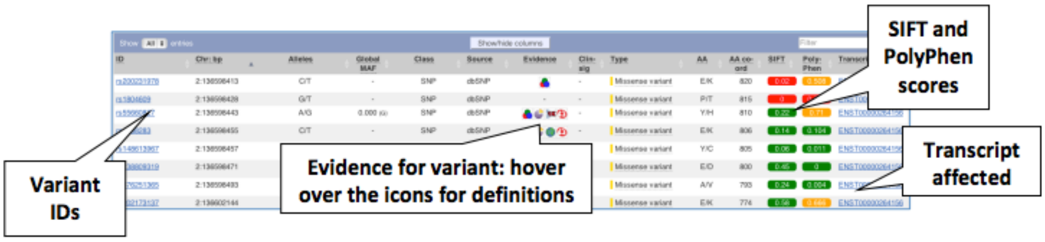
Switch to tree view

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) |
| 1 | Show | 🟧 | Splice donor variant | A splice variant that changes the 2 base region at the 5' end of an intron (SO:0001575) |
| 0 | - | 🟧 | 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) |
| 0 | - | 🟪 | 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) |
| 1 | Show | 🟨 | 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.

Click on Show to expand a detailed table for any of the consequence types available such as Missense variants:



SIFT and PolyPhen scores

Transcript affected

Variant IDs

Evidence for variant: hover over the icons for definitions

Show All entries Show/hide columns Filter

| ID | Chr: bp | Alleles | Global MAF | Class | Source | Evidence | Clin. sig | Type | AA | AA co-ord | SIFT | Poly-Phen | Transcript |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs200231976 | 2:136598413 | C/T | - | SNP | dbSNP | ● | - | Missense variant | E/K | 820 | 0.02 | 0.508 | [illegible] |
| rs1804609 | 2:136598428 | G/T | - | SNP | dbSNP | - | - | Missense variant | P/T | 815 | 0 | [illegible] | ENST00[illegible] |
| rs55960[illegible]7 | 2:136598443 | A/G | 0.000 (G) | SNP | dbSNP | ● ● ● ● | - | Missense variant | Y/H | 810 | 0.22 | 0.71 | ENST00000264156 |
| [illegible]283 | 2:136598455 | C/T | - | SNP | dbSNP | ● ● | - | Missense variant | E/K | 806 | 0.14 | 0.104 | ENST00000264156 |
| rs148613967 | 2:136598457 | | | | | | | Missense variant | Y/C | 805 | 0.06 | 0.011 | ENST00000264156 |
| [illegible]38809319 | 2:136598471 | | | | | | | Missense variant | E/D | 800 | 0.45 | 0 | ENST00000264156 |
| [illegible]76251365 | 2:136598493 | | | | | | | Missense variant | A/V | 793 | 0.24 | 0.004 | ENST[illegible] |
| [illegible]32173137 | 2:136602144 | | | | | | | Missense variant | E/K | 774 | 0.58 | 0.666 | ENST00000264156 |

Let’s have a look at some of the regulatory data that is available for the *ESPN* gene in human. You are still in the Gene tab.

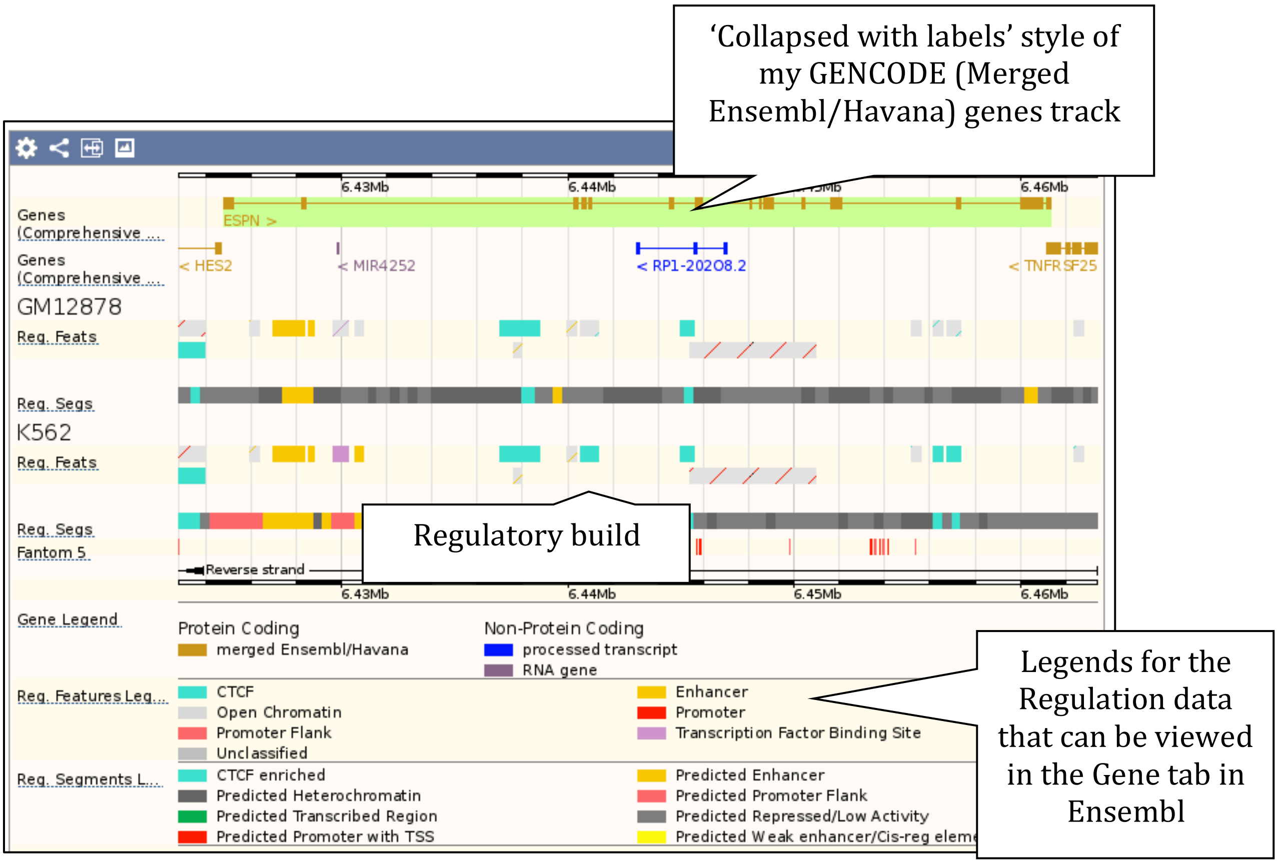
In the left hand side menu of the Gene page, click on the *Regulation* link.

This page shows all the available data from the Regulatory build of Ensembl. The source of this data is mainly from the ENCODE project.

This view can also be configured to show cell specific data. Lets choose to display data for GM12878 and K562 cells only.

More information on the sources of this data (including what the cells are) can be found below:

http://www.ensembl.org/info/genome/funcgen/regulation_sources.html



'Collapsed with labels' style of my GENCODE (Merged Ensembl/Havana) genes track

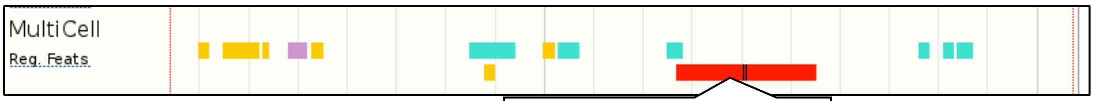
Regulatory build

Legends for the Regulation data that can be viewed in the Gene tab in Ensembl

The different blocks show the regions where the available biochemical data (e.g. ChIP-sequencing to assess Transcription Factor Binding Sites, or TFBS, and modified histones) maps to on the human genome. Check the different colours in the legend to find out more about this annotation e.g. Enhancer, Promoter, etc…

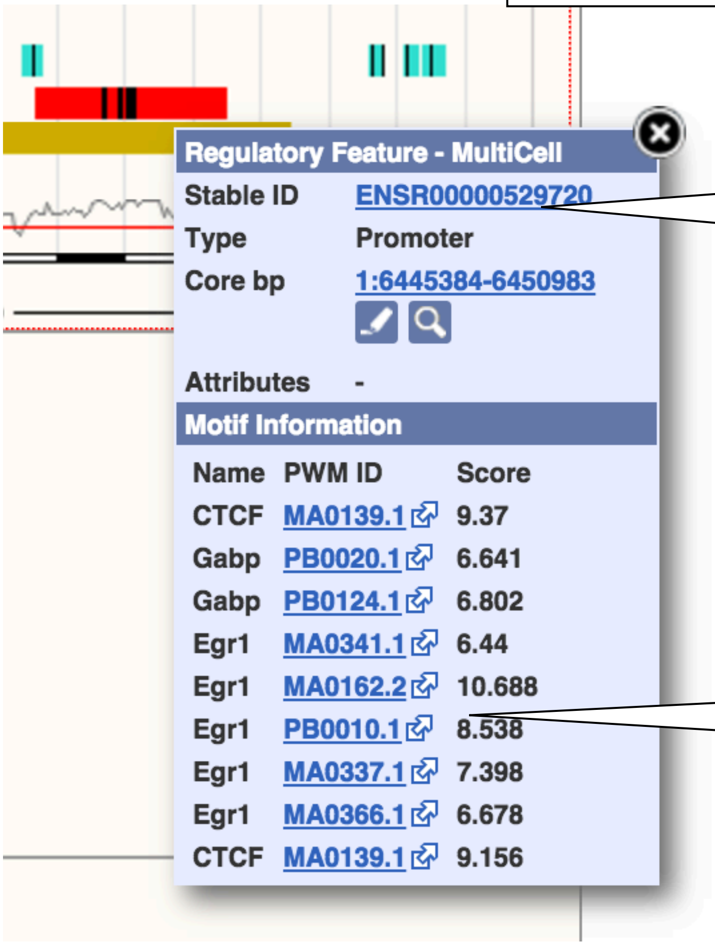
We provide a summary of these data across 18 cells in humans (MultiCell) and five cells in mouse as part of our Regulatory build:

http://www.ensembl.org/info/genome/funcgen/regulatory_build.html



Black lines indicate TF motifs

Click on the regulatory features ) to learn more.



**Regulatory Feature - MultiCell**

| | |
|---|---|
| Stable ID | ENSR00000529720 |
| Type | Promoter |
| Core bp | 1:6445384-6450983 |
| Attributes | - |

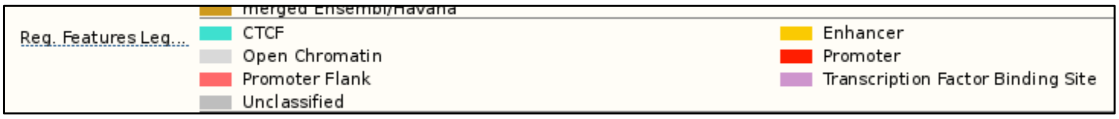
**Motif Information**

| Name | PWM ID | Score |
|---|---|---|
| CTCF | MA0139.1 | 9.37 |
| Gabp | PB0020.1 | 6.641 |
| Gabp | PB0124.1 | 6.802 |
| Egr1 | MA0341.1 | 6.44 |
| Egr1 | MA0162.2 | 10.688 |
| Egr1 | PB0010.1 | 8.538 |
| Egr1 | MA0337.1 | 7.398 |
| Egr1 | MA0366.1 | 6.678 |
| CTCF | MA0139.1 | 9.156 |

Regulatory feature ID, ENSR00000529720. Click on it to go to the Regulation tab

TF motifs within Regulatory feature ENSR00000529720

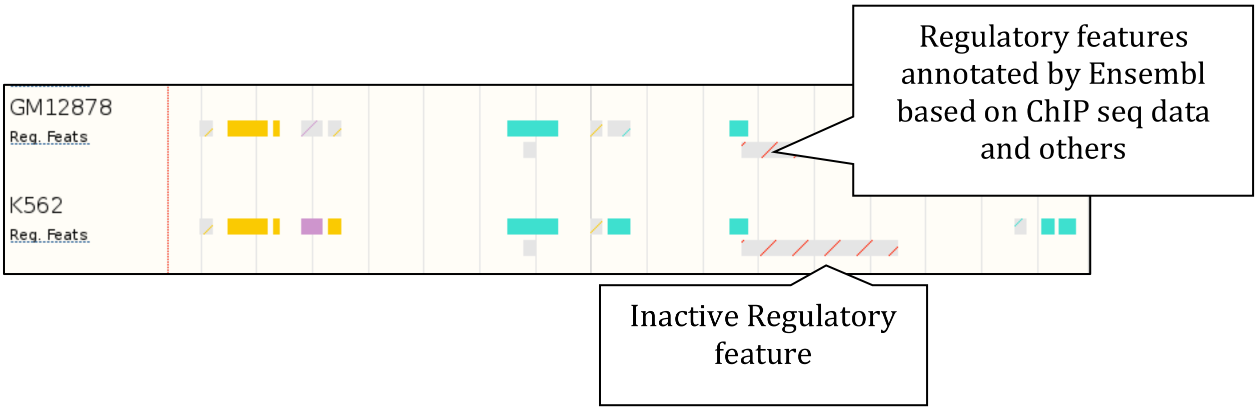
Note the different colours to indicate specific annotation at the Gene Regulation level: regions can be associated with promoter, enhancer, CTCF, TFBS, open chromatin activity, etc. See the legend for more details:



Reg. Features Leg...

| | |
|---|---|
| merged Ensembl/Havana | |
| CTCF | Enhancer |
| Open Chromatin | Promoter |
| Promoter Flank | Transcription Factor Binding Site |
| Unclassified | |

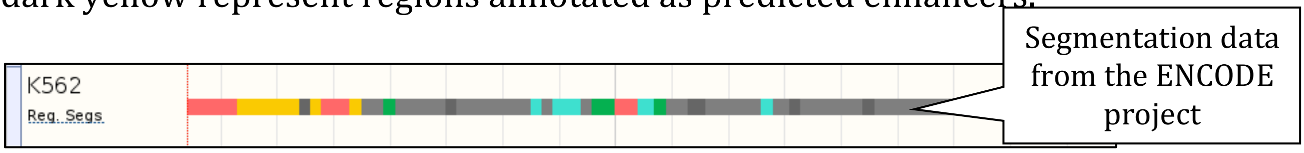
We also provide cell–specific data (e.g. Reg. Feats for CD4, Reg. Feats GM12878, etc). Click on 'Configure this page' to add cells specific data.



Regulatory features annotated by Ensembl based on ChIP seq data and others

Inactive Regulatory feature

In addition to the Regulatory Features, Segmentation data from the ENCODE project can be displayed in this view. See the legend for more details on the different colours for this data track. For example, dark yellow represent regions annotated as predicted enhancers.

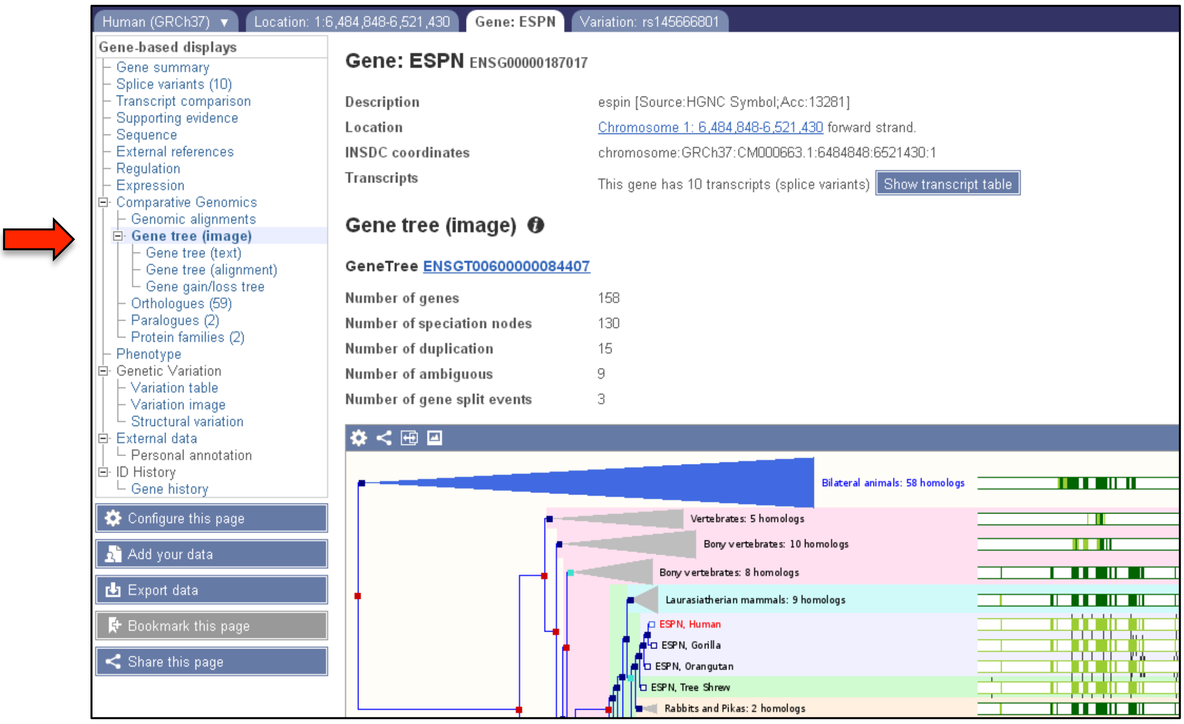


Segmentation data from the ENCODE project

Details on the Segmentation data in Ensembl can be found below:

http://www.ensembl.org/info/genome/funcgen/regulatory_segmentation.html

Let's now have a look at some of the views on Comparative Genomics in the Ensembl Browser. In the left hand menu in the Gene tab, click on the 'Gene tree' link. This will display the Ensembl gene in the context of a phylogenetic tree used to determine orthologues and paralogues.



**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 |

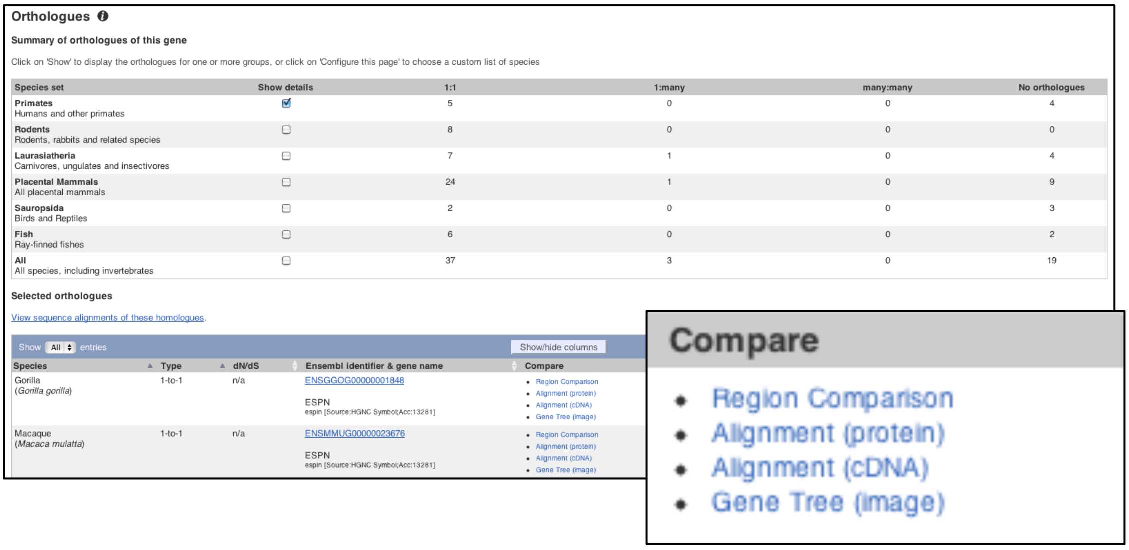
**Gene tree (image)**

GeneTree ENSGT00600000084407

| | |
|---|---|
| Number of genes | 158 |
| Number of speciation nodes | 130 |
| Number of duplication | 15 |
| Number of ambiguous | 9 |
| Number of gene split events | 3 |

Bilateral animals: 58 homologs
Vertebrates: 5 homologs
Bony vertebrates: 10 homologs
Bony vertebrates: 8 homologs
Laurasiatherian mammals: 9 homologs
ESPN, Human
ESPN, Gorilla
ESPN, Orangutan
ESPN, Tree Shrew
Rabbits and Pikas: 2 homologs

Click the *Orthologues* link at the left of this page to view homologues detected by this tree. Note the links under Compare.



**Orthologues**

**Summary of orthologues of this gene**

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

| Species set | Show details | 1:1 | 1:many | many:many | No orthologues |
|---|---|---|---|---|---|
| **Primates**<br>Humans and other primates | ☑ | 5 | 0 | 0 | 4 |
| **Rodents**<br>Rodents, rabbits and related species | □ | 8 | 0 | 0 | 0 |
| **Laurasiatheria**<br>Carnivores, ungulates and insectivores | □ | 7 | 1 | 0 | 4 |
| **Placental Mammals**<br>All placental mammals | □ | 24 | 1 | 0 | 9 |
| **Sauropsida**<br>Birds and Reptiles | □ | 2 | 0 | 0 | 3 |
| **Fish**<br>Ray-finned fishes | □ | 6 | 0 | 0 | 2 |
| **All**<br>All species, including invertebrates | □ | 37 | 3 | 0 | 19 |

**Selected orthologues**

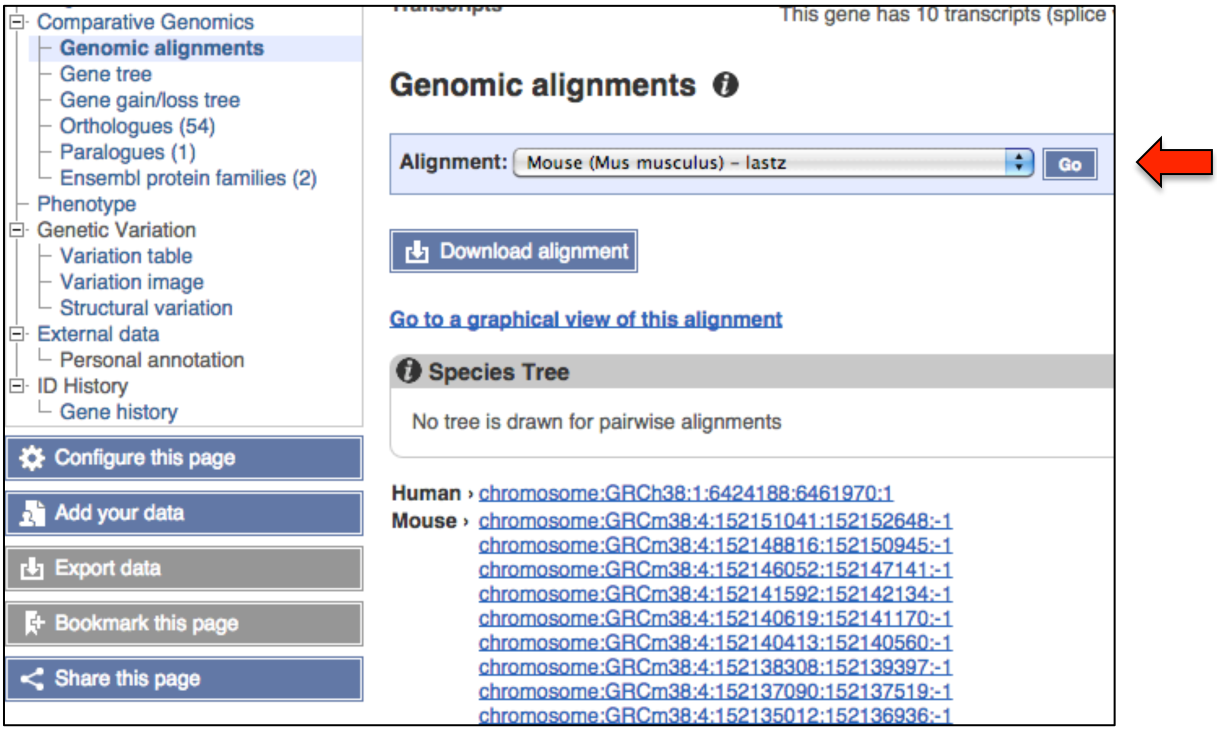
View sequence alignments of these homologues.

| Species | Type | dN/dS | Ensembl identifier & gene name | Compare |
|---|---|---|---|---|
| Gorilla<br>(*Gorilla gorilla*) | 1-to-1 | n/a | ENSGGOG00000001848<br>ESPN<br>espin [Source:HGNC Symbol;Acc:13281] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA)<br>• Gene Tree (image) |
| Macaque<br>(*Macaca mulatta*) | 1-to-1 | n/a | ENSMMUG00000023676<br>ESPN<br>espin [Source:HGNC Symbol;Acc:13281] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA)<br>• Gene Tree (image) |

**Compare**

- Region Comparison
- Alignment (protein)
- Alignment (cDNA)
- Gene Tree (image)

Let's have a look at the pairwise alignment between human and mouse. Click on *Genomic alignments*.



This gene has 10 transcripts (splice

Genomic alignments

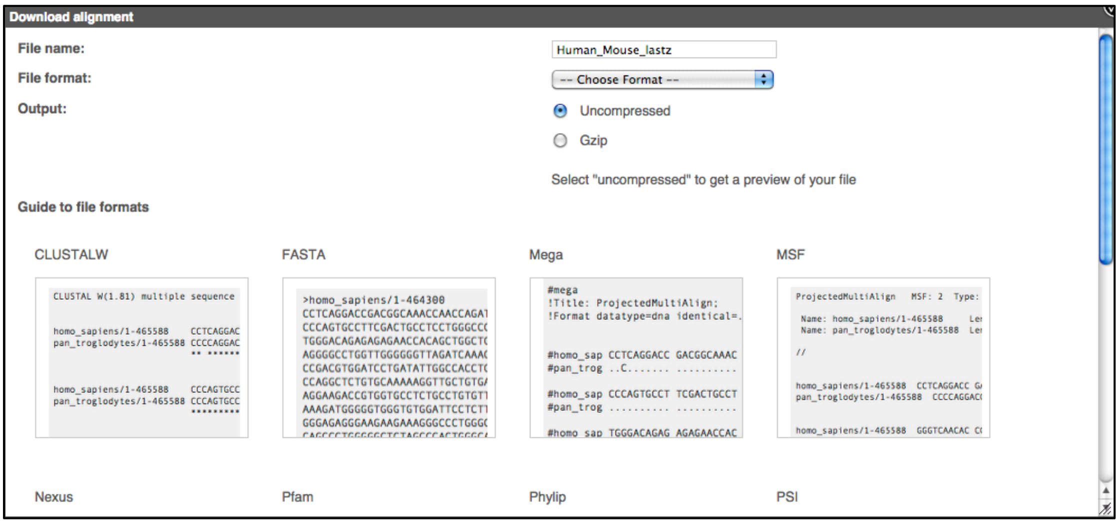
Alignment: Mouse (Mus musculus) – lastz

Species Tree

No tree is drawn for pairwise alignments

**Human ›** chromosome:GRCh38:1:6424188:6461970:1
**Mouse ›** chromosome:GRCm38:4:152151041:152152648:-1
chromosome:GRCm38:4:152148816:152150945:-1
chromosome:GRCm38:4:152146052:152147141:-1
chromosome:GRCm38:4:152141592:152142134:-1
chromosome:GRCm38:4:152140619:152141170:-1
chromosome:GRCm38:4:152140413:152140560:-1
chromosome:GRCm38:4:152138308:152139397:-1
chromosome:GRCm38:4:152137090:152137519:-1
chromosome:GRCm38:4:152135012:152136936:-1

You can download the alignment in multiple formats:



Download alignment

File name: Human_Mouse_lastz
File format: -- Choose Format --
Output: ● Uncompressed
○ Gzip

Select "uncompressed" to get a preview of your file

Guide to file formats

CLUSTALW

```
CLUSTAL W(1.81) multiple sequence

homo_sapiens/1-465588     CCTCAGGAC
pan_troglodytes/1-465588  CCCCAGGAC
                          ** ******

homo_sapiens/1-465588     CCCAGTGCC
pan_troglodytes/1-465588  CCCAGTGCC
                          *********
```

FASTA

```
>homo_sapiens/1-464300
CCTCAGGACCGACGGCAAACCAACCAGAT
CCCAGTGCCTTCGACTGCCTCCTGGGCCC
TGGGACAGAGAGAGAACCACAGCTGGCTC
AGGGGCCTGGTTGGGGGGTTAGATCAAAC
CCGACGTGGATCCTGATATTGGCCACCTC
CCAGGCTCTGTGCAAAAAGGTTGCTGTGA
AGGAAGACCGTGGTGCCTCTGCCTGTGTT
AAAGATGGGGGTGGGTGTGGATTCCTCTT
GGGAGAGGGAAGAAGAAAGGGCCCTGGGC
```

Mega

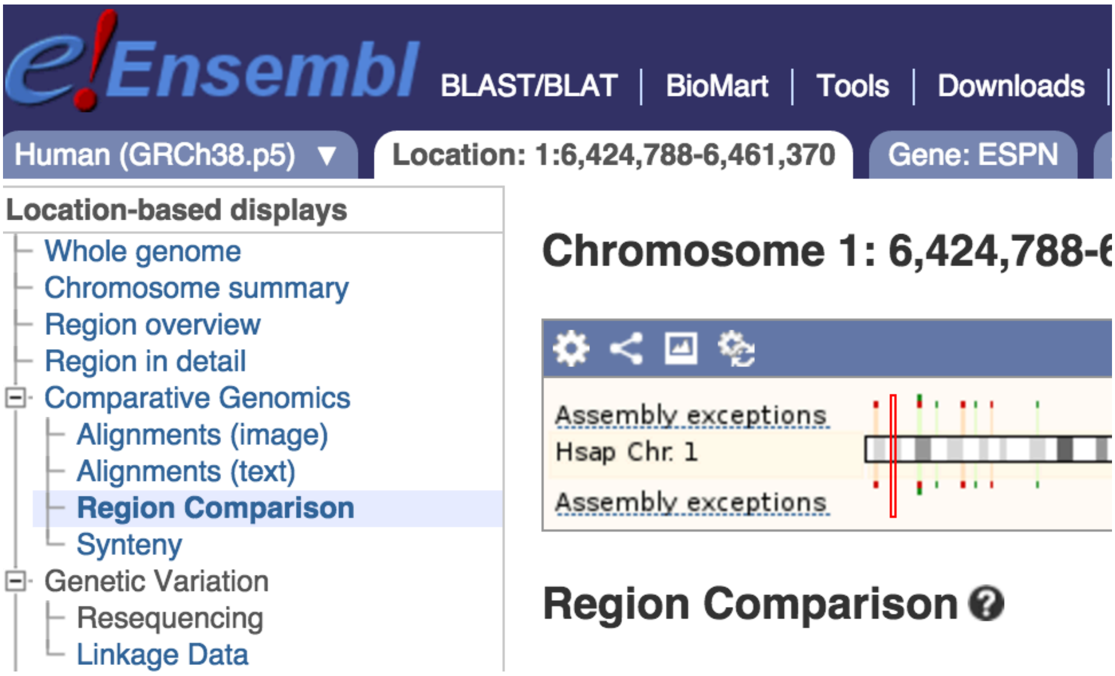
```
#mega
!Title: ProjectedMultiAlign;
!Format datatype=dna identical=.

#homo_sap CCTCAGGACC GACGGCAAAC
#pan_trog ..C....... ..........

#homo_sap CCCAGTGCCT TCGACTGCCT
#pan_trog .......... ..........

#homo_sap TGGGACAGAG AGAGAACCAC
```

MSF

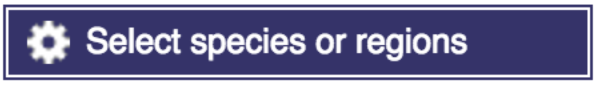
```
ProjectedMultiAlign   MSF: 2  Type:

 Name: homo_sapiens/1-465588     Le
 Name: pan_troglodytes/1-465588  Le

//

homo_sapiens/1-465588  CCTCAGGACC G
pan_troglodytes/1-465588  CCCCAGGAC

homo_sapiens/1-465588  GGGTCAACAC C
```

Nexus    Pfam    Phylip    PSI

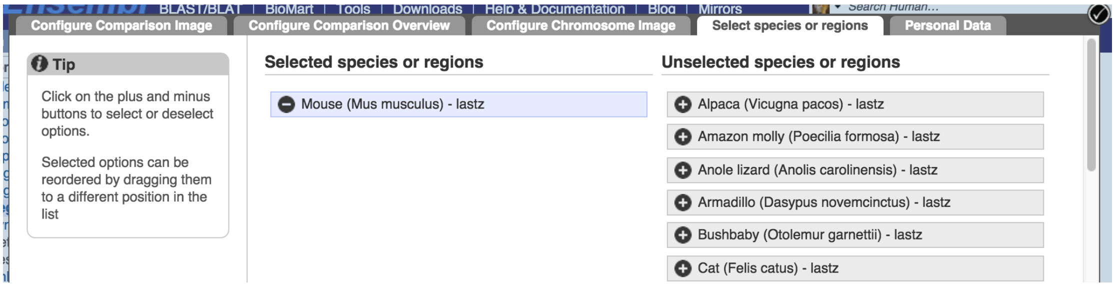
Other comparative genomics views are available from the Location tab such as Region comparison.



Click on the link in the left hand menu and choose the species to compare against my human *ESPN* gene



for example mouse


BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Search Human…
Configure Comparison Image
Configure Comparison Overview
Configure Chromosome Image
Select species or regions
Personal Data

Tip

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

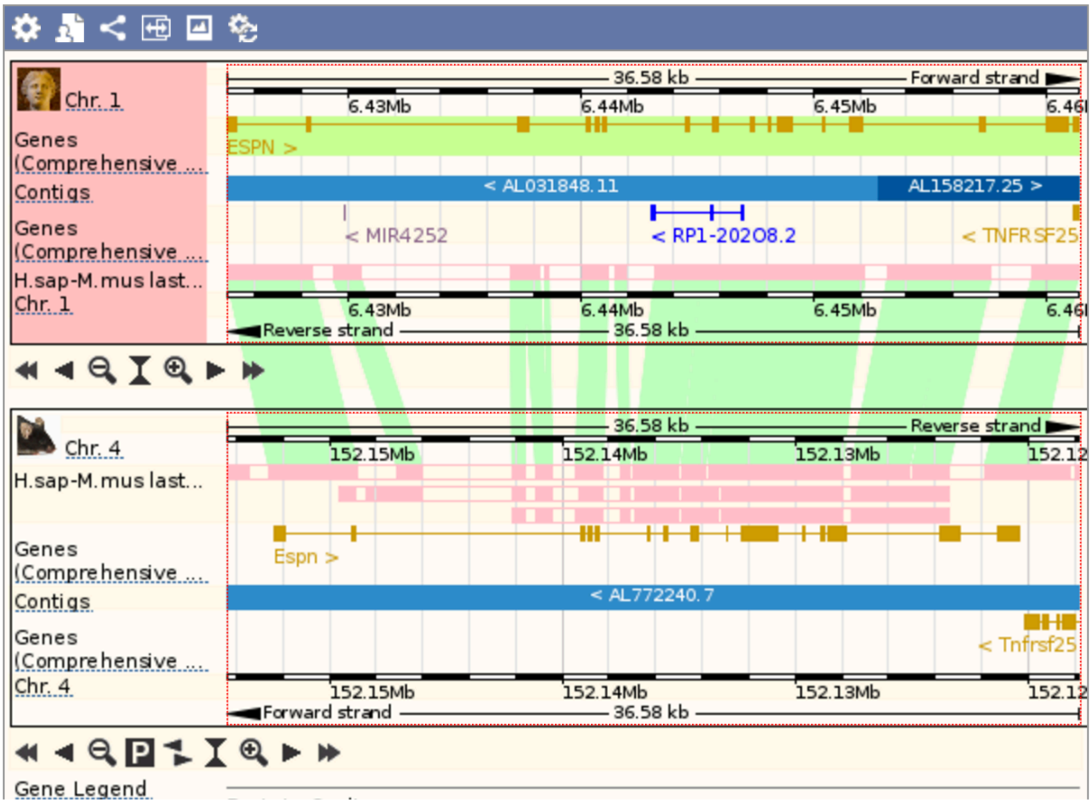
Selected species or regions

- Mouse (Mus musculus) - lastz

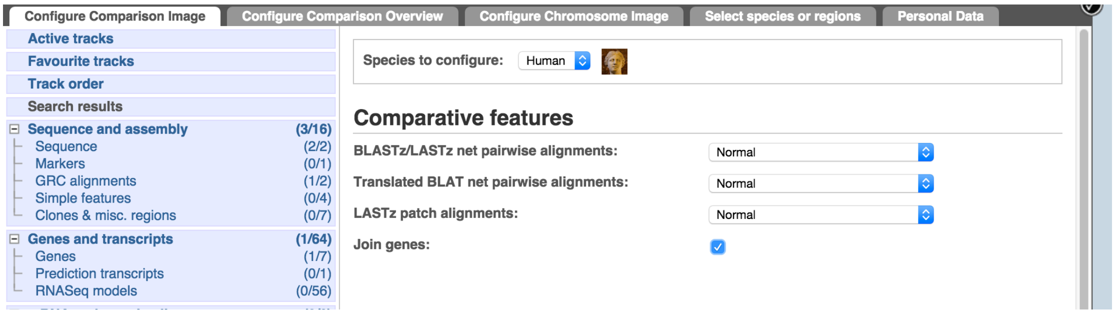
Unselected species or regions

- Alpaca (Vicugna pacos) - lastz
- Amazon molly (Poecilia formosa) - lastz
- Anole lizard (Anolis carolinensis) - lastz
- Armadillo (Dasypus novemcinctus) - lastz
- Bushbaby (Otolemur garnettii) - lastz
- Cat (Felis catus) - lastz

This is what you will see:



You can zoom in and out in this view and configure it to show a blue line (under Comparative features) to connect the orthologous genes. Tick the ‘Join genes’ box:


Configure Comparison Image
Configure Comparison Overview
Configure Chromosome Image
Select species or regions
Personal Data
Active tracks
Favourite tracks
Track order
Search results
Sequence and assembly (3/16)
Sequence (2/2)
Markers (0/1)
GRC alignments (1/2)
Simple features (0/4)
Clones & misc. regions (0/7)
Genes and transcripts (1/64)
Genes (1/7)
Prediction transcripts (0/1)
RNASeq models (0/56)

Species to configure: Human

Comparative features

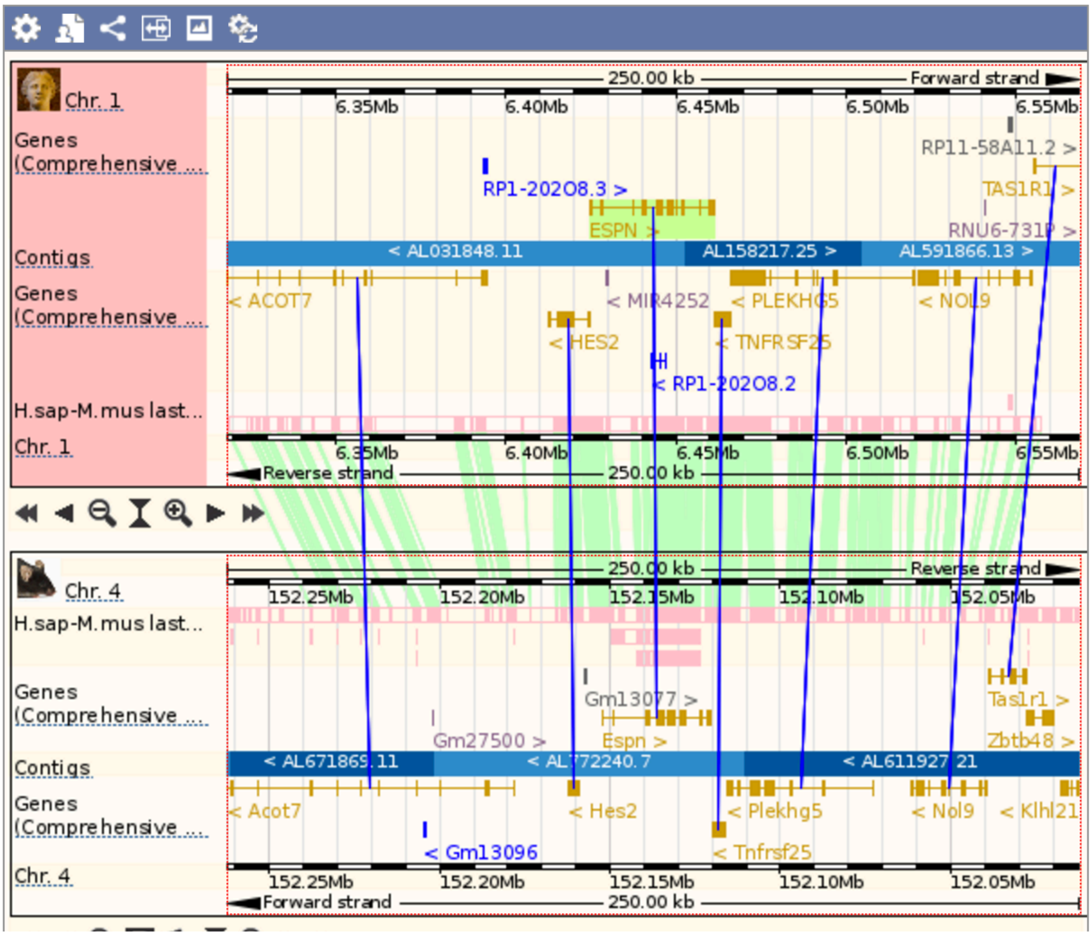
BLASTz/LASTz net pairwise alignments: Normal

Translated BLAT net pairwise alignments: Normal

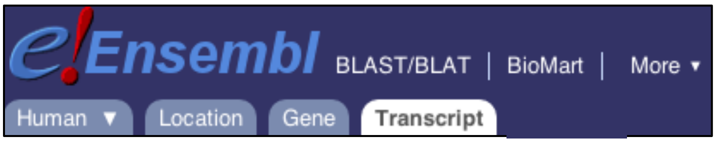
LASTz patch alignments: Normal

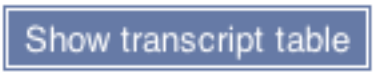
Join genes: ☑

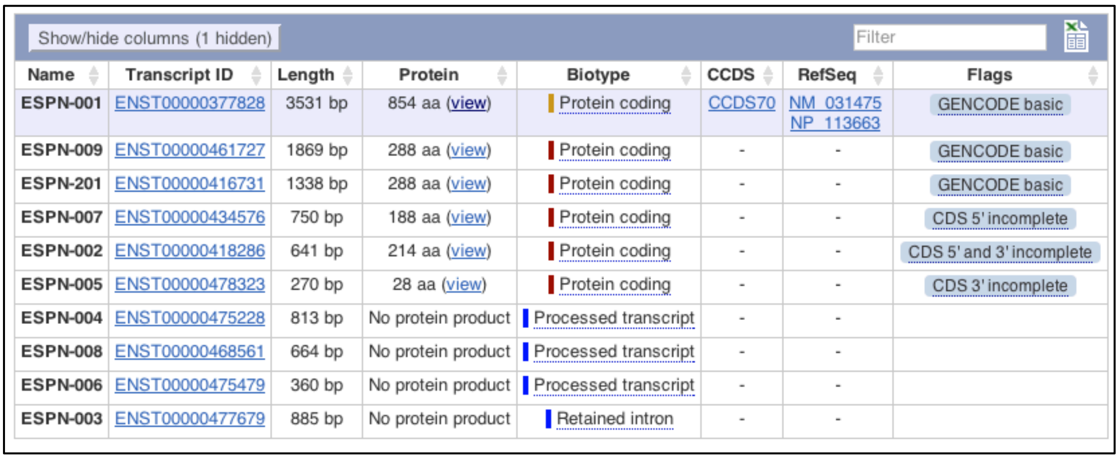
Save and close the new configuration to view an image like the one below:



Let’s now move to the Transcript tab to explore one splice isoform of the *ESPN* gene.





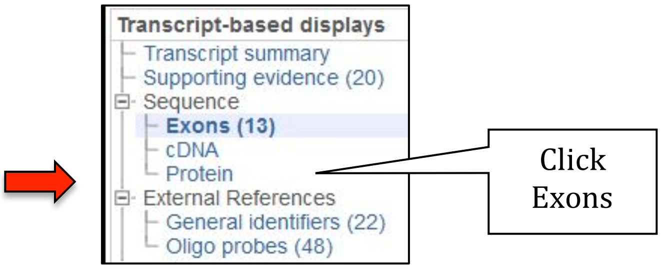


Show/hide columns (1 hidden)

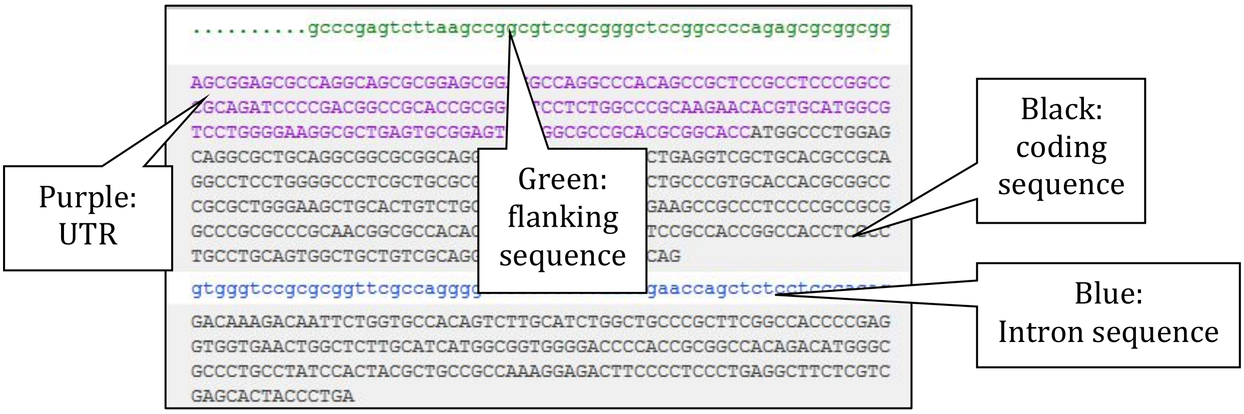
| Name | Transcript ID | Length | Protein | Biotype | CCDS | RefSeq | Flags |
|---|---|---|---|---|---|---|---|
| ESPN-001 | ENST00000377828 | 3531 bp | 854 aa (view) | Protein coding | CCDS70 | NM_031475<br>NP_113663 | GENCODE basic |
| ESPN-009 | ENST00000461727 | 1869 bp | 288 aa (view) | Protein coding | - | - | GENCODE basic |
| ESPN-201 | ENST00000416731 | 1338 bp | 288 aa (view) | Protein coding | - | - | GENCODE basic |
| ESPN-007 | ENST00000434576 | 750 bp | 188 aa (view) | Protein coding | - | - | CDS 5' incomplete |
| ESPN-002 | ENST00000418286 | 641 bp | 214 aa (view) | Protein coding | - | - | CDS 5' and 3' incomplete |
| ESPN-005 | ENST00000478323 | 270 bp | 28 aa (view) | Protein coding | - | - | CDS 3' incomplete |
| ESPN-004 | ENST00000475228 | 813 bp | No protein product | Processed transcript | - | - | |
| ESPN-008 | ENST00000468561 | 664 bp | No protein product | Processed transcript | - | - | |
| ESPN-006 | ENST00000475479 | 360 bp | No protein product | Processed transcript | - | - | |
| ESPN-003 | ENST00000477679 | 885 bp | No protein product | Retained intron | - | - | |

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

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.



Click Exons



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

AGCGGAGCGCCAGGCAGCGCGGAGCGG  GCCAGGCCCACAGCCGCTCCGCCTCCCGGCC
CGCAGATCCCCGACGGCCGCACCGCGG  TCCTCTGGCCCGCAAGAACACGTGCATGGCG
TCCTGGGGAAGGCGCTGAGTGCGGAGT  GGCGCCGCACGCGGCACCATGGCCCTGGAG
CAGGCGCTGCAGGCGGCGCGGCAGC    CTGAGGTCGCTGCACGCCGCA
GGCCTCCTGGGGCCCTCGCTGCGCC    CTGCCCGTGCACCACGCGGCC
CGCGCTGGGAAGCTGCACTGTCTGC    GAAGCCGCCCTCCCCGCCGCG
GCCCGCGCCCGCAACGGCGCCACAC    TCCGCCACCGGCCACCTCGCC
TGCCTGCAGTGGCTGCTGTCGCAGC    CAG
gtgggtccgcgcggttcgccagggg    gaaccagctctcctcccagag
GACAAAGACAATTCTGGTGCCACAGTCTTGCATCTGGCTGCCCGCTTCGGCCACCCCGAG
GTGGTGAACTGGCTCTTGCATCATGGCGGTGGGGACCCCACCGCGGCCACAGACATGGGC
GCCCTGCCTATCCACTACGCTGCCGCCAAAGGAGACTTCCCCTCCCTGAGGCTTCTCGTC
GAGCACTACCCTGA
```

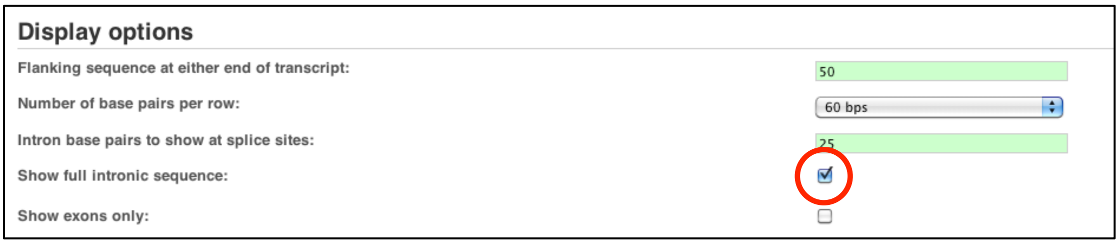
Purple: UTR

Green: flanking sequence

Black: coding sequence

Blue: Intron sequence

You may want to change the display (to show more flanking sequence or to show full introns, for example). 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

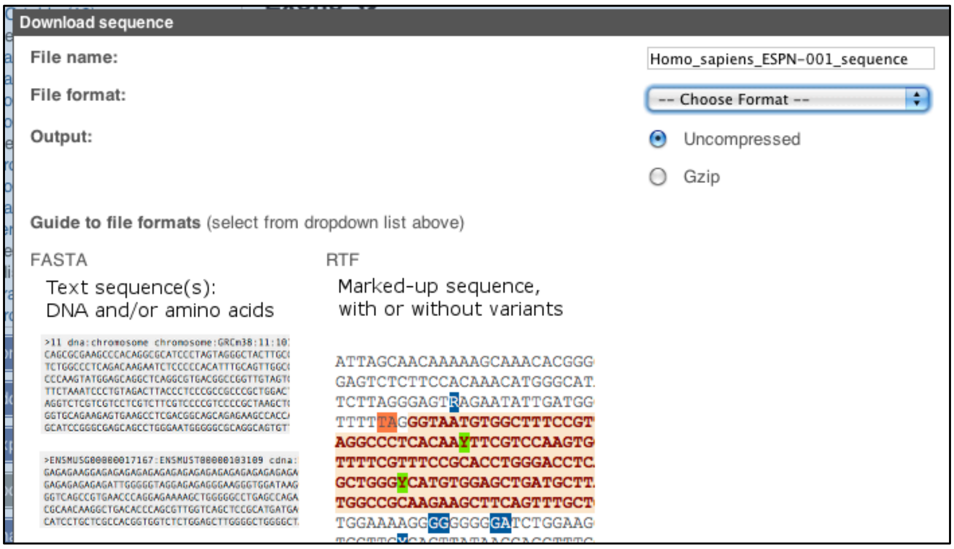
Number of base pairs per row: 60 bps

Intron base pairs to show at splice sites: 25

Show full intronic sequence: ☑

Show exons only: □

If you would like to export the sequence either as FASTA or RTF (Rich Text Format), click on



Download sequence

File name: Homo_sapiens_ESPN-001_sequence

File format: -- Choose Format --

Output: ● Uncompressed ○ Gzip

**Guide to file formats** (select from dropdown list above)

FASTA

Text sequence(s):
DNA and/or amino acids

```
>11 dna:chromosome chromosome:GRCm38:11:10
CAGCGCGAAGCCCACAGGCGCATCCCTAGTAGGGCTACTTGCC
TCTGGCCCTCAGACAAGAATCTCCCCCACATTTGCAGTTGGCC
CCCAAGTATGGAGCAGGCTCAGGCGTGACGGCCGGTTGTAGTC
TTCTAAATCCCTGTAGACTTACCCTCCCGCCGCCCGCTGGAC
AGGTCTCGTCGTCCTCGTCTTCGTCCCCGTCCCCGCTAAGCTC
GGTGCAGAAGAGTGAAGCCTCGACGGCAGCAGAGAAGCCACCA
GCATCCGGGCGAGCAGCCTGGGAATGGGGGCGCAGGCAGTGT
```

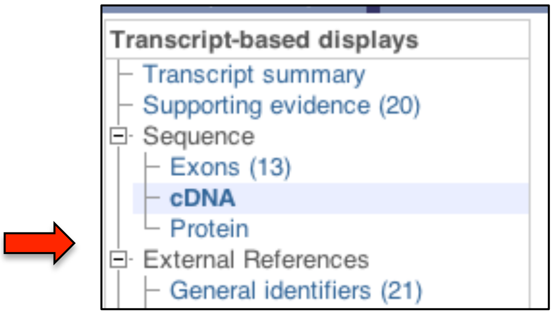
```
>ENSMUSG00000017167:ENSMUST00000103109 cdna:
GAGAGAAGGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA
GAGAGAGAGAGATTGGGGGTAGGAGAGAGGGAAGGGTGGATAAG
GGTCAGCCGTGAACCCAGGAGAAAAGCTGGGGGCCTGAGCCAGA
CGCAACAAGGCTGACACCCAGCGTTGGTCAGCTCCGCATGATGA
CATCCTGCTCGCCACGGTGGTCTCTGGAGCTTGGGGCTGGGGCT
```

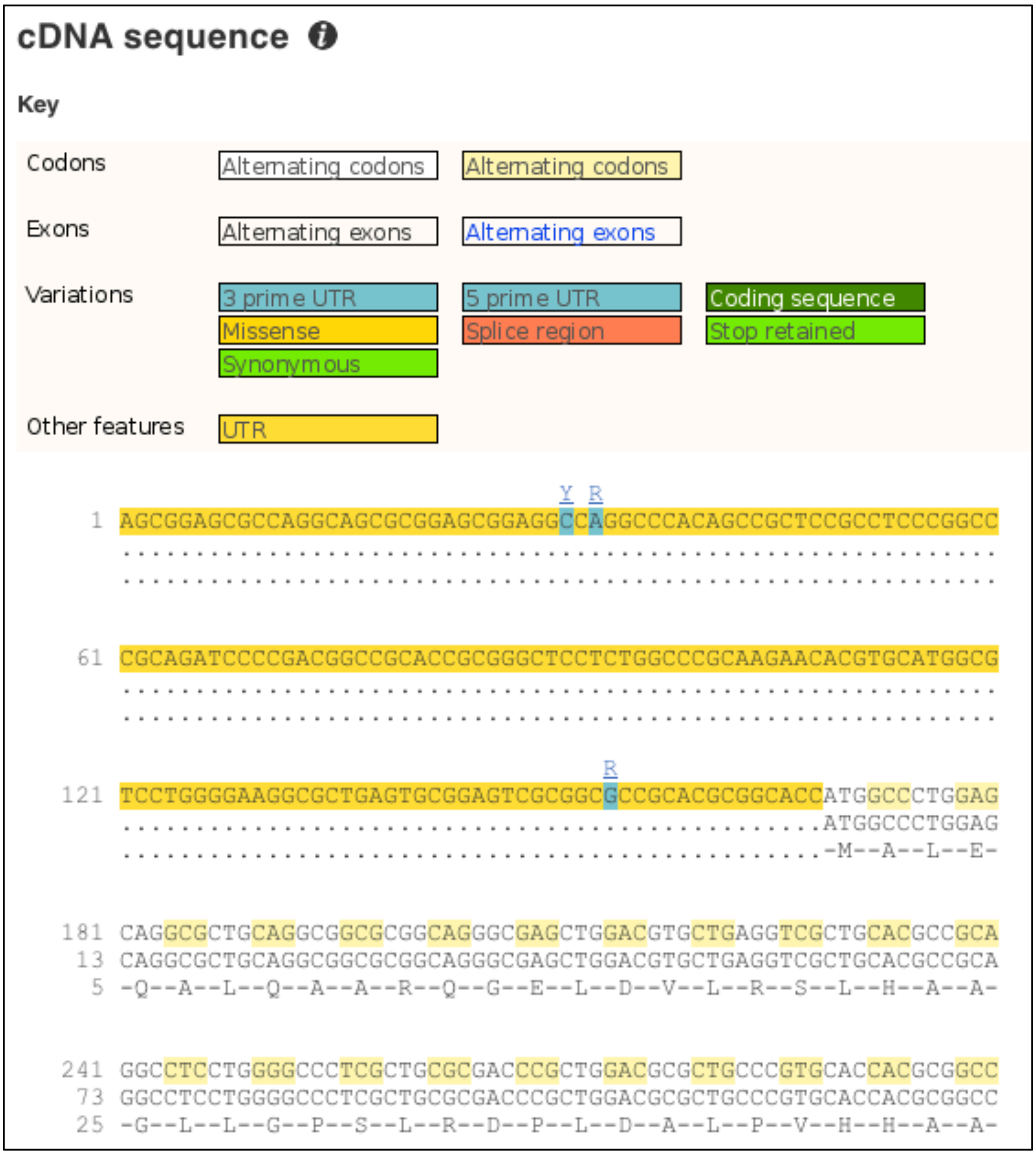
RTF

Marked-up sequence,
with or without variants

```
ATTAGCAACAAAAAGCAAACACGGG
GAGTCTCTTCCACAAACATGGGCAT
TCTTAGGGAGTRAGAATATTGATGG
TTTTTAGGGTAATGTGGCTTTCCGT
AGGCCCTCACAAYTTCGTCCAAGTG
TTTTCGTTTCCGCACCTGGGACCTC
GCTGGGYCATGTGGAGCTGATGCTT
TGGCCGCAAGAAGCTTCAGTTTGCT
TGGAAAAGGGGGGGGGATCTGGAAG
```

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





cDNA sequence

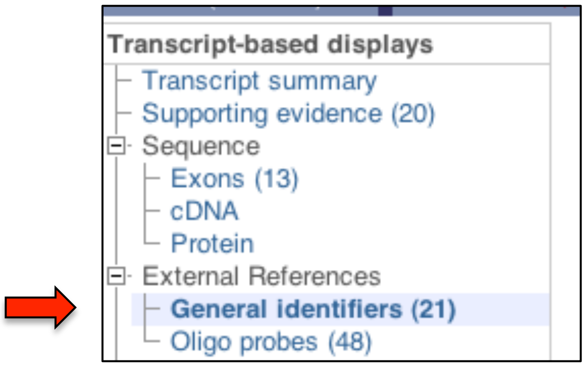
Key

| | | | |
|---|---|---|---|
| Codons | Alternating codons | Alternating codons | |
| Exons | Alternating exons | Alternating exons | |
| Variations | 3 prime UTR<br>Missense<br>Synonymous | 5 prime UTR<br>Splice region | Coding sequence<br>Stop retained |
| Other features | UTR | | |

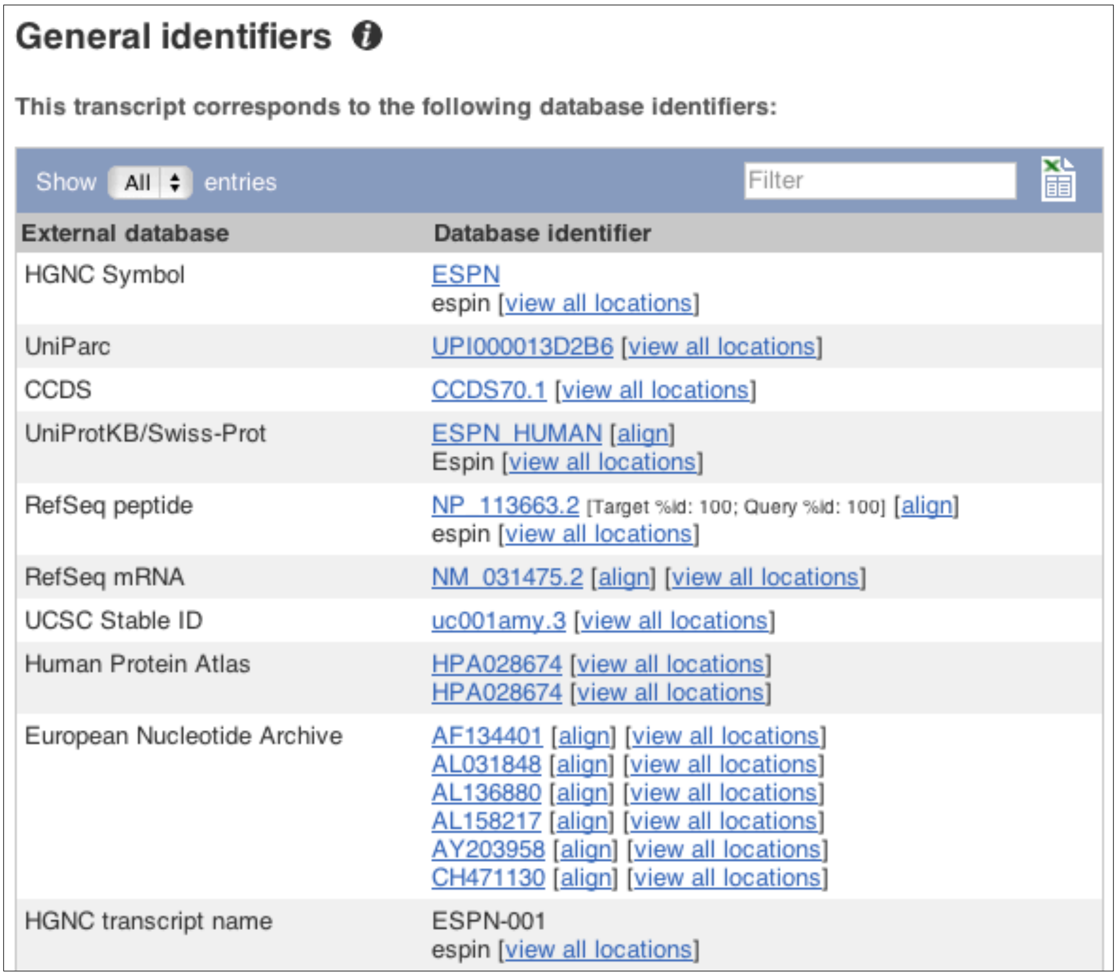
```
                                    Y R
  1 AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCC
    ............................................................
    ............................................................

 61 CGCAGATCCCCGACGGCCGCACCGCGGGCTCCTCTGGCCCGCAAGAACACGTGCATGGCG
    ............................................................
    ............................................................

                                        R
121 TCCTGGGGAAGGCGCTGAGTGCGGAGTCGCGGCGCCGCACGCGGCACCATGGCCCTGGAG
    ................................................ATGGCCCTGGAG
    ................................................-M--A--L--E-

181 CAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCTGGACGTGCTGAGGTCGCTGCACGCCGCA
 13 CAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCTGGACGTGCTGAGGTCGCTGCACGCCGCA
  5 -Q--A--L--Q--A--A--R--Q--G--E--L--D--V--L--R--S--L--H--A--A-

241 GGCCTCCTGGGGCCCTCGCTGCGCGACCCGCTGGACGCGCTGCCCGTGCACCACGCGGCC
 73 GGCCTCCTGGGGCCCTCGCTGCGCGACCCGCTGGACGCGCTGCCCGTGCACCACGCGGCC
 25 -G--L--L--G--P--S--L--R--D--P--L--D--A--L--P--V--H--H--A--A-
```

UnTranslated Regions (UTRs) are highlighted in 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. This page shows data from other resources, such as HGNC, RefSeq, CCDS, EntrezGene, OMIM, UniProtKB, and others, that match to the Ensembl transcript and protein.



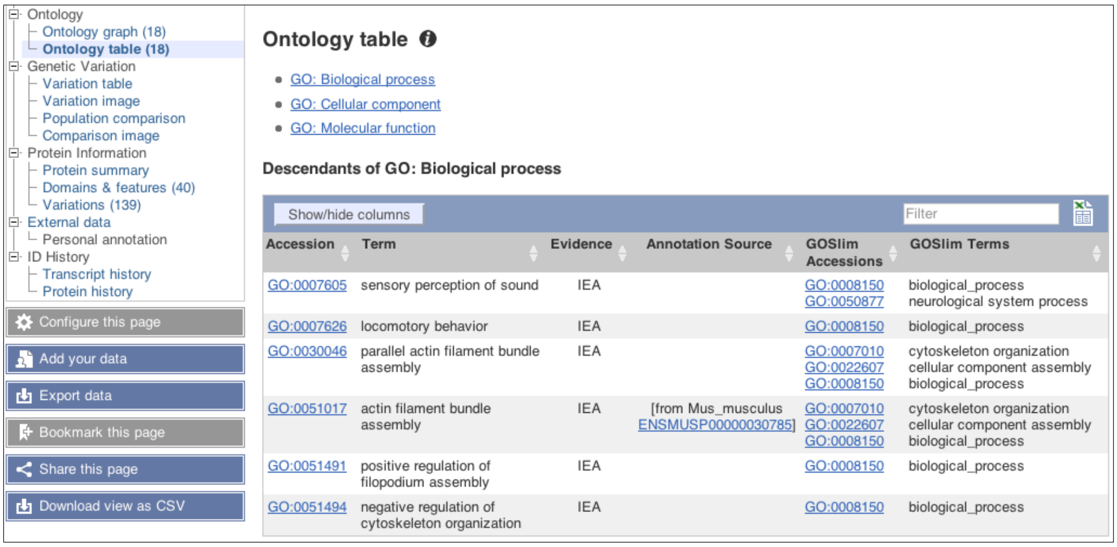
Links to matches to this Ensembl *ESPN* transcript in other databases


General identifiers

This transcript corresponds to the following database identifiers:

| External database | Database identifier |
|---|---|
| HGNC Symbol | ESPN<br>espin [view all locations] |
| UniParc | UPI000013D2B6 [view all locations] |
| CCDS | CCDS70.1 [view all locations] |
| UniProtKB/Swiss-Prot | ESPN_HUMAN [align]<br>Espin [view all locations] |
| RefSeq peptide | NP_113663.2 [Target %id: 100; Query %id: 100] [align]<br>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]<br>HPA028674 [view all locations] |
| European Nucleotide Archive | AF134401 [align] [view all locations]<br>AL031848 [align] [view all locations]<br>AL136880 [align] [view all locations]<br>AL158217 [align] [view all locations]<br>AY203958 [align] [view all locations]<br>CH471130 [align] [view all locations] |
| HGNC transcript name | ESPN-001<br>espin [view all locations] |

Now, 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**

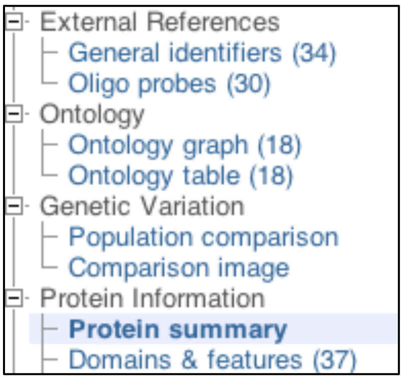
- GO: Biological process
- GO: Cellular component
- GO: Molecular function

**Descendants of GO: Biological process**

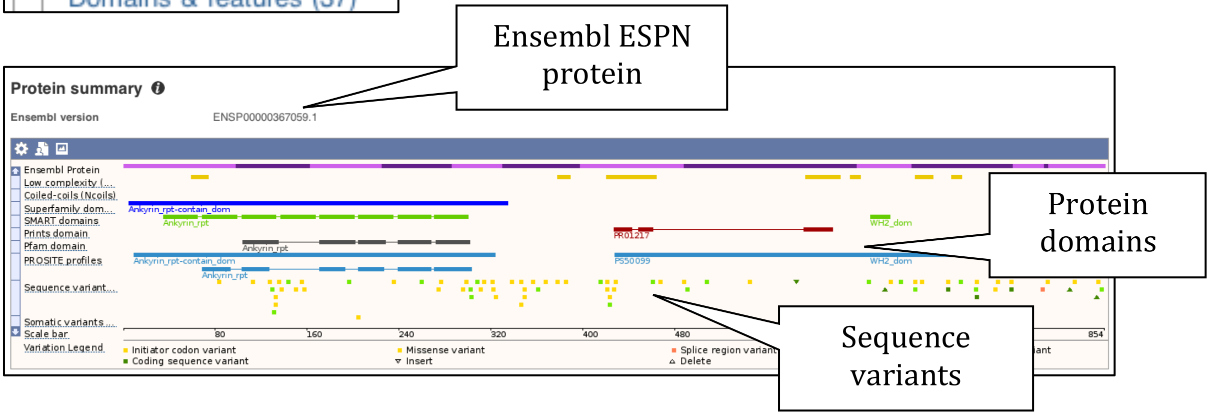
| Accession | Term | Evidence | Annotation Source | GOSlim Accessions | GOSlim Terms |
|---|---|---|---|---|---|
| GO:0007605 | sensory perception of sound | IEA | | GO:0008150<br>GO:0050877 | biological_process<br>neurological system process |
| GO:0007626 | locomotory behavior | IEA | | GO:0008150 | biological_process |
| GO:0030046 | parallel actin filament bundle assembly | IEA | | GO:0007010<br>GO:0022607<br>GO:0008150 | cytoskeleton organization<br>cellular component assembly<br>biological_process |
| GO:0051017 | actin filament bundle assembly | IEA | [from Mus_musculus ENSMUSP00000030785] | GO:0007010<br>GO:0022607<br>GO:0008150 | cytoskeleton organization<br>cellular component assembly<br>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.

Click on *Protein summary* to view domains from Pfam, PROSITE, Superfamily, InterPro, and more.



- External References
  - General identifiers (34)
  - Oligo probes (30)
- Ontology
  - Ontology graph (18)
  - Ontology table (18)
- Genetic Variation
  - Population comparison
  - Comparison image
- Protein Information
  - **Protein summary**
  - Domains & features (37)



**Protein summary** ⓘ

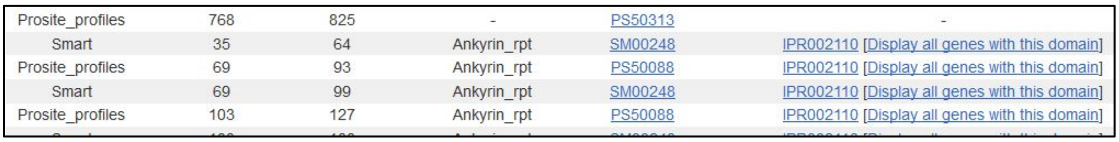
Ensembl version ENSP00000367059.1

Ensembl ESPN protein

Protein domains

Sequence variants

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



| | | | | | |
|---|---|---|---|---|---|
| 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] |
| Prosite_profiles | 103 | 127 | Ankyrin_rpt | PS50088 | IPR002110 [Display all genes with this domain] |

Our last task is to export genomic sequence and perform a similarity search using the new Ensembl BLAST/BLAT tool.

You can export a sequence from different views in Ensembl but lets go back to the Location tab.

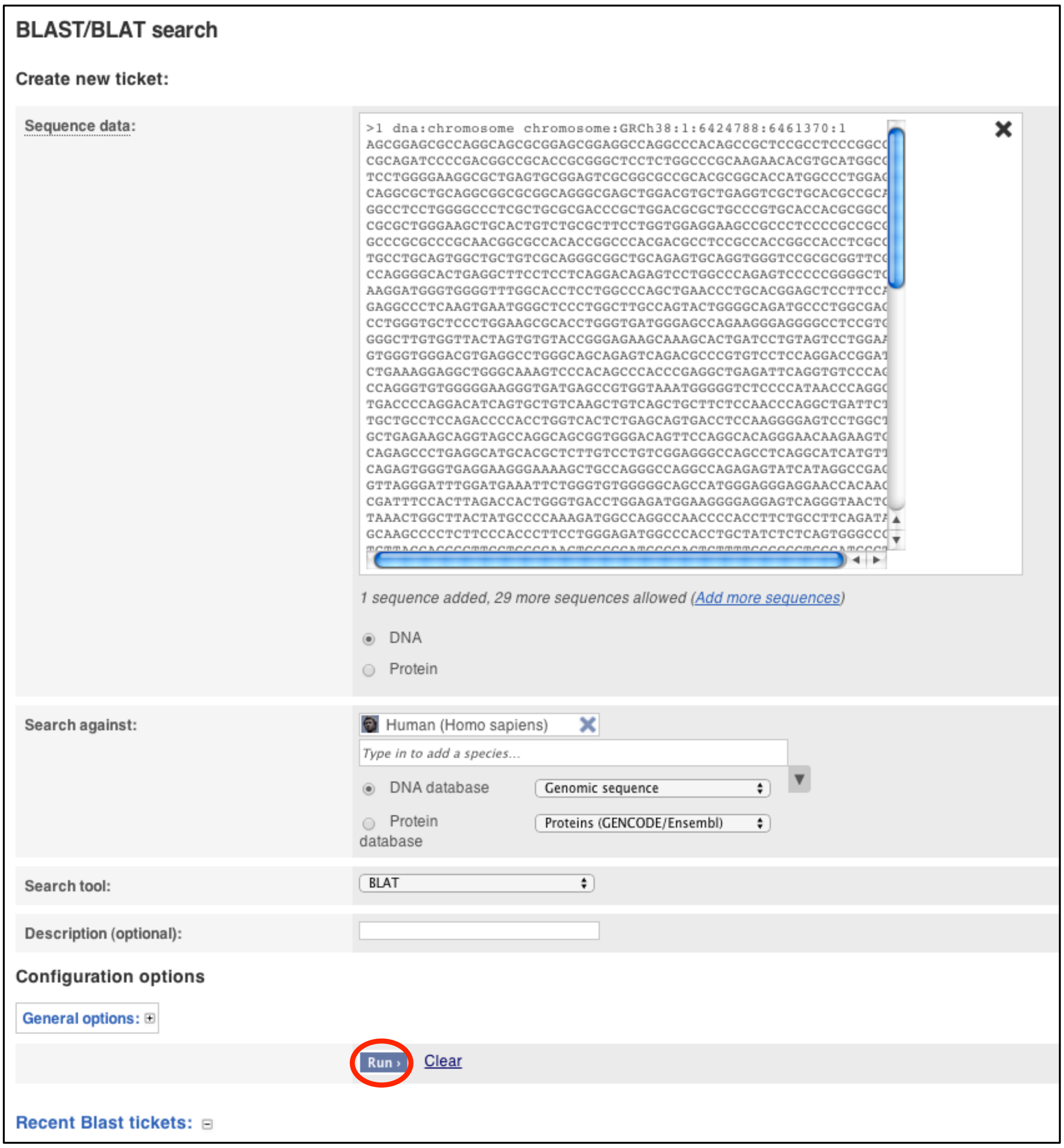
Click on the Export data option, select the default parameters (e.g. Fasta sequence as output) and click Next then HTML. Please note that you can export the genomic sequence as either unmasked or masked (soft or hard masked).

This is a snapshot what you will see in your browser:

```
>1 dna:chromosome chromosome:GRCh37:1:6484848:6521430:1
AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCCCGCAGATCCC
CGACGGCCGCACCGCGGGCTCCTCTGGCCCGCAAGAACACGTGCATGGCGTCCTGGGGAAGGCGCTGAG
TGCGGAGTCGCGGCGCCGCACGCGGCACCATGGCCCTGGAGCAGGCGCTGCAGGCGGCGCGGCAGGGCG
AGCTGGACGTGCTGAGGTCGCTGCACGCCGCAGGCCTCCTGGGGCCCTCGCTGCGCGACCCGCTGGACGC
GCTGCCCGTGCACCACGCGGCCCGCGCTGGGAAGCTGCACTGTCTGCGCTTCCTGGTGGAGGAAGCCGCC
CTCCCCGCCGCGGCCCGCGCCCGCAACGGCGCCACACCGGCCCACGACGCCTCCGCCACCGGCCACCTCGC
CTGCCTGCAGTGGCTGCTGTCGCAGGGCGGCTGCAGAGTGCAGGTGGGTCCGCGCGGTTCGCCAGGGGC
ACTGAGGCTTCCTCCTCAGGACAGAGTCCTGGCCCAGAGTCCCCCGGGGCTCAAGGATGGGTGGGGTTT
GGCACCTCCTGGCCCAGCTGAACCCTGCACGGAGCTCCTTCCAGAGGCCCTCAAGTGAATGGGCTCCCTG
GCTTGCCAGTACTGGGGCAGATGCCCTGGCGAGCCTGGGTGCTCCCTGGAAGCGCACCTGGGTGATGGG
AGCCAGAAGGGAGGGGCCTCCGTG
```

To use this sequence for similarity searches, you can select the header and a few lines of the nucleotide sequence and then copy it on the clipboard. Click on the BLAST/BLAT link in the bar at the top of the main Ensembl homepage. Paste the sequence into the appropriate box and select BLAT as the search algorithm.

Click Run.



BLAST/BLAT search

Create new ticket:

Sequence data:

```
>1 dna:chromosome chromosome:GRCh38:1:6424788:6461370:1
AGCGGAGCGCCAGGCAGCGCGGAGCGGAGGCCAGGCCCACAGCCGCTCCGCCTCCCGGCC
CGCAGATCCCCGACGGCCGCACCGCGGGCTCCTCTGGCCCGCAAGAACACGTGCATGGCC
TCCTGGGGAAGGCGCTGAGTGCGGAGTCGCGGCGCCGCACGCGGCACCATGGCCCTGGAG
CAGGCGCTGCAGGCGGCGCGGCAGGGCGAGCTGGACGTGCTGAGGTCGCTGCACGCCGCA
GGCCTCCTGGGGCCCTCGCTGCGCGACCCGCTGGACGCGCTGCCCGTGCACCACGCGGCC
CGCGCTGGGAAGCTGCACTGTCTGCGCTTCCTGGTGGAGGAAGCCGCCCTCCCCGCCGCC
GCCCGCGCCCGCAACGGCGCCACACCGGCCCACGACGCCTCCGCCACCGGCCACCTCGCC
TGCCTGCAGTGGCTGCTGTCGCAGGGCGGCTGCAGAGTGCAGGTGGGTCCGCGCGGTTCG
CCAGGGGCACTGAGGCTTCCTCCTCAGGACAGAGTCCTGGCCCAGAGTCCCCCGGGGCTC
AAGGATGGGTGGGGTTTGGCACCTCCTGGCCCAGCTGAACCCTGCACGGAGCTCCTTCCA
GAGGCCCTCAAGTGAATGGGCTCCCTGGCTTGCCAGTACTGGGGCAGATGCCCTGGCGAG
CCTGGGTGCTCCCTGGAAGCGCACCTGGGTGATGGGAGCCAGAAGGGAGGGGCCTCCGTG
GGGCTTGTGGTTACTAGTGTGTACCGGGAGAAGCAAAGCACTGATCCTGTAGTCCTGGAA
GTGGGTGGGACGTGAGGCCTGGGCAGCAGAGTCAGACGCCCGTGTCCTCCAGGACCGGAT
CTGAAAGGAGGCTGGGCAAAGTCCCACAGCCCACCCGAGGCTGAGATTCAGGTGTCCCAG
CCAGGGTGTGGGGGAAGGGTGATGAGCCGTGGTAAATGGGGGTCTCCCCATAACCCAGGC
TGACCCCAGGACATCAGTGCTGTCAAGCTGTCAGCTGCTTCTCCAACCCAGGCTGATTCT
TGCTGCCTCCAGACCCCACCTGGTCACTCTGAGCAGTGACCTCCAAGGGGAGTCCTGGCT
GCTGAGAAGCAGGTAGCCAGGCAGCGGTGGGACAGTTCCAGGCACAGGGAACAAGAAGTG
CAGAGCCCTGAGGCATGCACGCTCTTGTCCTGTCGGAGGGCCAGCCTCAGGCATCATGTT
CAGAGTGGGTGAGGAAGGGAAAAGCTGCCAGGGCCAGGCCAGAGAGTATCATAGGCCGAG
GTTAGGGATTTGGATGAAATTCTGGGTGTGGGGGCAGCCATGGGAGGGAGGAACCACAAC
CGATTTCCACTTAGACCACTGGGTGACCTGGAGATGGAAGGGGAGGAGTCAGGGTAACTC
TAAACTGGCTTACTATGCCCCAAAGATGGCCAGGCCAACCCCACCTTCTGCCTTCAGATA
GCAAGCCCCTCTTCCCACCCTTCCTGGGAGATGGCCCACCTGCTATCTCTCAGTGGGCCC
```

*1 sequence added, 29 more sequences allowed (Add more sequences)*

● DNA

○ Protein

Search against: Human (Homo sapiens)

Type in to add a species…

● DNA database: Genomic sequence

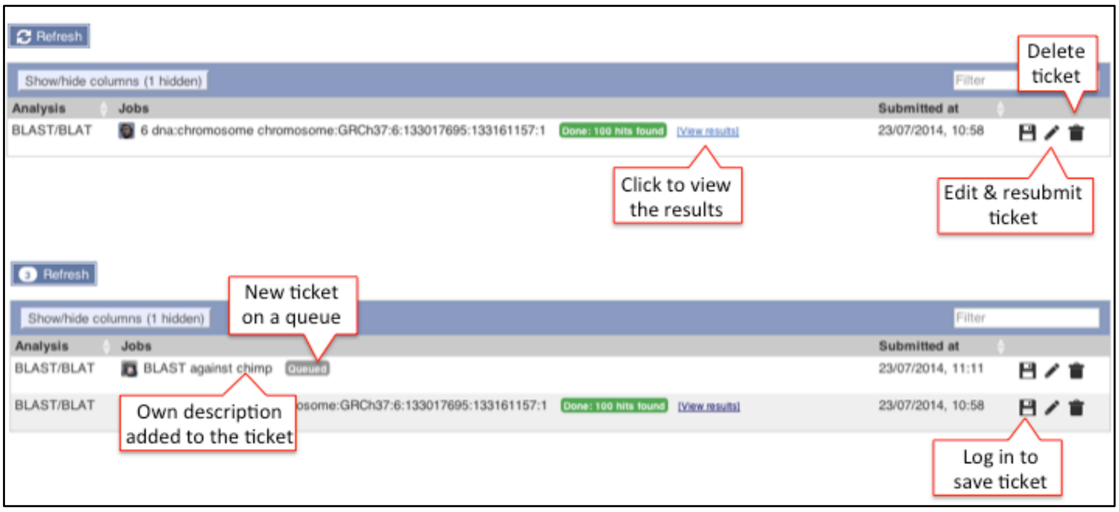
○ Protein database: Proteins (GENCODE/Ensembl)

Search tool: BLAT

Description (optional):

Configuration options

The table with results should look like these examples. The jobs are given a ticket number and highlighted as green when successfully completed.



| Analysis | Jobs | | | Submitted at |
|---|---|---|---|---|
| BLAST/BLAT | 6 dna:chromosome chromosome:GRCh37:6:133017695:133161157:1 | Done: 100 hits found | [View results] | 23/07/2014, 10:58 |

Delete ticket

Click to view the results

Edit & resubmit ticket

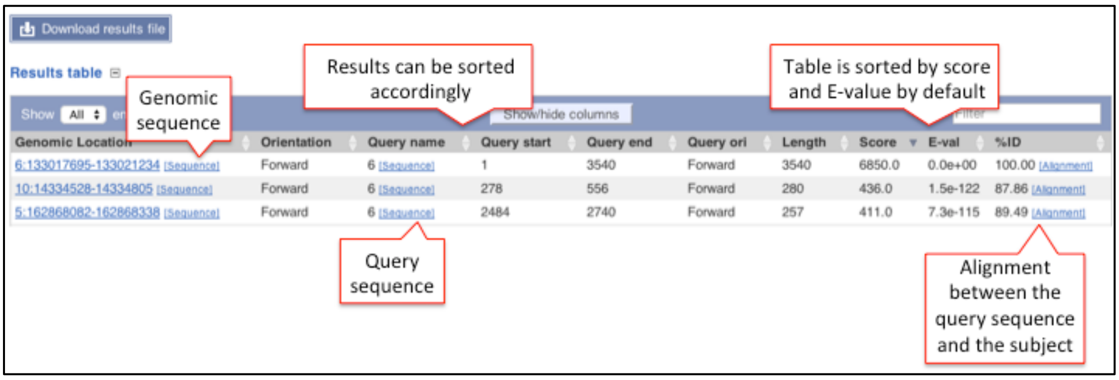
| Analysis | Jobs | | | Submitted at |
|---|---|---|---|---|
| BLAST/BLAT | BLAST against chimp | Queued | | 23/07/2014, 11:11 |
| BLAST/BLAT | osome:GRCh37:6:133017695:133161157:1 | Done: 100 hits found | [View results] | 23/07/2014, 10:58 |

New ticket on a queue

Own description added to the ticket

Log in to save ticket

Click on view results for a table or karyotype view.



| Genomic Location | Orientation | Query name | Query start | Query end | Query ori | Length | Score | E-val | %ID |
|---|---|---|---|---|---|---|---|---|---|
| 6:133017695-133021234 [Sequence] | Forward | 6 [Sequence] | 1 | 3540 | Forward | 3540 | 6850.0 | 0.0e+00 | 100.00 [Alignment] |
| 10:14334528-14334805 [Sequence] | Forward | 6 [Sequence] | 278 | 556 | Forward | 280 | 436.0 | 1.5e-122 | 87.86 [Alignment] |
| 5:162868082-162868338 [Sequence] | Forward | 6 [Sequence] | 2484 | 2740 | Forward | 257 | 411.0 | 7.3e-115 | 89.49 [Alignment] |

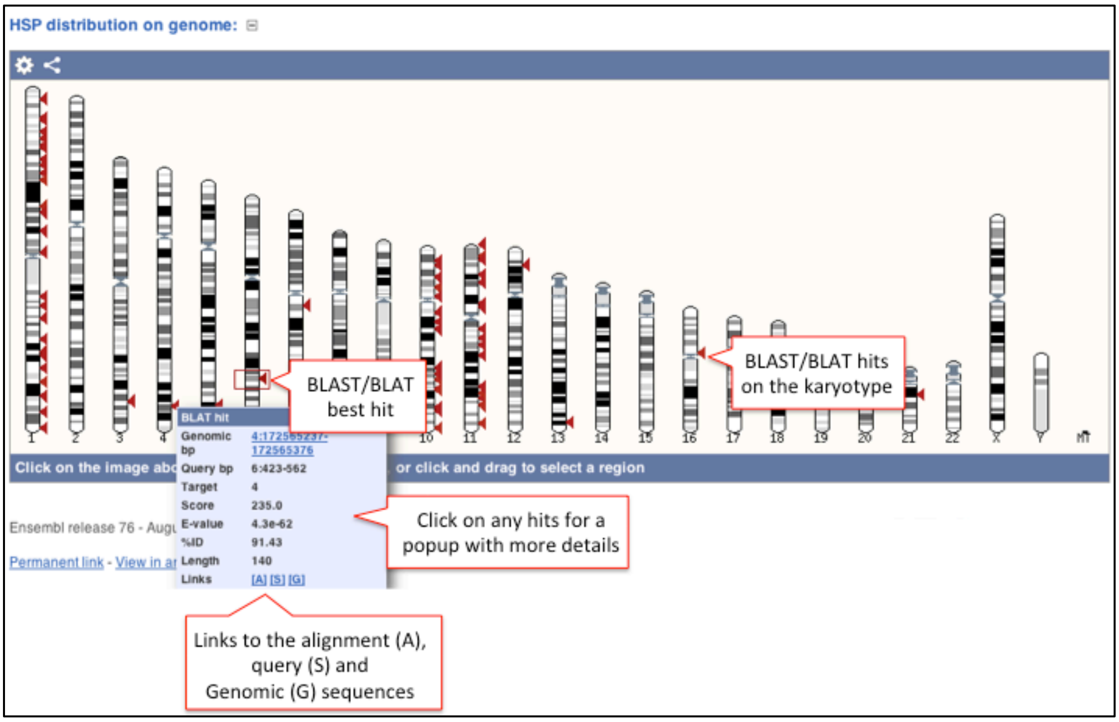
Genomic sequence

Results can be sorted accordingly

Table is sorted by score and E-value by default

Query sequence

Alignment between the query sequence and the subject


HSP distribution on genome:
1
2
3
4
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
MT
Click on the image abo
or click and drag to select a region
Ensembl release 76 - Augu
Permanent link - View in ar

| BLAT hit | |
|---|---|
| Genomic bp | 4:172565237-172565376 |
| Query bp | 6:423-562 |
| Target | 4 |
| Score | 235.0 |
| E-value | 4.3e-62 |
| %ID | 91.43 |
| Length | 140 |
| Links | [A] [S] [G] |

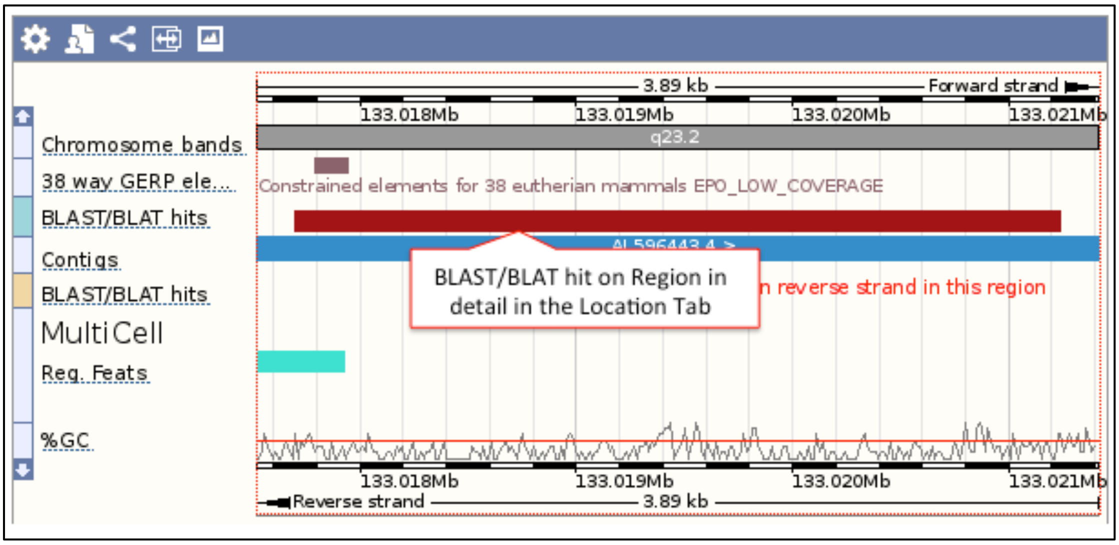
BLAST/BLAT best hit

BLAST/BLAT hits on the karyotype

Click on any hits for a popup with more details

Links to the alignment (A), query (S) and Genomic (G) sequences

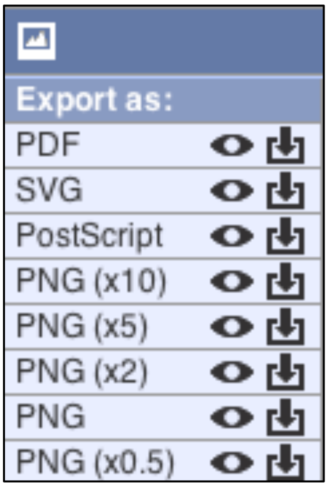
In the Results table, click on the Genomic location link to view the BLAST/BLAT hit in the Region in detail (Locatin tab).



BLAST/BLAT hit on Region in detail in the Location Tab

Click on the red bar for some summary statistics such as the score, %ID, and other BLAST/BLAT values.

Export Image for your lab notebook or publications, or Share it with your colleagues and collaborators!



Export as:
- PDF
- SVG
- PostScript
- PNG (x10)
- PNG (x5)
- PNG (x2)
- PNG
- PNG (x0.5)

END OF THE BROWSER WALKTHROUGH

# EXERCISES

## ENSEMBL BROWSER

### Exercise 1 – Exploring a genomic region in human

a) Go to the region from bp 31,873,863 to 32,623,863 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 (i.e. BAC clones upon which the current genomic assembly was based) that contain the complete *BRCA2* gene?

d) Add the track with RefSeq gene models. This track is known is Ensembl as 'RefSeq human import' Has RefSeq annotated the *BRCA2* gene? If so, how many transcripts have been annotated? Do they differ from the Ensembl transcripts?

e) Save a picture of the main panel so that you can use it in your publication (.png format).

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

g) Delete all tracks you have added to the Location tab.

### Exercise 2 – Exploring a gene and its transcripts in human

a) Find the human *F9* (coagulation factor IX) gene. On which chromosome and which strand of the genome is this gene located? How many transcripts (splice variants) have been annotated for it?

b) What is the longest transcript? How long is the protein it encodes? Has this transcript been annotated automatically (by Ensembl) or manually (by Havana)? How many exons does it have? Are there exons completely or partially untranslated?

c) In which part (i.e. the N-terminal –start- or C-terminal –end) of the protein encoded by ENST00000218099 does its peptidase activity reside? Does the protein contain any transmembrane domains?

d) Have missense variants been discovered for the protein encoded by ENST00000218099?

e) What is the regulatory feature annotated in the region of this gene? Find its ID. Is this feature active in all cells Ensembl has got data for?

f) How many orthologues are predicted for this gene in rodents? How much sequence identity does the mouse protein have to the human one? Note the Target %id and Query %id. View the alignment (protein) between the two sequences.

g) Go to the orthologue in mouse and find the genomic alignment between mouse and human. Can you configure the view to show both START and STOP codons?

---

## *ADDITIONAL EXERCISES: BROWSER*

***If you have finished the exercises above, you may want to do these extra ones***

### *Extra exercise 1 – Mouse assembly, protein domains*

*a) What is the latest assembly of the mouse genome in Ensembl? When was it produced? What is the total number of non-coding genes in the current release? Can you find the previous assembly of the mouse genome (i.e. NCBIM37)?*

*b) Find the Brca1 gene in mouse. Does this gene have any transcript (s) with annotation that has been agreed between the EBI, Sanger, NCBI and UCSC? What is the protein sequence for this transcript? Are there any domains or features present in the amino acid sequence? Can you download this information as a table?*

---

### *Extra exercise 2 – Exploring a gene in Ensembl Bacteria*

*Start in http://bacteria.ensembl.org/index.html and select the Streptomyces lividans 1326 genome.*

*a) What GO: molecular function terms are associated with the 'era' gene?*

*b) What domains can be found in the protein product of this gene? How many different domain prediction methods agree with each of these domains?*

---

### *Extra exercise 3 – miRNA genes in A. thaliana*

*MicroRNAs (miRNAs) are small non-coding RNA molecules (ca. 22 nucleotides) found in plants and animals, which function in transcriptional and post-transcriptional regulation of gene expression. A well-studied miRNA family in plants is the MIR395 family (See also: http://en.wikipedia.org/wiki/MicroRNA and http://en.wikipedia.org/wiki/Mir-395_microRNA_precursor_family).*

*a) How many members does the MIR395 family in Arabidopsis thaliana have?*

*b) How are the MIR395 genes organised? Are they clustered? Are they all located on the same strand of the genome? How are they positioned relative to each other?*

---

# ENSEMBL TOOLS: BIOMART

*Follow these instructions to guide yourself through BioMart and answer the following query:*

You have three questions about this set of human genes: *ESPN, MYH9, USH1C, CHD7, CISD2, THRB, DFNB31*

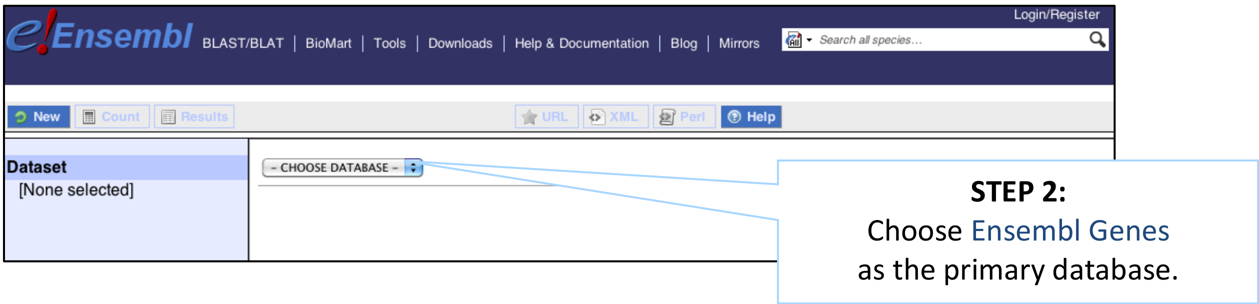
Note: These are HGNC gene symbols.

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?

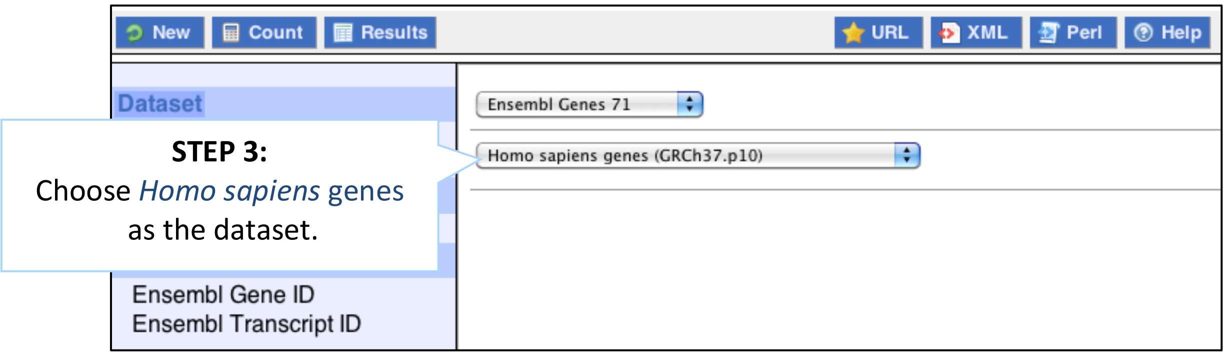
3) What are their cDNA sequences?

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



Dataset
[None selected]

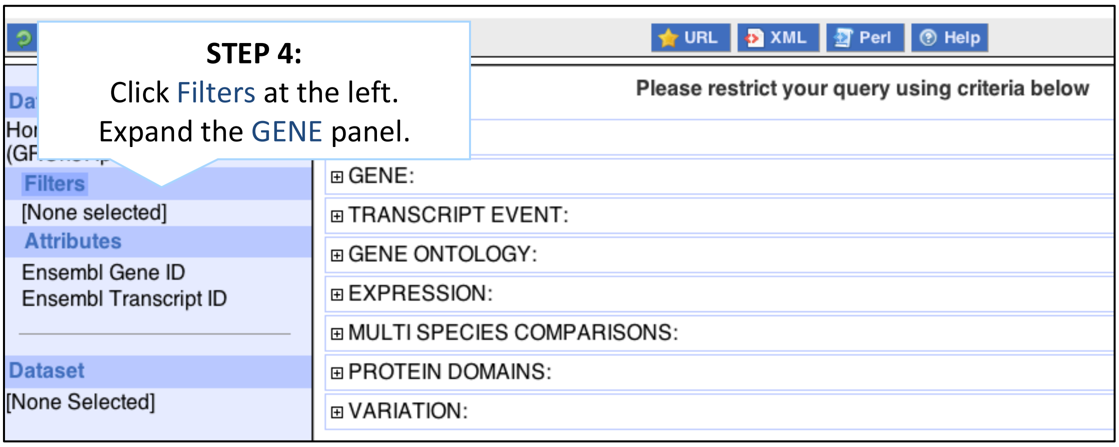
**STEP 2:**
Choose Ensembl Genes
as the primary database.



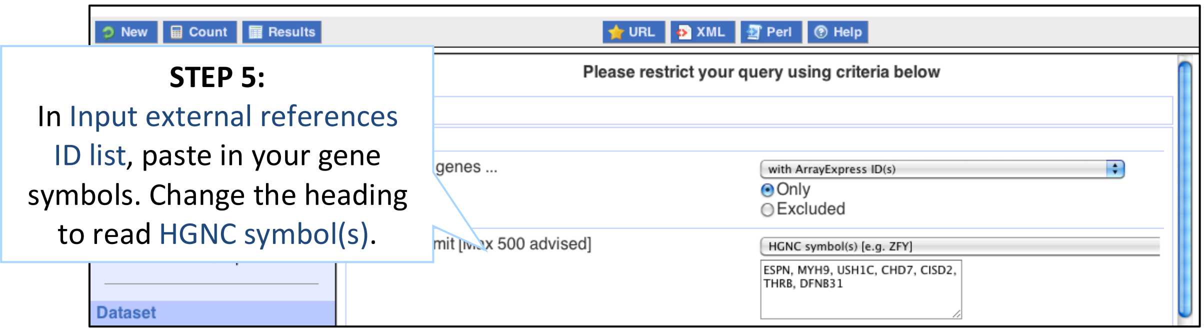
Ensembl Genes 71

Homo sapiens genes (GRCh37.p10)

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

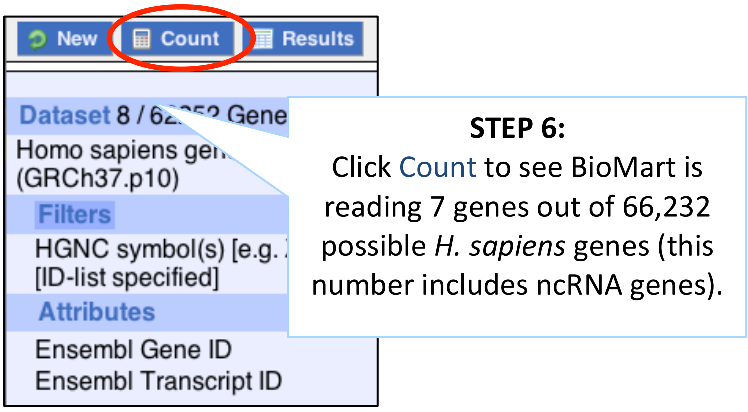


**STEP 4:**
Click Filters at the left.
Expand the GENE panel.

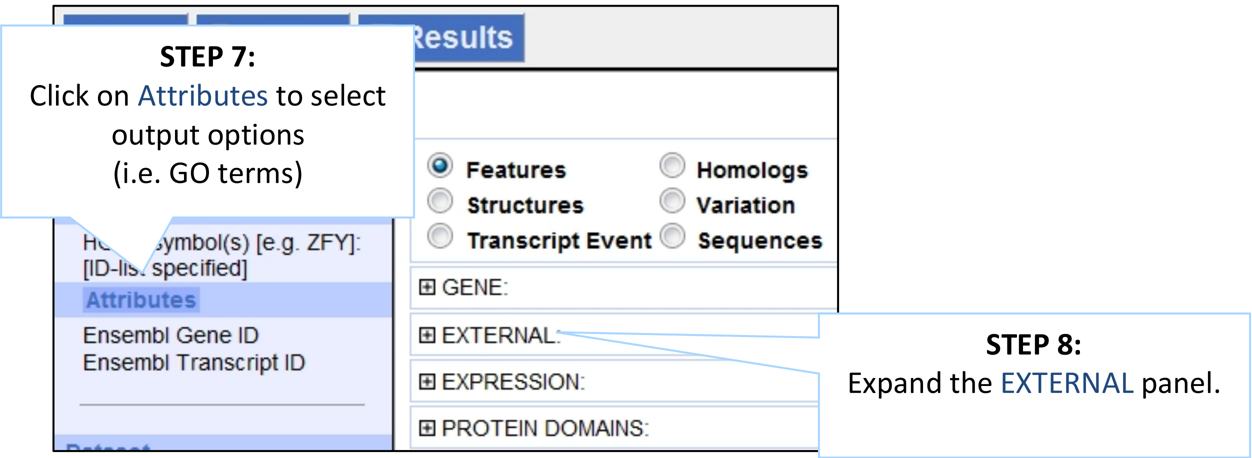


**STEP 5:**
In Input external references ID list, paste in your gene symbols. Change the heading to read HGNC symbol(s).

New
Count
Results
URL
XML
Perl
Help
Please restrict your query using criteria below
genes ...
with ArrayExpress ID(s)
● Only
○ Excluded
HGNC symbol(s) [e.g. ZFY]
ESPN, MYH9, USH1C, CHD7, CISD2, THRB, DFNB31
Dataset



**STEP 6:**
Click Count to see BioMart is reading 7 genes out of 66,232 possible *H. sapiens* genes (this number includes ncRNA genes).

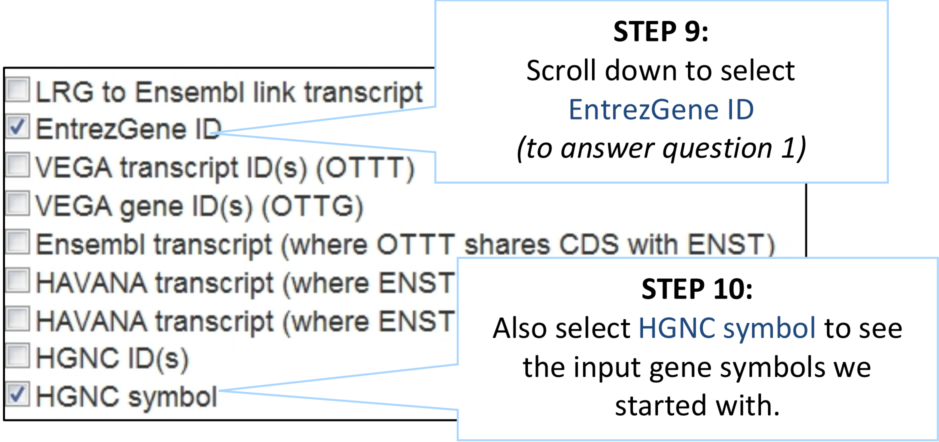


● Features ○ Homologs
○ Structures ○ Variation
○ Transcript Event ○ Sequences

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

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

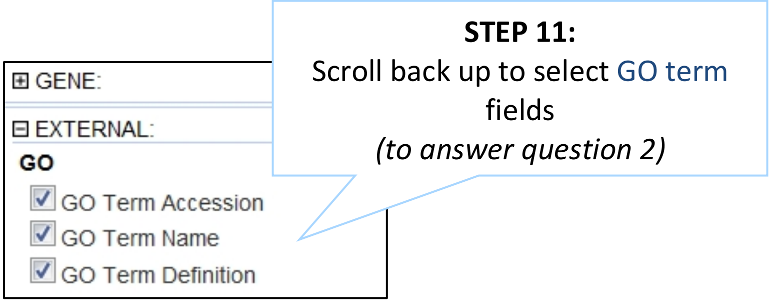
**STEP 8:**
Expand the EXTERNAL panel.



□ 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

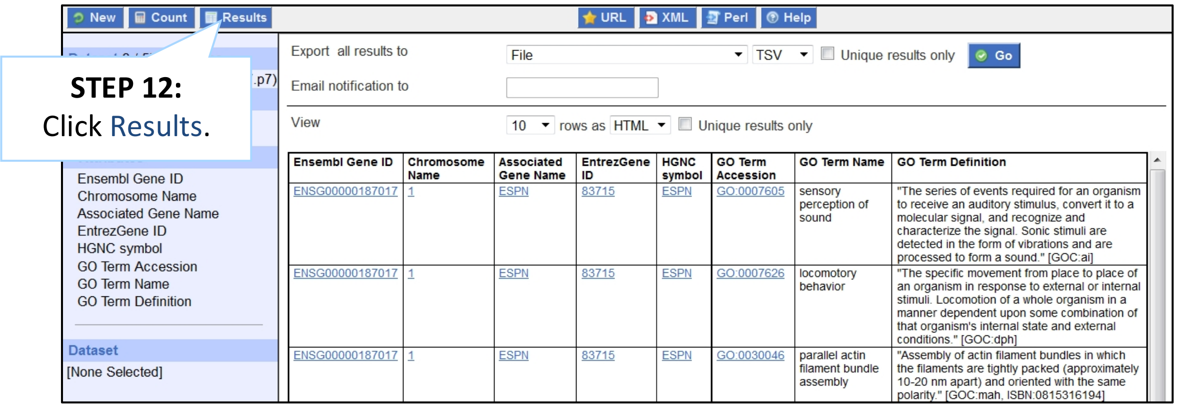
**STEP 9:**
Scroll down to select EntrezGene ID
*(to answer question 1)*

**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.

Attributes
Ensembl Gene ID
Chromosome Name
Associated Gene Name
EntrezGene ID
HGNC symbol
GO Term Accession
GO Term Name
GO Term Definition

Dataset
[None Selected]

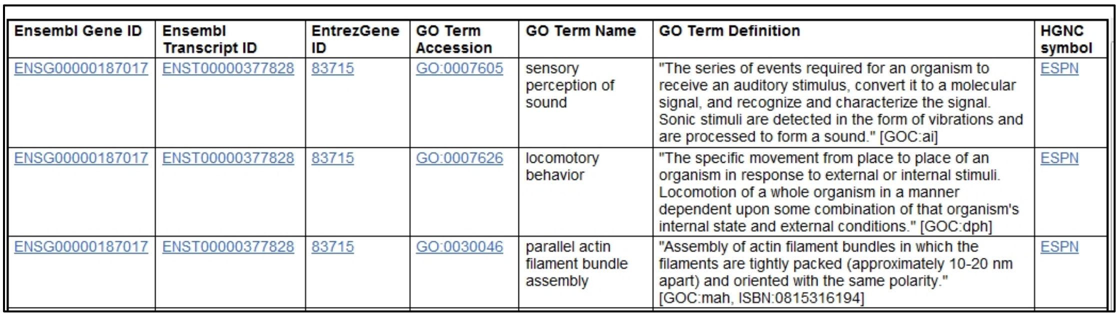
Export all results to File TSV □ Unique results only

Email notification to

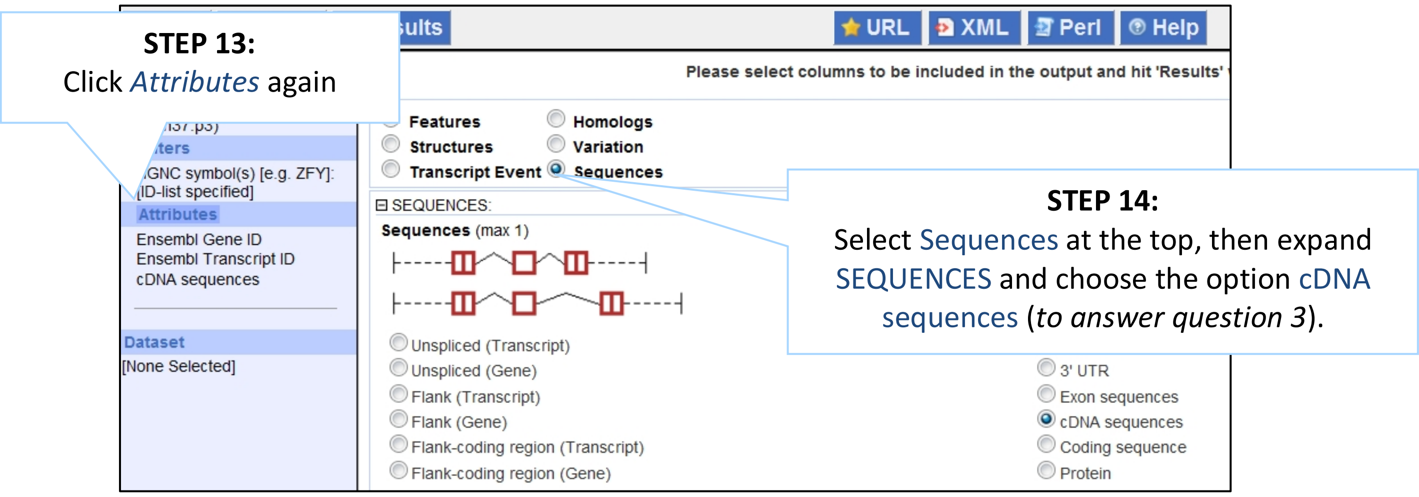
View 10 rows as HTML □ Unique results only

| Ensembl Gene ID | Chromosome Name | Associated Gene Name | EntrezGene ID | HGNC symbol | GO Term Accession | GO Term Name | GO Term Definition |
|---|---|---|---|---|---|---|---|
| ENSG00000187017 | 1 | ESPN | 83715 | ESPN | 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] |
| ENSG00000187017 | 1 | ESPN | 83715 | ESPN | 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] |
| ENSG00000187017 | 1 | ESPN | 83715 | ESPN | 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] |

Why are there multiple rows for one gene ID? Look at the first few rows of your results table.



| 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 |



**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*).

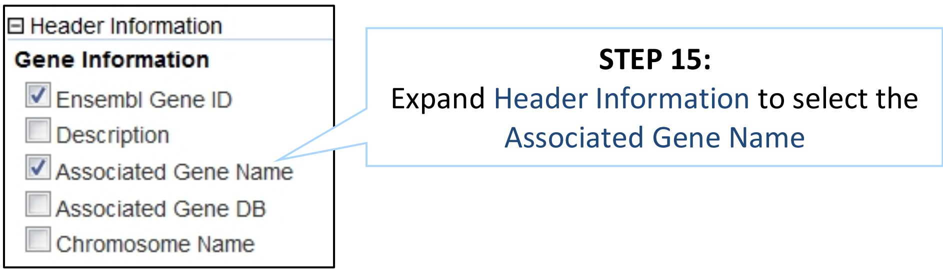
Please select columns to be included in the output and hit 'Results'

○ 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

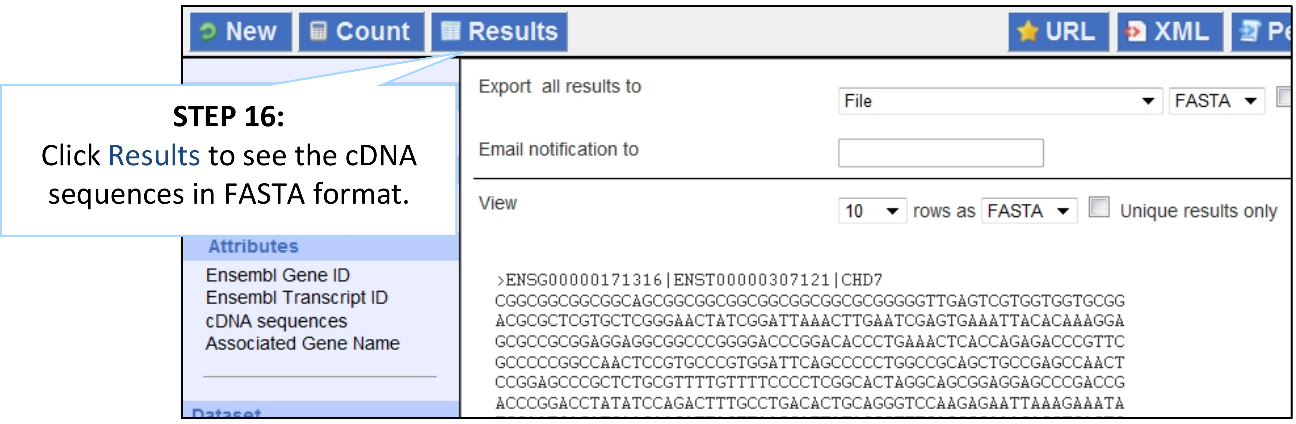


**STEP 15:**
Expand Header Information to select the Associated Gene Name

⊟ Header Information

Gene Information

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



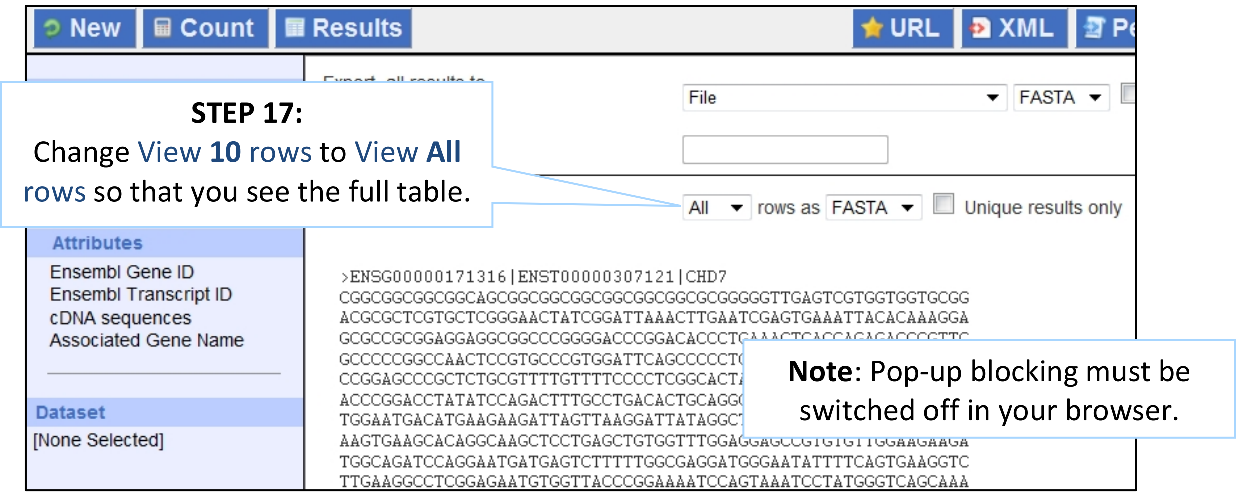
**STEP 16:**
Click Results to see the cDNA sequences in FASTA format.

Export all results to File FASTA

Email notification to

View 10 rows as FASTA □ Unique results only

```
>ENSG00000171316|ENST00000307121|CHD7
CGGCGGCGGCGGCAGCGGCGGCGGCGGCGGCGCGGGGGTTGAGTCGTGGTGGTGCGG
ACGCGCTCGTGCTCGGGAACTATCGGATTAAACTTGAATCGAGTGAAATTACACAAAGGA
GCGCCGCGGAGGAGGCGGCCCGGGGACCCGGACACCCTGAAACTCACCAGAGACCCGTTC
GCCCCCGGCCAACTCCGTGCCCGTGGATTCAGCCCCCTGGCCGCAGCTGCCGAGCCAACT
CCGGAGCCCGCTCTGCGTTTTGTTTTCCCCTCGGCACTAGGCAGCGGAGGAGCCCGACCG
ACCCGGACCTATATCCAGACTTTGCCTGACACTGCAGGGTCCAAGAGAATTAAAGAAATA
```



All rows as FASTA □ Unique results only

```
>ENSG00000171316|ENST00000307121|CHD7
CGGCGGCGGCGGCAGCGGCGGCGGCGGCGGCGGCGCGGGGGTTGAGTCGTGGTGGTGCGG
ACGCGCTCGTGCTCGGGAACTATCGGATTAAACTTGAATCGAGTGAAATTACACAAAGGA
GCGCCGCGGAGGAGGCGGCCCGGGGACCCGGACACCCTCAAACTCACCAGAGACCCCGTTC
GCCCCCGGCCAACTCCGTGCCCGTGGATTCAGCCCCCTC
CCGGAGCCCGCTCTGCGTTTTGTTTTCCCCTCGGCACTA
ACCCGGACCTATATCCAGACTTTGCCTGACACTGCAGGC
TGGAATGACATGAAGAAGATTAGTTAAGGATTATAGGCT
AAGTGAAGCACAGGCAAGCTCCTGAGCTGTGGTTTGGAGGAGCCGTGTGTTGGAAGAAGA
TGGCAGATCCAGGAATGATGAGTCTTTTTGGCGAGGATGGGAATATTTTCAGTGAAGGTC
TTGAAGGCCTCGGAGAATGTGGTTACCCGGAAAATCCAGTAAATCCTATGGGTCAGCAAA
```

**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 get these results using 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

## Exercise 1 – Export sequences in FASTA format from mouse

Retrieve the sequences of all mouse genes that are located on chromosome 17, that are protein coding and that encode for proteins containing transmembrane domains. Do a count after selection of each filter to check the number of genes remaining in your dataset. Export the results of the protein sequences (FASTA) as Compressed web file and get the results notified to you by email.

---

## Exercise 2 – Convert IDs of human genes

BioMart is a very handy tool when you want to map IDs between different databases. The following is a list of 19 accession numbers from the UniProtKB/Swiss-Prot database (http://www.uniprot.org/) of human proteins that are supposedly involved in the sensory perception of pain.

Q99608, P34913, P28482, P28223, Q96LB1, P10997, P01210, P25929, P17481, P43681, P29460, Q9HC23, P20366, Q9Y2W7, Q99572, Q99571, Q00535, P01138, P17787

Can you convert this list of UniProtKB/Swisse-Prot IDs into HGNC symbols? You may want to include as attributes Ensembl Gene IDs and 'Description' (under the field GENE).

---

## *ADDITIONAL EXERCISE: BIOMART*

***If you have finished the exercises above, you may want to do these extra ones***

### *Extra exercise 4: Convert UniProt IDs into Ensembl IDs for Arabidopsis proteins*

*BioMart is a very handy tool when you want to map between different databases.*

*The following is a list of IDs from the UniProtKB/Swiss-Prot database* *(http://www.uniprot.org/) of Arabidopsis thaliana proteins that are believed to be involved in flavonoid metabolism*

*http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0009812:*

*P42813, Q9LS08, Q9ZST4, Q9SYM2, P51102, Q9LPV9, Q9FE25, Q96323, Q9FKW3, P13114, P41088, Q9S818, Q96330, O22203, Q39224, O22264, Q9SD85, Q9LYT3, Q9FJA2, Q43128, P43254, O04153, Q43125, Q9S9P6, Q94C57, Q9LNE6, Q9FK25, Q9SYM5, Q9ZQ95*

*Go to plants.ensembl.org and click on the link BioMart at the top of the page. Generate a list that shows, to which Ensembl Gene IDs these UniProtKB/Swiss-Prot IDs map to. Also include the Gene name, Gene description and Pfam ID.*

---

### *Extra exercise 5: Retrieve a list of SNPs from the tomato genome (Solanum lycopersicum)*

*The region between coordinates 21,394,819 and 21,397,868 on chromosome 6 in tomato contains a gene involved in oxidation-reduction process (GO:0055114).*

*Can I use BioMart to retrieve all the SNPs that cause a change at the amino acid level of this gene (those SNPs are known as missense variants) including their IDs and possible alleles?*

---

### *Extra exercise 6: Find genes associated with array probes in Ensembl release 78*

*In the paper 'Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers' (Forrest et al. Environ Health Perspect. 2005 June; 113(6): 801–807) the effect of benzene exposure on peripheral blood mononuclear cell gene expression in a population of shoe factory workers with well-characterized occupational exposures was examined using microarrays. The microarray used was the Affymetrix U133A/B GeneChip (also called 'U133 plus 2').*
*The top probe sets that are up-regulated by benzene exposure were:*

*207630_s_at, 221840_at, 219228_at, 204924_at 227613_at, 223454_at, 228962_at, 214696_at, 210732_s_at, 212371_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*

*Use the archive version of BioMart on the previous release of Ensembl, i.e. release 78. The list of our archives can be found below:*

*http://www.ensembl.org/info/website/archives/index.html*

*Choose Ensembl 78 and then select BioMart.*

*a) Generate a list of the genes to which these probe sets map. Include the Ensembl Gene ID, name and description as well as the probe set name.*

*b) As a first step towards analysing them for possible regulatory features they have in common, retrieve the 250 bp upstream of the transcripts of these genes. Include the Ensembl Gene and Transcript ID, name and description in the sequence header.*

*c) In order to be able to study these human genes in mouse, generate a list of the human genes and their mouse orthologues. Include the Ensembl Gene ID for both the human and mouse genes and the homology type in your list.*

*d) Generate the same list as in (c), but now also include the name and description of both the human and mouse genes. As the name and description of the mouse genes are not available as attributes in the Ensembl human genes dataset, you have to add the Ensembl mouse genes as a second dataset.*

# ENSEMBL TOOLS: THE VEP

You may have identified in your lab, previously unknown variants (e.g. SNPs, SVs) and would like to find out which genes they map to and what are the effects of them on genes and proteins.

In Ensembl, we have got a tool that allows you to do just that: annotate your own SNPs/SVs in any given genome, even if Ensembl has not got variation data for it (e.g. Bacteria).

This tool is called Variant Effect Predictor or VEP.

The VEP can be run on a web interface, via a Perl script or via our REST API. More information can be found below:

www.ensembl.org/vep

Let's try some exercises using the web interface of the VEP based on the talk and demo given by the instructor.

**Exercise 1 – Using VEP to predict the consequence of SNPs on the previous human assembly, GRCh37. Go to grch37.ensembl.org**

An analysis of 5,000 patients from a Taipei cohort has identified few variants associated with lung cancer:

chr 15, genomic coordinate 78889339, alleles G/A, forward strand
chr 22, genomic coordinate 30332586, alleles T/C, forward strand
chr 6, genomic coordinate 31721033, alleles G/A, forward strand
chr 5, genomic coordinate 1260624, alleles G/A, forward strand
chr 17, genomic coordinate 63554591, alleles A/G, forward strand
chr 5, genomic coordinate 1254510, alleles C/T, forward strand

Use the VEP to answer the following:

a) Which genes and transcripts do these variants map to?

b) What are the consequence terms for these variants?

c) Are there deleterious variants according to the SIFT/PolyPhen predictions? Are these two predictions always in agreement with each other?

---

## Exercise 2 – The VEP tool and variants on the bread wheat genome.

An analysis of 5,000 individuals from two different populations of bread wheat (*T. aestivum*) has identified thousands of polymorphic loci. See a list of a few of them below:

chr 2D, genomic coordinate 89551917, alleles G/A, forward strand
chr 2D, genomic coordinate148408765, alleles G/T, forward strand
chr 3D, genomic coordinate113574123, alleles C/A, forward strand
chr 3D, genomic coordinate 93827883, alleles G/A, forward strand
chr 3B, genomic coordinate 727928129, alleles C/T, forward strand
chr 3B, genomic coordinate 736734474, alleles C/T, forward strand
chr 6A, genomic coordinate 196872409, alleles T/G, forward strand
chr 6A, genomic coordinate 196153918, alleles A/G, forward strand
chr 6A, genomic coordinate 196774882, alleles G/C, forward strand

Can you use the VEP tool to answer the following?

a) Are any of these variants known in the public domain? Can you list a few of the IDs of these existing variants?

b) Which genes and transcripts do these variants map to?

c) Which consequence types can be found for these variants? Do any of them cause a change at the amino acid level?

---

## Exercise 3 – VEP in Bacterial genomes

Find the genome for *Bacteroides fragilis* 638R and launch the VEP tool. Use the VEP to predict the effects of a 7 bp deletion of TCTACAA on the supercontig FQ312004 at the position 258140-258146.

# QUICK GUIDE TO DATABASES AND PROJECTS

Here is a list of databases and projects you will come across in these exercises. 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, GenBank and DDBJ –** They 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 all of them.

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

**NCBI Entrez Gene –** NCBI's gene collection.

**NCBI RefSeq –** NCBI's collection of 'reference sequences'. It includes genomic DNA, mRNAs and proteins. NM represents 'Known mRNA' (e.g. NM_005476) and NP (e.g. NP_005467) is 'Known proteins'.

**UniProtKB –** the "Protein knowledgebase" is a comprehensive set of protein sequences. It is divided into two parts: Swiss-Prot and TrEMBL

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

**UniProt TrEMBL –** the automatically annotated, unreviewed set of protein sequences (EMBL-Bank translated). Varying quality.**VEGA –** Vertebrate Genome Annotation database providing a selection of manually-curated genes, transcripts, and proteins (human, mouse, zebrafish, gorilla, wallaby, pig, chimpanzee and dog).

**HAVANA –** Human and Vertebrate Analysis and annotation group at the Wellcome Trust Sanger Institute. It's the main contributor to the manual annotation presented in VEGA.

## GENE NAMES

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

**ZFIN –** The Zebrafish Model Organism Database
(Zebrafish International Resource Center).

**MGI –** Mouse Genome Informatics

**For other species you may want to try/use WikiGenes**

## ANNOTATION OF PROTEINS

**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.

**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

**dbSNP –** A collection of sequence polymorphisms; mainly single nucleotide polymorphisms, along with small insertion and deletions.

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