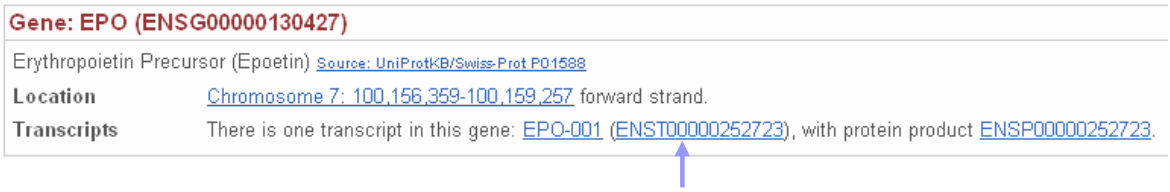
Module 3 – Genes and Transcripts

You will learn about

- Ensembl IDs and gene structure
- Homologues
- DAS sources and GO terms
- More links from the Gene page

Searching for a gene in Ensembl provides a link to the **gene** page. If you'd like to follow along on the live-site, search for 'human EPO gene' and click on the Ensembl gene identifier (**ENSG0000130427**)- this will link you to the 'GeneView' page for the Erythropoeitin precursor. Please note that since this tutorial was constructed updates may have been made, so to emulate this tutorial exactly, please use the archived site for version 52 (as used in this example) which can be found here:
http://Dec2008.archive.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000130427

The **Gene** page provides links to a wealth of information. At the top of the page we find the ID, name and chromosomal location of the gene. The name for this gene is assigned by the **HGNC** (HUGO Gene Nomenclature Committee) and the description is from the corresponding gene entry in UniProt/Swiss-Prot. Furthermore, this gene is a member of the human **CCDS** set. These are consensus coding sequences agreed upon by Ensembl, Havana, NCBI and UCSC.



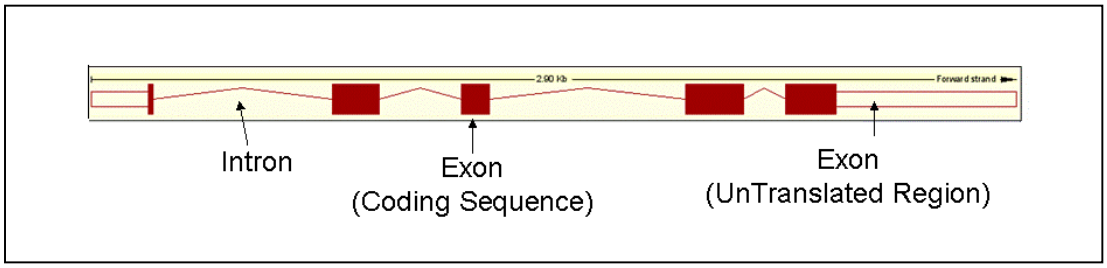
All transcripts (splice isoforms) annotated by Ensembl or Havana are shown. For this gene, only one transcript, **ENST00000252723**, is annotated (see arrow in figure above). The Ensembl ID follows the convention of all identifiers from the Ensembl '**Genebuild**' procedure. These identifiers start with ENS (for Ensembl) followed by 'G' for gene, 'T' for transcript, 'E' for Exon, 'P' for peptide and 'F' for family. The following 11-digit number is stable (i.e. it is not changed upon a new release without a corresponding change in gene sequence or annotation). The following convention is for human Ensembl genes:

ENS**G**… Gene
ENS**T**… Transcript
ENS**E**… Exon
ENS**P**… Protein

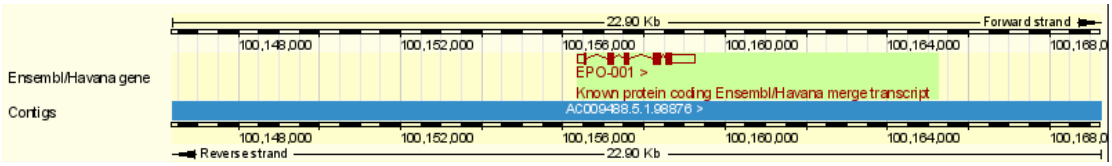
For species other than human, a three-letter sequence follows 'ENS'. For example, a mouse (*Mus musculus*) gene will be named 'ENSMUSG…' following this convention.

ENSMUSG… Mouse Gene
ENSMUST… Mouse Transcript
ENSMUSE… Mouse Exon
ENSMUSP… Mouse Protein

Further down the **gene** page, the gene structure is drawn for each transcript. The gene structure is as follows: exons are represented by boxes, and lines connecting the boxes are introns. Filled boxes reflect coding sequence, unfilled boxes reflect UTRs (UnTranslated Regions).

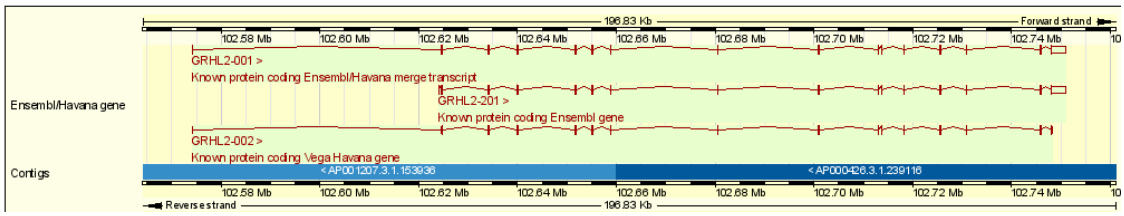


An Ensembl gene may have multiple transcripts reflecting alternative splicing, in this case the EPO gene only shows one transcript in Ensembl.



The transcript is displayed on top of the blue bar (the genomic assembly) as it is on the forward strand of the chromosome.

In another case, multiple transcripts are shown for the human **GRHL2** gene. In this example, three transcripts of varying lengths are found above the genome. Transcripts that have overlapping coding sequence are grouped under the same gene name and ID.



Click on any transcript for links back to the Gene page, or for more information about the transcript, such as the corresponding protein. These transcripts are clickable in every page of Ensembl.

The EPO transcript shown above and the GRHL2 transcripts in the second figure are listed as 'known protein coding' genes. When a transcript matches back to a known identifier in a public database such as UniProtKB or NCBI RefSeq, it is assigned a 'known status'. A 'novel'

transcript, which we do not see in these examples, would not match back to a record in UniProt or NCBI RefSeq for the same species (rather, a different species).

A 'Vega/Havana' transcript has been imported from the manual curators at the Wellcome Trust Sanger Institute. An Ensembl/Havana merge indicates the exact same coding sequence was determined by the Ensembl annotation pipeline and the Havana manual curators. These are high quality transcripts, as two projects relying on mRNA and protein in scientific databases independently find the same transcript.

In these examples, EPO has an Ensembl/Havana merge. This is also indicated in the 'Prediction method' above the transcript diagram. GRHL2 has a merged transcript, an Ensembl transcript and a Vega/Havana transcript. All three are displayed in Ensembl pages.

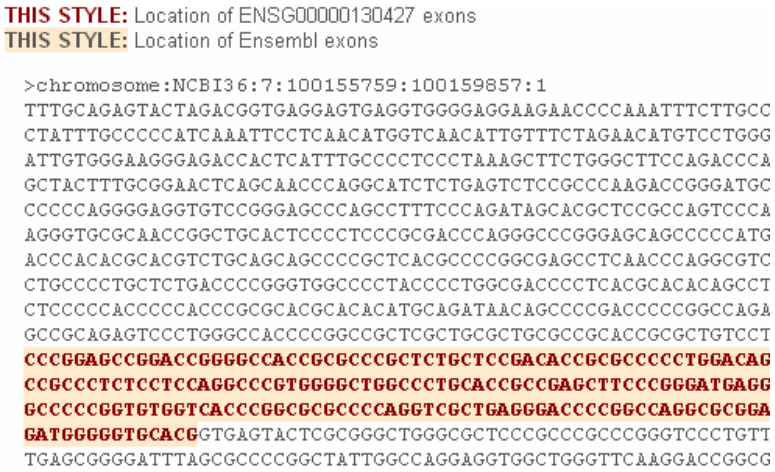
Click the 'help' icon on the Gene summary page for more about gene annotation in Ensembl.

Or, read more about the gene annotation process here:

http://www.ensembl.org/info/docs/genebuild/index.html

Let's explore some of the links from the **gene summary** page. Click on '**Supporting evidence**' to see mRNA and proteins underlying the Ensembl transcript. For EPO, the support comes from a CCDS sequence and also an NCBI RefSeq protein (NP_000790). Click on these IDs to learn more about the underlying evidence. Or, click on the features under 'Exon Support' at the right of the page, for more proteins and mRNAs that align to part of this Ensembl transcript. This is a good place to start if you have the question, 'where did this isoform come from?'

Now back to more links from **gene summary**. Click on '**Sequence**' at the left.



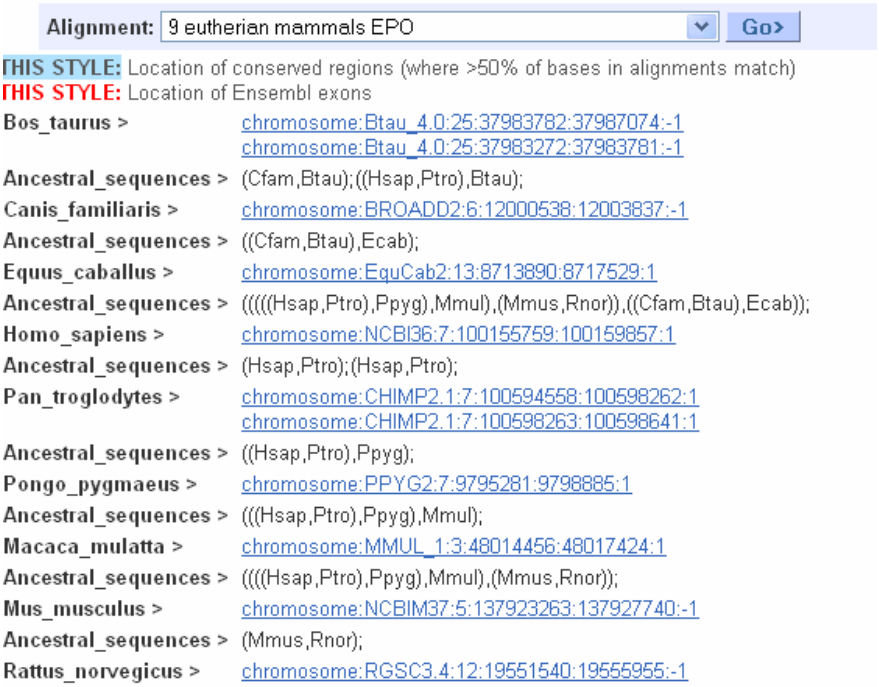
The genomic sequence in the region of the EPO gene is shown, with any exons highlighted. The sequence is in FASTA format. Let's look more closely at the FASTA header:

>chromosome:NCBI36:7:100155759:100159857:1

The symbol **>** always starts the FASTA header. Then, we see the name of the genomic assembly (**NCBI 36** is the most recent human genome sequence from the Human Genome Project.) The chromosome number is then indicated (**7**) and the base pairs (**100155759-100159857**). The final **1** shows the sequence is the forward strand of the chromosome. A **-1** would indicate the reverse strand.

Most pages in Ensembl can be configured. For example, click on **'Configure this page'** at the left to turn on line numbering. Choosing **'Relative to this sequence'** and clicking **'SAVE and close'** at the very top right hand corner of the panel will reload the Sequence view with line numbers (starting at 1 from the beginning of the sequence displayed). Note, you can also change the amount of flanking sequence shown and turn on variations along the sequence using 'Configure this page' in this view.

The next link we will explore from the **gene summary** page is the '**Genomic alignments**' view. This view portrays, on a sequence level, whole genome alignments between species in Ensembl. Most of these are pairwise alignments between two species, however there are a small subset of multi-species alignments available. 'Select an alignment' at the top. Let's select '9 eutherian mammals'. EPO does not actually refer to our gene, but the method used to determine the whole genome alignments (developed in the Ensembl project. Click on the Help on this page for more.)

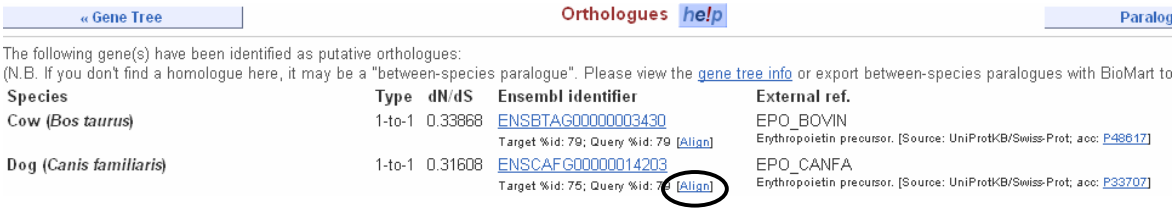


Shown here are the chromosomes and regions aligned for each of the species in this 9-way alignment. Also, ancestral sequences are predicted.

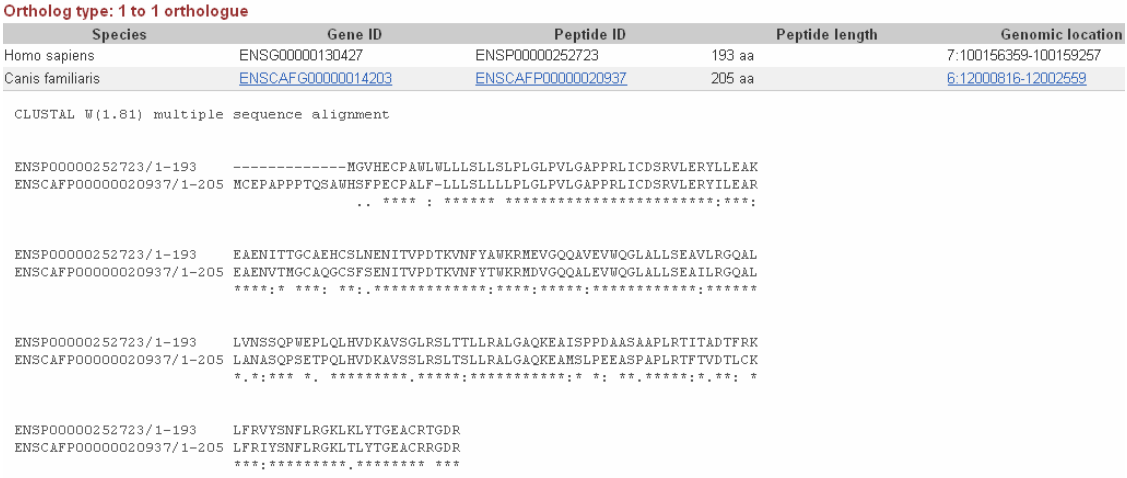
Homologues are predicted using a **phylogenetic gene tree**, which is based on the longest translation of every gene for all species in Ensembl, and can be viewed by clicking on the **'Gene tree'** link at the left hand side of the page.

The gene tree is used to examine homology, and is the result of multiple sequence alignments using all the species in Ensembl. This gene tree shows speciation events leading to **orthologues**, and duplication events leading to **paralogues**.

Click the '**Orthologues**' link from the **gene summary** page. Clicking on the 'Align' link for the *Canis familiaris* EPO gene (circled below) shows the sequence alignment between the human EPO gene and the dog.



This leads to a page like the following screenshot showing the alignment of the two proteins:



The alignment format can be customised. The peptide alignment in clustal format as shown above is the default, but this can be changed to DNA alignments in FASTA, Pfam, PSI formats and more. Use **'Configure this page'** to alter the format.

FInally, click on **'External Data'** for the '**DAS** Sources' available. DAS stands for 'Distributed Annotation System' and allows information held in external databases to be shown alongside the Ensembl annotation. For example, click on **'Configure this page'** and select **'ArrayExpress Wharehouse'** to see if any transcription profiles are available for the EPO gene. The DAS sources available will vary between releases and species as these are external databases.

We've explored many of the links accessible from the gene page. Let's look closer at the EPO transcript. Click on '**ENST00000252723**' at the top of any view. Alternatively, click on the transcript tab.

To see the exon and intron sequences in a colour-coded display, click the '**Exons**' link at the left. For more sequence options, see the **cDNA** and **Protein** links.

Click on '**General identifiers**' at the left of the page. These are IDs in other databases that match to the EPO transcript sequence.

This Ensembl/Havana merge transcript entry corresponds to the following database identifiers:

HGNC Symbol:	EPO [view all locations]
CCDS:	CCDS5705.1 [view all locations]
Human Protein Atlas:	CAB010336 [view all locations] CAB010336 [view all locations]
WikiGene:	EPO [view all locations]
UniProtKB/Swiss-Prot:	EPO_HUMAN [Target %id: 100; Query %id: 100] [align] [view all locations]
RefSeq peptide:	NP_000790.2 [Target %id: 100; Query %id: 100] [align] [view all locations]
RefSeq DNA:	NM_000799 [align] [view all locations]
EntrezGene:	EPO [view all locations]

Links are shown to the databases housing the matching sequences. Here we find the HGNC and Swiss-Prot records from which the EPO name and description were taken. This page is updated with every release (every two months) so keep an eye on it for new sequence matches!

Scroll down to see matches to **PDB** structures, 3-D structures of proteins in the Protein Data Bank.

The link below '**General Identifiers**' (the '**Oligo probes**' link) reveals a section where accession numbers from various **probes and sequences** from array platforms are listed for the gene. The accession numbers correspond to a sequence found on an array that possesses a matching sequence containing all or part of the Ensembl transcript.

This Ensembl/Havana merge transcript entry corresponds to the following database identifiers:

Agilent CGH:	A_14_P113914 [Target %id: 3; Query %id: 100] [view all locations]
Agilent Probe:	A_23_P145664 [Target %id: 4; Query %id: 100] [view all locations] A_23_P145669 [Target %id: 4; Query %id: 100] [view all locations]
Affymx Microarray Focus:	207257_at [view all locations]
Affymx Microarray HCG110:	1023_at [view all locations]
Affymx Microarray HuGeneFL:	X02158_rna1_at [view all locations]
Affymx Microarray U133:	207257_at [view all locations] 217254_s_at [view all locations] Hs.2303.1.S1_3p_a_at [view all locations] g4503588_3p_at [view all locations]
Affymx Microarray U95:	1023_at [view all locations]
GE Healthcare/Amersham Codelink WGA:	GE79554 [Target %id: 2; Query %id: 100] [view all locations]
Illumina V1:	GI_4503588-S [Target %id: 3; Query %id: 98] [view all locations]
Illumina V2:	ILMN_6125 [Target %id: 3; Query %id: 100] [view all locations]

Below '**Oligo probes'** is a section showing information about Gene Ontology (**GO**) terms assigned to the Ensembl entry by the Gene Ontology Consortium www.geneontology.org. GO terms are associated with Ensembl genes via the UniProt/Swiss-Prot, RefSeq and UniProt/TrEMBL entries to which the Ensembl gene has been mapped. Clicking on the classes shows more information about the function.

The following GO terms have been mapped to this entry via UniProt and/or RefSeq:

GO Accession	Evidence	Go Term
GO:0001666	IEA	response to hypoxia
GO:0005128	IEA	erythropoietin receptor binding
GO:0005179	IEA	hormone activity
GO:0005515	IPI	protein binding
GO:0005576	IEA	extracellular region
GO:0005615	TAS	extracellular space
GO:0006357	IEA	regulation of transcription from RNA polymerase II promoter

GO uses two or three-letter 'Evidence Codes' to show how the term is associated to a gene. The following codes are shown above and described below:

IC: Inferred by Curator
IDA: Inferred from Direct Assay
IEA: Inferred from Electronic Annotation
IEP: Inferred from Expression Pattern
IGI: Inferred from Genetic Interaction
IMP: Inferred from Mutant Phenotype
IPI: Inferred from Physical Interaction
ISS: Inferred from Sequence or Structural Similarity
NAS: Non-traceable Author Statement
ND: No biological Data available
RCA: inferred from Reviewed Computational Analysis
TAS: Traceable Author Statement
NR: Not Recorded

Remember, to get a description of these codes within the Ensembl browser, you could have clicked on the 'Help' button (circled in the figure below) at the top of the **Gene ontology** page.

The left hand navigation panel can be used to investigate other pages such as sequence **variations** within a population, or across strains or breeds. Also, have a look at protein domains. The location tab will be dealt with in **module 4**.

Glossary

CCDS-consensus coding sequence set A set of coding sequences that have been agreed on between Ensembl, VEGA, UCSC and NCBI (RefSeq). http://www.ncbi.nlm.nih.gov/CCDS/

DAS- Distributed Annotation System A platform that allows external data to be portrayed on the browser. www.biodas.org

Genebuild The process of determining the Ensembl geneset using the annotation pipeline.

GO – The Gene Ontology A project that assigns genes into functional classifications or gene ontologies. http://www.geneontology.org/

HGNC symbol The gene name assigned by the HUGO Gene Nomenclature Committee (for human). http://www.genenames.org

Orthologue *(In Ensembl)* A homologous (related) gene in which the most recent divergence was a speciation event. Determined using phylogenetic comparisons (trees) across all Ensembl species.

Paralogue *(In Ensembl)* A homologous (related) gene in which the most recent divergence was a duplication event. Determined using phylogenetic comparisons (trees) across all Ensembl species.

PDB- Protein Data Bank www.rcsb.org/pdb A database of experimentally-determined protein and nucleic acid structures.

Supporting Evidence *(In Ensembl)* The mRNA and protein that form the basis of the Ensembl transcript.

Variations *(In Ensembl)* Population variations in human and other species are mainly obtained from NCBI dbSNP. Strain variations are calculated for rat and mouse from resequencing projects.

What to do next

To read more about gene annotation in Ensembl, see our article here:

http://www.ensembl.org/info/docs/genebuild/index.html

For more about comparative genomics in Ensembl, go here:

http://www.ensembl.org/info/docs/compara/index.html

To explore a chromosomal region, go to **module 4**. For a more in depth walk-through of the Ensembl browser, see the web-site walk-through in our course booklet here:

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

Watch the related video!

www.ensembl.org/info/website/tutorials/index.html

'Supporting Evidence for a Gene'