

Bioinformatics: Life Sciences on Your Computer

Ensembl

This is supplemental reading for the video lecture on Ensembl.

Ensembl

Ensembl (<http://www.ensembl.org>) contains sequence info from a genomic perspective. The database contains known and predicted genes with information on exons and splice variants. The current version is Release 75, which was rolled out in February 2014. There are species in Ensembl, mostly chordates, but some other model organisms, including yeast and *C. elegans*.

Other species can be found in [Ensembl Genomes](#):

Ensembl Metazoa: <http://metazoa.ensembl.org/index.html>

Ensembl Protists: <http://protists.ensembl.org/index.html>

Ensembl Bacteria: (includes archaea) <http://bacteria.ensembl.org/index.html>

Ensembl Plants: <http://plants.ensembl.org/index.html>

Ensembl Fungi: <http://fungi.ensembl.org/index.html>

Ensembl project

Ensembl is a joint project between the EMBL – European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute to produce accurate, automatic genome annotation, originally focused on selected eukaryotic genomes. The goal is to integrate various biological data in an open source manner and to develop open source software for analysis. The [Ensembl Project](#) dates back to 1999. A lot of the annotation is automated, but there is a staff that builds gene sets.

Many video tutorials are available. There is a three-hour course for beginners, but I recommend the YouTube video called [The Ensembl Genome Browser](#).

Species homepage

There is a search window on Ensembl front page, that will search all species. You could include the organism in the search itself (e.g. human EPO gene). However, if you go to the species home page, you would not need to put the name of the organism in the search box (e.g. EPO gene) because it will just search that species.

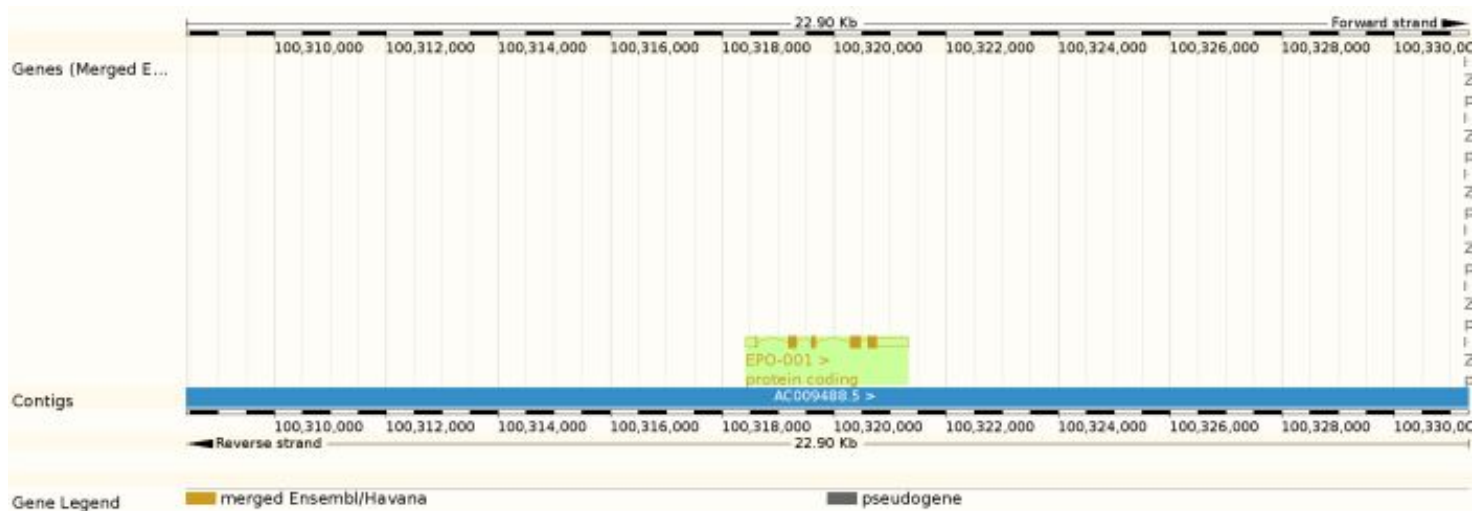
So, type in 'EPO' on human page. Click "Gene" then click "Human." Click the first hit, ENSG00000130427. Ensembl gene IDs begin with "ENSG" for human and there are a lot of leading zeroes. A three-letter code inserted for non-human species. Mouse genes begin with "ENSMUSG" and horse genes begin with "ENSECAG."

Gene Summary page

When you open a gene record, you access its Gene Summary page. At the top, you can find a description, location, transcripts. The description refers to the official HGNC symbol. There is now a link that reads "Show transcript table" that must be clicked to see the transcript table. Click that link. You will see a table with known transcripts for that genes and links to the protein produced by that transcript (if any). Human transcripts begin with ENST and human proteins begin with ENSP.

Below the transcript table is the actual gene summary, including gene name, synonyms and other information. The graphic (below) shows the gene model. It's above the contig so this gene is on the

forward strand. Empty boxes refer to the UTR and filled boxes represent the coding regions. The connecting lines are introns.



Gene status is as follows:

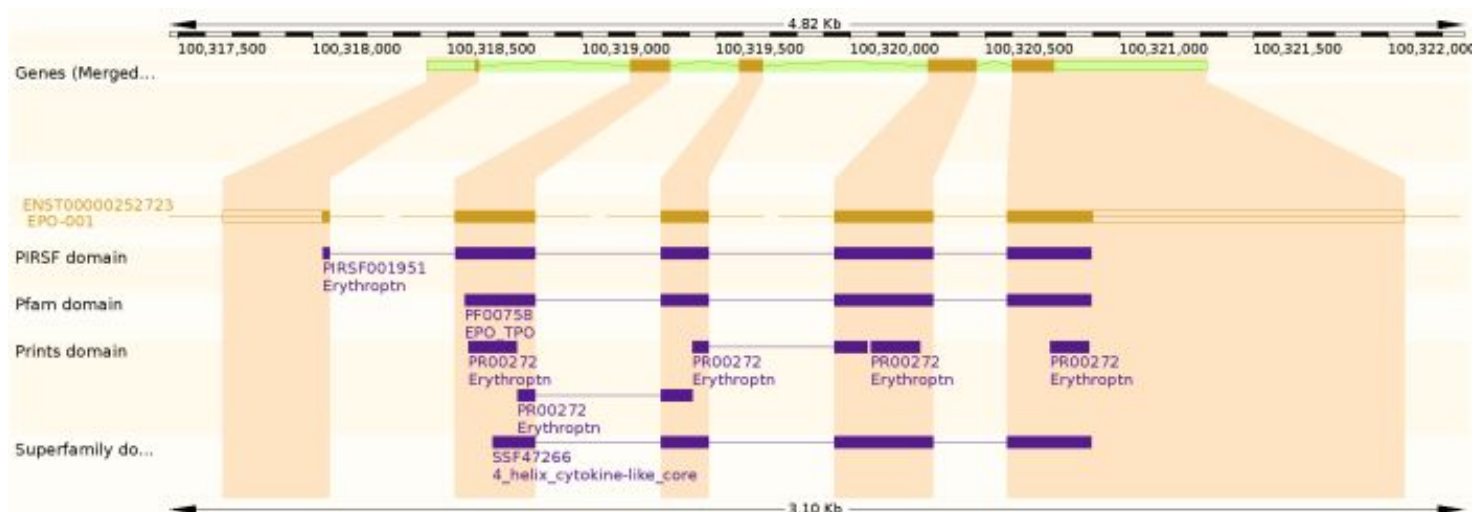
Known gene: Transcript match in same species

Novel gene: Only sequence matches for an alternate species

Merged gene: Has at least one known merged transcript between automated annotation and a curated annotation

Splice variants page

All transcripts for a gene displayed. The image below represents an "easy" gene in that there is only one known transcript. The colors represent the known/novel/merged status and the tan highlights trace exons. If there were more than one transcript, exons across transcripts may be compared. The exons are expanded at bottom to show protein motifs and matches to databases like Pfam, PRINTS and Superfamily.

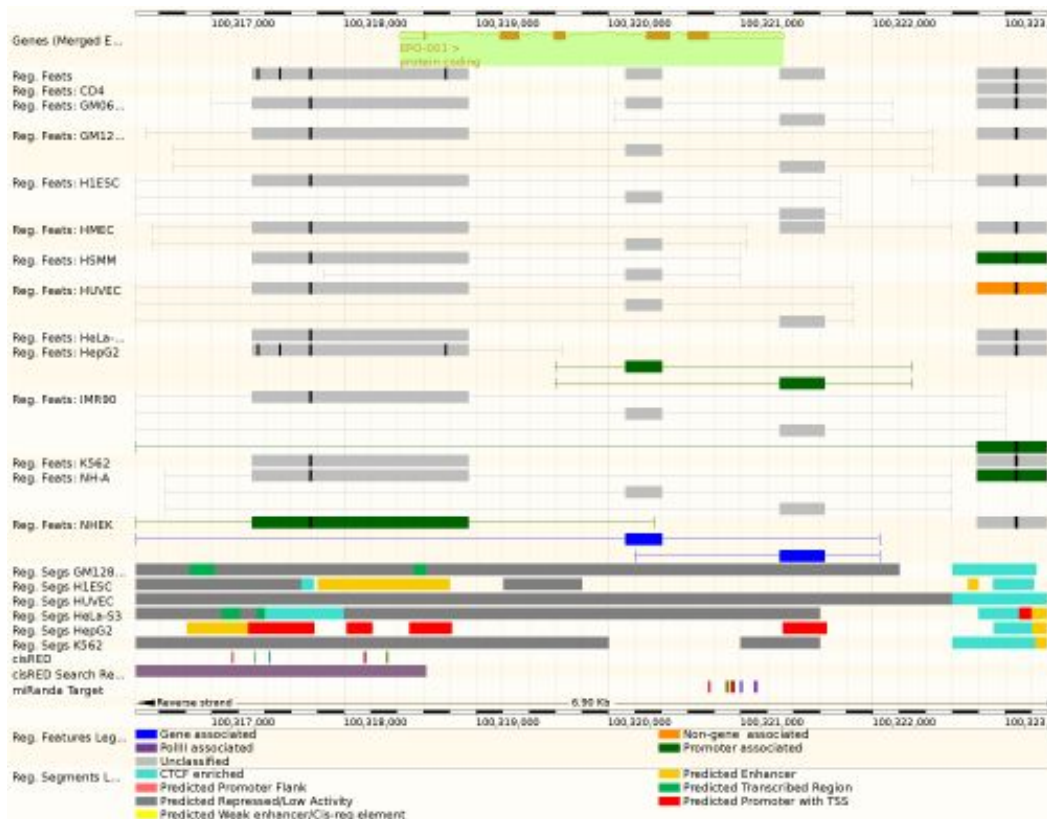


Discussion topic: Access this page with a gene that has more than one transcript. How is that information different?

Regulation page

The regulation page shows species-specific regulatory feature information. The regulatory features from

Ensembl Regulatory Build and include the graphical locations of sequences of regulatory elements. Those can include miRNA targets from miRanda.



Orthologues Page

The orthologue page shows homologues inferred from genetrees. These have been determined for each gene using all species in Ensembl. Orthologues are defined as genes for which most common ancestor node is a speciation event. Only one copy is found in each species if the Type is 1-to-1. Between-species paralogues are not shown--they are in a separate Paralogues tab. The links for orthologues point to gene records in all available species for which there is an orthologue.

Variation Table page

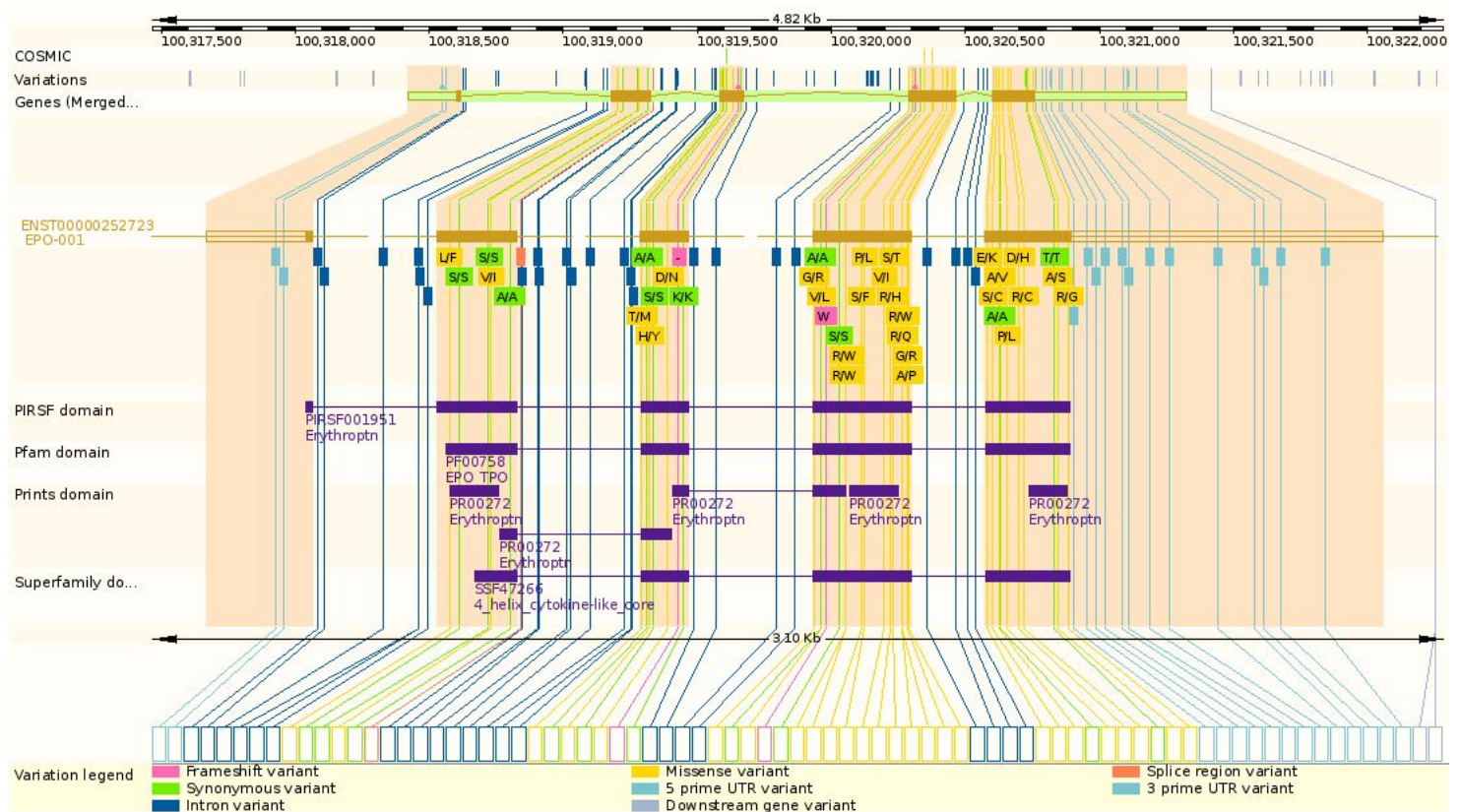
The variation table shows SNPs with a summary of variation data in table. It breaks down variants by type, Missense variant (non-synonymous), 3' UTR, etc. The table includes the SNP identifier, the nucleotide change and the amino acid change if it's a coding SNP.

On the variation image page, all variations for a gene are shown graphically and displayed as vertical lines, They are color-coded according to position of variation in transcript.

- Light blue - UTR
- Dark blue - Intronic
- Green - Synonymous Coding
- Yellow - Non-synonymous Coding
- Red - Splice site
- Pink - Frameshift coding

In the image below, note that the green does not change the amino acid but the yellow does show the

change.



Genomic Alignments

If you click Genomic alignments on the left menu, that allows for aligning genomic sequence to orthologues. Pick the desired orthologues in the pull-down menu. It shows the alignment. Red nucleotides represent exons.

Discussion topic: Try the transcript page (ENST) and protein page (ENSP). What information is there?

Discussion topic: Try adding a track through Configure this page. How hard is it?