

# Access to genes and genomes with **Ensembl**



**Course Manual**

March 2009 (v53)



## I) Introduction

Ensembl is one of the world's primary resources for genomic research, a resource through which scientists can access the human genome as well as the genomes of other model organisms. Because of the complexity of the genome and the many different ways in which scientists want to use it, Ensembl has to provide many levels of access with a high degree of flexibility. Through the Ensembl website a wet-lab researcher with a simple web browser can for example perform BLAST searches against chromosomal DNA, download a genomic sequence or search for all members of a given protein family. But Ensembl is also an all-round software and database system that can be installed locally to serve the needs of a genomic centre or a bioinformatics division in a pharmaceutical company enabling complex data mining of the genome or large-scale sequence annotation.

### The need for automatic annotation

Recent years have seen the release of huge amounts of sequence data from genome sequencing centres (figure 1). However, this raw sequence data is most valuable to the laboratory biologist when provided along with quality annotation of the genomic sequence.

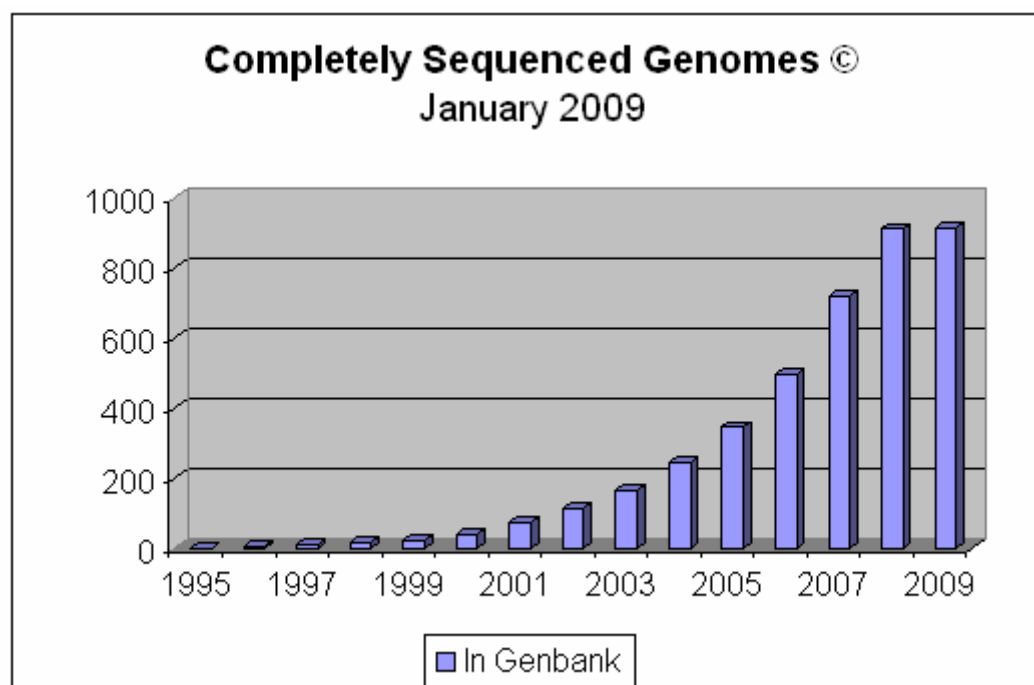


Figure 1. Completely sequenced genomes as of January, 2009 (figure taken from <http://www.genomesonline.org>).

This information can be the starting point for planning experiments, interpreting Single Nucleotide Polymorphisms, inferring the function of gene products, predicting regulatory sites for gene expression and so on. The currently agreed 'gold standard' for the annotation of eukaryotic genomes is annotation made by a human being. This so-called "manual annotation" is based on information derived from sequence homology searches, the results of various *ab initio* gene prediction methods and literature searches. Annotation of large genomes (such as mouse and human) that meet this standard is slow and labour intensive, taking large teams of annotators years to complete. As a result, the annotation can almost never be entirely up-to-date and free of inconsistencies (as the annotation process usually begins before the sequencing process is complete). Hence, an automated annotation system is desirable since it is a relatively rapid process that allows frequent updates to accommodate new data. To meet this need, we produced the Ensembl annotation system by observing how annotators build gene structures and condensing this process into a set of rules.

### **The start of Ensembl**

Ensembl's genesis was in response to the acceleration of the public effort to sequence the human genome in 1999. At that point it was clear that if annotation of the draft sequence was to be available in a timely fashion it would have to be automatically generated and that new software systems would be needed to handle genome data sets that were much larger, much more fragmented and much more rapidly changing than anything previous dealt with.

Ensembl was conceived in three parts: as a scalable way of storing and retrieving genomic data; as a web site for genome display; and as an automatic annotation method based around a set of heuristics. It was initially written for the draft human genome, which was sequenced clone-by-clone but has also been successfully used for whole genome shotgun assemblies. The storage and display parts of Ensembl are used for all the genomes currently present in Ensembl, while the automatic gene annotation has been run for most of the genomes with the exception of Takifugu, Tetraodon, Fruitfly, *C. elegans* and Yeast.

Over the past few years Ensembl has grown into a large scale enterprise, with substantial computing resources enabling it to process and provide live database access to currently more than 25 different genomes (figure 2) and a bimonthly update frequency to its website. It has a large community of users in both industry and academia, using it as a base for their individual organisation's experimental and computational genome based investigations, some of which maintain their own local installations.

Ensembl is a collaboration between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute, both located on the Wellcome Trust Genome Campus in Hinxton, Cambridge, UK. Ensembl is funded principally by the Wellcome Trust, with additional funding from the European

Molecular Biology Laboratory (EMBL), the National Institutes of Health – National Institute of Allergy and Infectious Disease (NIH-NIAID) and the Biotechnology and Biological Sciences Research Council (BBSRC).

### **The Ensembl software and database system**

As a software/database system Ensembl can be best described as a hybrid of a scripting programming language (Perl) and a relational database (MySQL, pronounced “My Ess Que Ell”).

Ensembl Perl software inherits from a tradition of biological object-design developed through BioPerl (<http://www.bioperl.org/>). This means that developers at Ensembl aimed at creating reusable pieces of software that would faithfully describe biological entities such as gene, transcript, protein, genomic clone or chromosome. Rules of usage and design of Ensembl and BioPerl objects can be best learned while using them, browsing their code and through a bit of trial-and-error. There is a comprehensive BioPerl tutorial available at the BioPerl website.

The Ensembl database is based on a relational database called MySQL. SQL in MySQL stands for ‘Structured Query Language’, a universal database programming language shared by many relational databases. Because MySQL is available free of charge for non-commercial developers, every academic centre can install its own local copy of MySQL (MySQL server) and download Ensembl data from the Ensembl ftp site. Simple queries of the database can be handled using the SQL language (see appendix), but for complex queries demanded by most biological analyses the Ensembl MySQL server is best accessed using Ensembl Perl objects.

### **The Ensembl annotation pipeline**

The Ensembl analysis and annotation pipeline is based on a rule set of heuristics that a human annotator would use. All Ensembl gene predictions are based on experimental evidence, which is imported via manually curated UniProt/Swiss-Prot, partially manually curated NCBI RefSeq and automatically annotated UniProt/TrEMBL records. Untranslated regions (UTRs) are annotated to the extent supported by EMBL mRNA records. As there is no guarantee that UTR sequences in EMBL records are complete there is similarly no guarantee that the Ensembl genome analysis and annotation pipeline has enough biological evidence to predict complete UTR regions. For a limited number of species regulatory regions are annotated, but this annotation isn’t very extensive yet as the set of well-characterised promoters is still small and there is currently no algorithm yielding reliable results on a genomic scale.

### **The Ensembl website**

Ensembl provides access to genomic information with a number of visualisation tools. The Ensembl website gives you the possibility to directly download data, whether it is the DNA sequence of a genomic contig you are trying to identify novel genes in, or positions of SNPs in a gene you are

working on. The key Ensembl web pages are covered in the web-site walk-through. An updated version of the website is released bimonthly. Old versions are accessible on the 'Archive!' website, dating back two years. Apart from that the 'Pre!' website provides displays of genomes that are still in the process of being annotated. There is also an ftp site to download large amounts of data from the Ensembl database, as well as the data-mining tool BioMart, that allows rapid retrieval of information from the databases, and BLAST/BLAT sequence searching and alignment.

### Further reading

Hubbard, T.J.P. *et al.*

#### **Ensembl 2009**

Nucleic Acids Res. Jan 2009 37: D690-697 (*database issue*)

Vilella, A.J. *et al.*

#### **EnsemblCompara GeneTrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates.**

Genome Res. 2009 Feb 19(2):327-35

Smedley, D. *et al.*

#### **BioMart – biological queries made easy**

BMC Genomics 2009 Jan 14;10:22

Flicek, P. *et al.*

#### **Ensembl 2008**

Nucleic Acids Res. Jan 2008; 36: D707 - D714

Spudich, G., Fernández-Suárez, X. M., and Birney, E.

#### **Genome Browsing with Ensembl: a practical overview**

Brief Funct Genomic Proteomic, 2007 Sept; 6: 202-219

Fernández Suárez X. M. and Schuster M.

#### **Using the Ensembl Genome Server to Browse Genomic Sequence Data.**

*Current Protocols in Bioinformatics*, UNIT 1.15, January 2007.

Hubbard, T.J.P. *et al.*

#### **Ensembl 2007**

Nucleic Acids Res. 2007 (*Database Issue*)

Birney, E. *et al.*<sup>1</sup>

#### **An Overview of Ensembl.**

Genome Research 14(5): 925-928 (2004)

Ashurst, J. L. *et al.*

#### **The Vertebrate Genome Annotation (Vega) database.**

Nucl. Acids Res. 33:D459-D465 (2005)

---

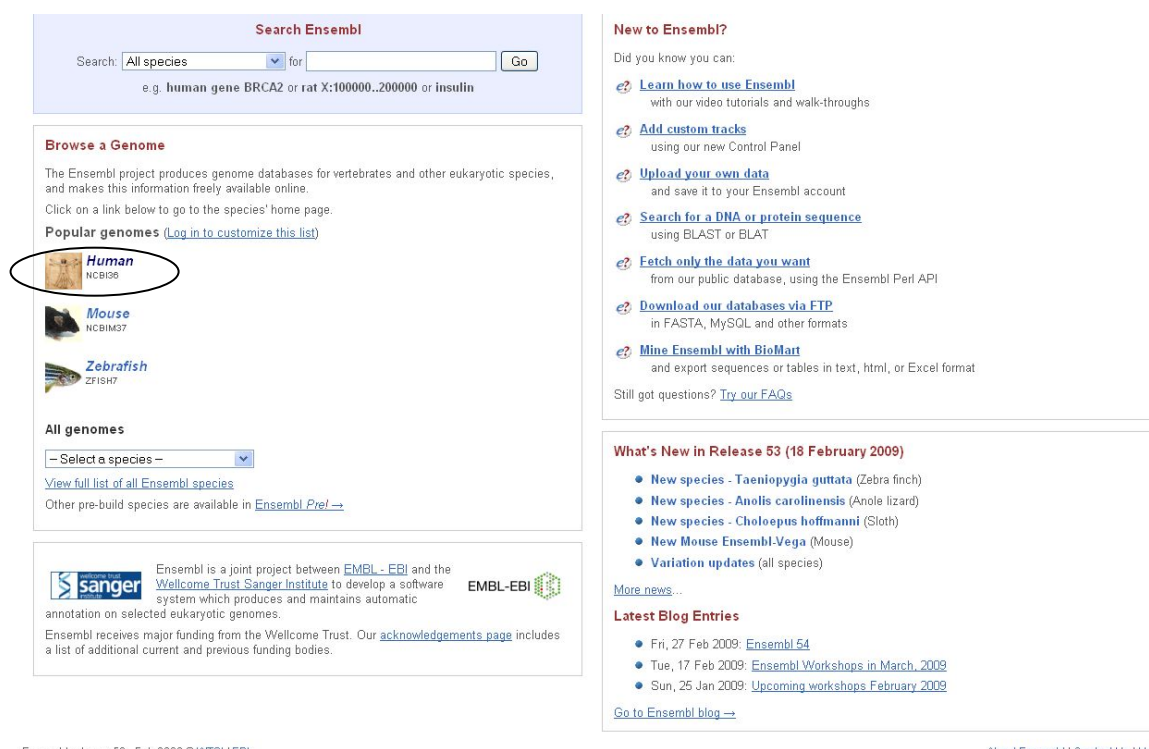
<sup>1</sup> This paper was part of the May 2004 issue of *Genome Research* which included an Ensembl special covering detailed aspects of the Ensembl web site, the underlying scalable database system for storing genome sequence and annotation information, as well as the automated genome analysis and annotation pipeline

## II) WALKING THROUGH THE WEBSITE

The instructor will guide you through the website using the **Interleukin-2 precursor (or IL2) gene**. The following points will be addressed:

- **The Gene Summary tab and gene-related links:**
  - Are there splice variants?
  - Can I view the genomic sequence with variations?
  - Find orthologues and paralogues
- **The Transcript tab and related links:**
  - What is the protein sequence?
  - What matching proteins and mRNAs are found in other databases?
  - Gene Ontology
- **The Location tab and related links:**
  - What's the conservation track?
  - How do I zoom in and change the gene focus.
  - Un-stacking a track (e.g. human cDNAs)
  - Adding a track (i.e. variations)
- **Exporting a sequence and running BLAT/BLAST**

Start by going to **[www.ensembl.org](http://www.ensembl.org)**



The screenshot shows the Ensembl website homepage. At the top, there is a search bar with the text "Search Ensembl" and a "Go" button. Below the search bar, there is a section titled "Browse a Genome" which includes a description of the Ensembl project and a list of popular genomes: Human (NCBI36), Mouse (NCBIM37), and Zebrafish (ZFISH7). The "Human" entry is circled in red. Below this, there is a section for "All genomes" with a dropdown menu to select a species. On the right side, there is a section titled "New to Ensembl?" with several links for new users, such as "Learn how to use Ensembl", "Add custom tracks", and "Upload your own data". At the bottom, there is a section titled "What's New in Release 53 (18 February 2009)" with a list of new species and variation updates.

Click on 'Human', or the picture circled above, which brings us to the species home page.



Home > Human

Search Ensembl Human

Search for:  Go  
 e.g. gene BRCA2 or 6:133017695-133161157 or muscular dystrophy

**Description** Assembly and Genebuild >

**Assembly**

This release is based on the NCBI 36 assembly of the [human genome](#) [November 2005]. The data consists of a reference assembly of the complete genome plus the Celera WGS and a number of alternative assemblies of individual haplotypic chromosomes or regions. [Full list of assemblies](#) →

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- [Nature 431, 931 - 945 \(21 October 2004\)](#)
- [WT Sanger Institute Press Release](#)

**Annotation**

Since release 38 (April 2006) the gene annotation presented has been a combined Ensembl-Havana, geneset which incorporates more than 18,000 full-length protein-coding transcripts annotated by the Havana team with the Ensembl automatic gene build. The human genome sequence is now considered sufficiently stable that since 2004 the major genome browsers have come together to produce a common set of identifiers where CDS annotations of transcripts can be agreed and these identifiers are also shown.

- [More information about the CCDS project.](#)

The [ENCODE](#) (ENCYClopedia Of DNA Elements) project aims to find functional elements in the human genome.

- [More information about the ENCODE resources](#) at Ensembl.

**Vega** Additional manual annotation of this genome can be found in [Vega](#)

Ensembl release 51 - Nov 2008 © WTSI / EBI  
[Permanent link](#) - [View in archive site](#)

[About Ensembl](#) | [Contact Us](#) | [Help](#)

Type 'gene IL2' into the search bar circled above and click the 'Go' button.

Home > Human

Genome

Search Ensembl

Species (1)  
 Homo sapiens (6)  
 Gene (6)

Feature type (6)  
 Gene (6)  
 Homo sapiens (6)

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

**Ensembl text search**

IL2 corporate/tree:"Top/Species/Homo sapiens/G" Search

Your query matched 6 entries in the search database

**Vega protein\_coding Gene: OTTHUMG00000133075 (HGNC Symbol: IL2)** [Region in detail]

Vega protein\_coding gene OTTHUMG00000133075 has 2 transcripts: OTTHUMT00000256715, OTTHUMT000001613867, OTTHUMT000001613868, OTTHUMT000001613869, OTTHUMT000001613870, O interleukin 2

The gene has the following external identifiers mapped to it:  
 CCDS: CCDS3726, CCDS3726.1  
 Ensembl Human Gene: ENSG00000109471  
 Ensembl transcript sharing CDS with Havana: ENST00000226730  
 EntrezGene: 3558, **IL2**  
 GO: GO:0005615, GO:0019209, GO:0006916, GO:0007267, GO:0030307, GO:0030101, GO:0007155, GO:0042104  
 HGNC Symbol: **IL2**, BPT  
 MIM gene: 147680  
 RefSeq DNA: NM\_000546  
 Sequence Publications  
 UniProtKB/Swiss-Prot:  
 Vega gene: OTTHUMG00000133075  
 Vega transcript: OTTHUMT00000256715, OTTHUMT000001613867, OTTHUMT000001613868, OTTHUMT000001613869, OTTHUMT000001613870  
 Vega protein: 49698, OTTHUMP00000164090

Source: e52; Species: Homo sapiens; Gene; Feature type: Gene; Homo sapiens;

**Ensembl protein\_coding Gene: ENSG00000109471 (HGNC (curated): IL2)** [Region in detail]

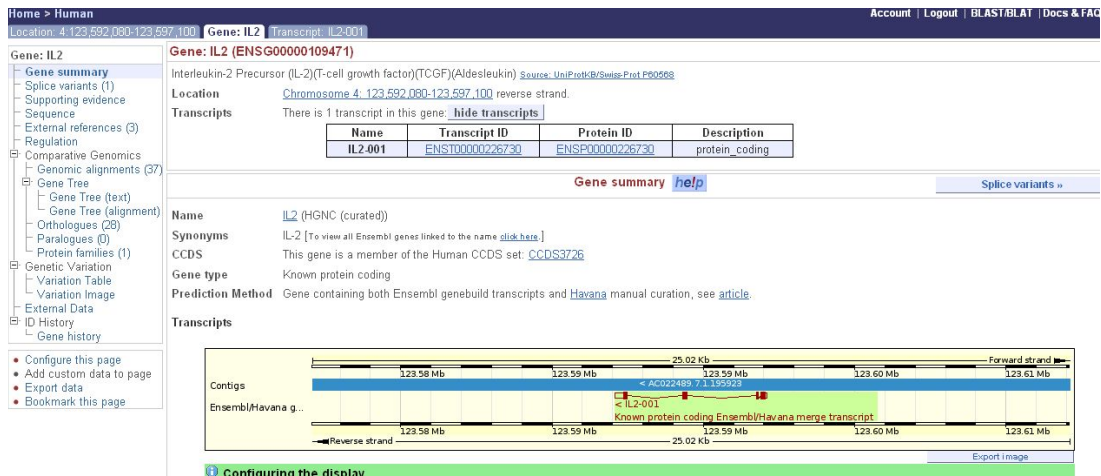
Ensembl protein\_coding gene ENSG00000109471 has 1 transcript: ENST00000226730, associated pep ENSE00000935280, ENSE00001138256, ENSE00001293064  
 Interleukin-2 Precursor (IL-2)(T-cell growth factor)(TCGF)(Aldesleukin) [Source:UniProtKB/Swiss-Prot;Ac  
 The gene has the following external identifiers mapped to it:  
 Affymx Microarray Focus: 207849\_at  
 Affymx Microarray HCG110: 1538\_s\_at  
 Affymx Microarray HuGeneFL: S77835\_s\_at, X00695\_s\_at  
 Affymx Microarray U133: 207849\_at, g10835148\_3p\_at  
 Affymx Microarray U95: 34021\_at, 1538\_s\_at

**Click ENSG00000109471**

Look through the search results for IL2, the gene symbol. Select the Ensembl gene (i.e. ENSG00000109471), rather than the manually curated Vega gene. (We will see Vega transcripts in the display, along with



Ensembl transcripts). The following 'Gene' tab will open:



Home > Human  
 Location: 4:123,592,080-123,597,100 Gene: IL2 Transcript: IL2-001 Account | Logout | BLAST/BLAT | Docs & FAQ

Gene: IL2  
 Gene: IL2 (ENSG00000109471)  
 Interleukin-2 Precursor (IL-2)(T-cell growth factor)(TCGF)(Aldesleukin) Sources: UniProtKB/Swiss-Prot P00568

Location: Chromosome 4: 123,592,080-123,597,100 reverse strand

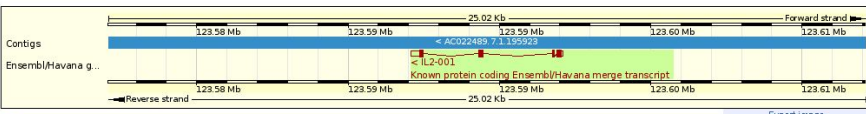
Transcripts: There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
IL2-001	ENST00000226730	ENSP00000226730	protein_coding

Gene summary [help](#) [Splice variants »](#)

Name: IL2 (HGNC (curated))  
 Synonyms: IL-2 [To view all Ensembl genes linked to the name [click here](#).]  
 CCDS: This gene is a member of the Human CCDS set: [CCDS3726](#)  
 Gene type: Known protein coding  
 Prediction Method: Gene containing both Ensembl genebuild transcripts and [Havana](#) manual curation, see [article](#).

Transcripts



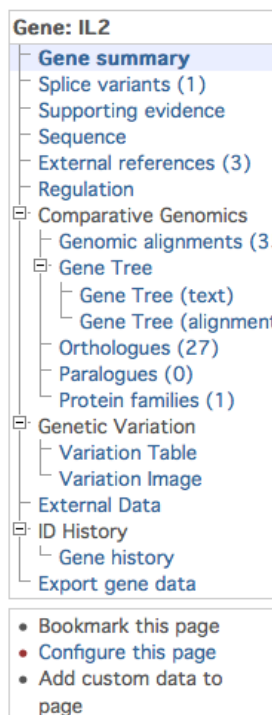
Contigs: 123.58 Mb, 123.59 Mb, 123.60 Mb, 123.61 Mb

Ensembl/Havana g...: IL2-001, Known protein coding Ensembl/Havana merge transcript

25.02 Kb Forward strand  
 Reverse strand

[Configure this page](#)  
[Add custom data to page](#)  
[Export data](#)  
[Bookmark this page](#)

Let's walk through some of the links in the left hand navigation column.



Gene: IL2

- Gene summary
- Splice variants (1)
- Supporting evidence
- Sequence
- External references (3)
- Regulation
- Comparative Genomics
  - Genomic alignments (3)
  - Gene Tree
    - Gene Tree (text)
    - Gene Tree (alignment)
  - Orthologues (27)
  - Paralogues (0)
  - Protein families (1)
- Genetic Variation
  - Variation Table
  - Variation Image
- External Data
- ID History
  - Gene history
- Export gene data

[Bookmark this page](#)  
[Configure this page](#)  
[Add custom data to page](#)

Click on **Supporting evidence** first, which will show biological sequence records (mRNA and protein) that have been used for the annotation of transcripts of a particular gene.

**Gene: IL2**  
 Interleukin-2 Precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin) [Source: UniProtKB/Swiss-Prot P60568](#)

**Location** [Chromosome 4: 123,592,080-123,597,100](#) reverse strand.

**Transcripts** There is one transcript in this gene: [IL2-001 \(ENST00000226730\)](#), with protein product [ENSP00000226730](#).

Transcript	CDS support	UTR support	Exon support
<a href="#">ENST00000226730</a>	<a href="#">CCDS3726 [align]</a> <a href="#">NP_000577.2 [align]</a>	<a href="#">NM_000586.3 [align]</a>	<a href="#">20 features</a>

Ensembl release S1 - Nov 2008 © [WTSI](#) / [EBI](#)  
[Permanent link](#) - [View in archive site](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)

How can we view the genomic sequence? Click **Sequence** at the left.

**THIS STYLE:** Location of ENSG00000109471 exons

**THIS STYLE:** Location of Ensembl exons

```
>chromosome:NCBI36:4:123591480:123597700:-1
TTAAATTAATAATAGCGTTAAACAGTACCTCAAGCTCAATAAGCATTTTAAGTATTCTAAT
CTTAGTATTCTCTAGCTGACATGTAAGAAGCAATCTATCTTATTGTATGCAATTAGCTC
ATTGTGTGGATAAAAAAGTAAAAACCATTCTGAAAACAGGAAACCAATACACTTCCATTTT
ATCAACAAATCTAAACATTTATTTTCATCTGTTTACTCTTGCTCTTGTCACCACACAA
TATGCTATTTCACATGTTTCAGTGTAGTTTTATGACAAAGAAAATTTCTGAGTTACTTTTG
TATCCCCACCCCTTAAAGAAAAGGAGGAAAACTGTTTCATACAGAAAGGCGTTAATTGCA
TGAATTAGAGCTATCACCTAAGTGTGGGCTAATGTAACAAAAGAGGATTTACCTACATC
CATTCAAGTCAGTCTTTGGGGTTTTAAAGAAAATTCCAAAGAGTCATCAGAAAGGAAAAAT
GAAAGGTAATGTTTTTCAGACAGGTAAAGTCTTTGAAAAATATGTGTAATATGTAATAACAT
TTTGACACCCCCATAATAATTTTTCCAGAAATTAACAGTATAAAATTCATCTCTTTGTTCAAG
AGTTCCCTATCACTCTCTTTAATCACTACTCACAGTAACTCAACTCCTGCCACAATGTA
CAGGATGCAACTCCTGTCTTGCACTAAGTCTTGCACTTGTCAAAAAGTGGCC
TACTTCAAGTTCTACAAAGAAAAACAAGCTACAAGTGGAGCATTACTGCTGGATTACA
GATGATTTTGAATGGAATTAATGTAAGTATATTTCCCTTTCTACTAAAATATTACATTT
AGTAACTAGCTGGAGATCATTTCCTTAATAACAATGCATTATACCTTTCTTAGAATTACAA
GAATCCAAACTCACAGGATGCTCACATTTAAGTTTTACATGCCAAGAAAGGTAAGTAC
AATATTTTATGTTCAATTTCTGTTTTAATAAAAATCAAAGTAATATGAAAAATTTGCACAG
ATGGGACTAATAGCAGCTCATCTGAGGTAAGAGTAACCTTAATTTGTTTTTTGAAAAAC
```

Upstream  
sequence

Exon  
sequence

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

Gene: IL2 (ENSG00000109471)

Configure page Custom Data Your account **SAVE and close**

Configure view  
 Configure

Configuration for: "Marked up gene sequence"

5' Flanking sequence: 600

3' Flanking sequence: 600

Number of base pairs per row: 60 bps

Additional exons to display: Core exons

Orientation of additional exons: Display exons in both orientations

Show variations: Yes and show links

Line numbering: Relative to this sequence

DAS sources

- ArrayExpress Warehouse  
Gene expression profile thumbnails from the ArrayExpress warehouse [Homepage]
- cbs\_func  
CBS Protein function and structure predictions [Homepage]
- cbs\_ptm  
CBS Post-translational modification site predictions [Homepage]
- cbs\_sort  
CBS Protein sorting predictions [Homepage]
- GAD  
Genetic Association Database - association of diseases to human Entrez genes. [Homepage]

Once you have selected changes (in this example, we display variations and show line numbers) click **Save and Close** at the top right (circled in red, above).

- THIS STYLE:** Location of ENSG00000109471 exons
- THIS STYLE:** Location of Ensembl exons
- THIS STYLE:** Location of SNPs
- THIS STYLE:** Location of deletions

```
>chromosome:NCBI36:4:123591480:123597700:-1
1  TTAAATTAATAAGCGTTAAACAATACCTCAAGCTCAATAAGCATTTTAAGTATTCTAAT 60 24:G/A;
61  CTTAGTATTCTCTAGCTGACATGTAAGAAGCAATCTATCTTATTGTATGCAATTAGCTC 120
121 ATTGGTGGATAAAAAGGTAACCATTCTSAACAGGAACCAATACACTTCCTGTTT 180 151:G/C; 180:T/A;
181 ATCAACAAATCAAACATTTATTCTTTTCATCTGTTTACTCTTGCTCTTGTCACCA 240 205:TTTT/-;
241 TATGCTATTCACATGTTCAAGTGTAGTTTTAGACAAAAGAAAATTTCTGAGTTACTTTG 300 271:T/G; 286:TC/-;
301 TATCTCCACCCCTTAAAGAAAGGAGGAAAACACTGTTTCATACAAAGGCGTTAATTGCA 360 305:C/T; 345:G/A;
361 TGAATTAGAGCTATCACCTAAGTGTGGGCTAATGTAACAAAAGAGGGATTTCACCTACATC 420
421 CATTCACTCAGTCTTTGGGGGTTTAAAGAAAATCCAAAAGAGTCATCAGAAGAGGAAAAAT 480
481 GAAGGTAATGTTTTTTCAGACGGTAAAGTCTTTGAAAATATGTGTAATATGTAACACAT 540 502:A/T;
541 TTTGACACCCCCATAAATATTTTCCAGAATTAACAGTATAAATTGCATCTCTTGTTCAG 600
601 AGTTCCTATCACTCTCTTAATCACTACTCACAGTAACTCAACTCCTGCCAATGTA 660
661 CAGGATGCAACTCCTGCTTTGCACTAAGCTTTGCACTTTGCACTTTGCAAAACAGTGCA 720
721 TACTTCAAGTTCTCAAAGAAAACAGCTACAACCTGGAGCATTACKCTGGATTACA 780 768:T/G; 769:G/T;
781 GATGATTTGAATGGAATTAATGTAAGTATATTCTTCTTACTAAAATATTACATTT 840
841 AGTAATCTAGCTGGAGATCATTCTTAATACAATGCATTATACTTTCTTAGAATTACAA 900 870:A/T;
901 GAATCCAAACTCACAGGATGCTCAATTTAAGTTTTACATGCCCAAGGAGTAAAGTAC 960
961 AATATTTATGTTAATTTCTGTTTAAATAAAAATTCAAAAGTAATATGAAAATTTGACAG 1020 968:TA/-; 974:C/T; 1016:C/A
1021 ATGGGACTAATAGCAGCTCATCTGAGGTAAGAGTAACTTTAATTTGTTTTTTTGAAC 1080
1081 CCAAGTTTGATAATGAAGCCTCTATTAAAACAGTTTTACCTATATTTTTAATATATATT 1140
1141 GTGTGTTGGTGGGGTGGGAAGAAAACATAAAAATAAATATTCTCACTTTATCGATAAGAC 1200
1201 AATCTAAACAAAATGTTCAATTTATGGTTTCAATTTAAAATGTAACACTCTAAAATATT 1260
1261 TGATTATGCTATTTAGTATGTAATATACAAAATCTATTTCAAAGGAGCCCACTTTTA 1320 1285:A/T; 1319:A/-;
1321 1380
```

Variations in the sequence are highlighted in green, and links to variation pages are shown at the right. Line numbers have been added.

Now let's click on **Gene tree**, which will display the current gene in the context of a phylogenetic tree of orthologous and paralogous genes.

Ensembl Home > Human Location: 4:123,592,080-123,597,100 Gene: IL2 Transcript: IL2-001 Login / Register | BLAST/BLAT | Docs & FAQs

**Gene: IL2**  
 Interleukin-2 Precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin) Source: UniProtKB/Swiss-Prot P60568  
 Location: Chromosome 4: 123,592,080-123,597,100 reverse strand.  
 Transcripts: There is one transcript in this gene: IL2-001 (ENST00000226730), with protein product ENSP00000226730.

**Gene: IL2 (ENSG00000109471)**

Interleukin-2 Precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin) Source: UniProtKB/Swiss-Prot P60568  
 Location: Chromosome 4: 123,592,080-123,597,100 reverse strand.  
 Transcripts: There is one transcript in this gene: IL2-001 (ENST00000226730), with protein product ENSP00000226730.

Genomic alignments | **Gene Tree** | Orthologues

View options:  
 • View current gene only  
 • View paralogs of current gene  
 • View all duplication nodes  
 • View fully expanded tree

Use the 'configure page' link in the left panel to set the default. Further options are available from menus on individual tree nodes.

Ensembl release 51 - Nov 2008 © WTSI / EBI  
 Permanent link - View in archive site

Click **View fully expanded tree** at the bottom.

Now lets view genetic variation mapped onto all transcripts of a gene.

Click on **Variation image** at the left.

Ensembl Home > Human Location: 4:123,592,080-123,597,100 Gene: IL2 Transcript: IL2-001 Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

**Gene: IL2**  
 Interleukin-2 Precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin) Source: UniProtKB/Swiss-Prot P60568  
 Location: Chromosome 4: 123,592,080-123,597,100 reverse strand.  
 Transcripts: There is one transcript in this gene: IL2-001 (ENST00000226730), with protein product ENSP00000226730.

Variation Table | **Variation Image** | External Data

Variations  
 ENST transcript  
 Protein domains  
 Variations

Configuring the display  
 Tip: use the 'Configure this page' link on the left to customise the protein domains and types of variations displayed above.  
 Please note the default 'Context' settings will probably filter out some intronic SNPs.  
 None of the variations are filtered out by the Source, Class and Type filters.  
 21 intronic variations are removed by the Context filter.

Click any variation, then **Variation properties** to learn more about it. A fourth tab will open:

**Variation: rs41304870**

**Variation: rs41304870**

**Variation type** SNP (source [dbSNP](#))

**Synonyms** None currently in the database

**Alleles** T/C (Ambiguity code: Y)

**Location** [4:123592534](#) (forward strand)

**Validation status** Unknown

**Linkage disequilibrium data** No linkage data for this SNP

**Flanking Sequence**

```

TAACATTC AACATAAATAAATATTTTGGGATAAATAAGTGAAACCATTTTAGAGCCCC
TAGGGCTTACAAAAAGAAATCATAAAAGATCCATATTTATAGTTTTAAGATTAAGAATAAT
AGTTACAATAGGTAGCAAACCATACATTCACAAATAAATAAATAAATAAATAAATAAATAA
AAATAGAAGGCCGTGATATGTTTTAAGTGGGAAGCACATTAATTATCAAGTCAGTTGAGA
TGATGCTTTTGACAAAAGGTAATCCATCTGTTTCAGAAAATTCACAAATGGTTGCTGTCAT
CAGCATATTCACACATGAAATGTTTTCAGATCCCTATAAAAAGAAAAATGTTAATTTTTT
AAAGTACAGAGTAGTTTACCTTATATACAGTTATTCCTCCAA
TGCCTTTATTTCAAGCTTACCAAAACATATTAATTATTCACATTTCTTGAATGATCAA
AATGAATGCCAAAATTACATATTTTGTATGATACCAAATGATAGTAACACAGAAGCTGA
GATTTTCTCATTGTTAATTCAGTCAATGAACACAAAAATTTTTCCTAAATGGGGTG
CAAAATAAAGATTTAAATAAATAAATAGATCTAGACAAATATAAATGCTGGATGTGAATAA
ACTTGATCGCTTTTCTCTGAAATGTACACCTATATTTGGTAAAGTAAATGTATGAATGG
ATATGCACATTATCTTTCTTTCTTAAATCTTGATTAGAGTCCCTATTTTCTACT
TAAAGTAATCTCCTTATATTT
    
```

(Variant highlighted)

Ensembl release 52 - Dec 2008 © [WTSI](#) / [EBI](#)  
[Permanent link](#) - [View in archive site](#)

Now, let's focus on one transcript. Select the transcript from the header section by clicking on the *Transcript tab* for IL2. This will lead to the Transcript summary display.

**Transcript: IL2-001 (ENST00000226730)**

Interleukin-2 Precursor (IL2)(T-cell growth factor)(TCGF)(Aldesleukin) [Source:UniProtKB/Swiss-Prot;Acc:P60568]

**Location** [Chromosome 4: 123,592,080-123,597,100](#) reverse strand.

**Gene** This transcript is a product of gene [ENSG00000109471](#)

**Protein** [ENSP00000226730](#) is the protein product of this transcript

**Statistics** Exons: 4 Transcript length: 794 bps Translation length: 153 residues

**CCDS** This transcript is a member of the Human CCDS set: [CCDS3726](#)

**Type** Known protein coding

**Prediction Method** Transcript where the Ensembl genebuild transcript and the [Vega](#) manual annotation have the same sequence, for every base pair. See [article](#).

**Alternative transcripts** This Ensembl/Havana merge transcript entry corresponds to the following database identifiers:  
Havana transcript having same CDS: [OTTHUMT00000256715](#) [\[view all locations\]](#)

Ensembl release 52 - Dec 2008 © [WTSI](#) / [EBI](#)  
[Permanent link](#) - [View in archive site](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)

Again, the left hand navigation column provides several options for this particular transcript.

**Transcript-based displays**

- Transcript summary
- Exons (4)
- Supporting evidence (23)
- Sequence
  - cDNA
  - Protein
- External References
  - General identifiers (91)
  - Oligo probes (11)
  - Gene ontology (23)
- Genetic Variation
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (13)
  - Variations (2)
- External Data
- ID History
  - Transcript history
  - Protein history
- Export transcript data

- Bookmark this page
- Configure this page
- Add custom data to page

Choose the **Exons** option first, which displays full exon sequences and either full or shortened introns. Use the **Configure this page** link to change the display (for example, show more flanking sequence, show full introns).

**Ensembl**  
 Home > Human  
 Location: chr4:123,592,080-123,597,100 | Gene: IL2 | Transcript: IL2-001 | Version: rs41304370

Transcript-based displays

Transcript: IL2-001 (ENST00000226730)  
 Interleukin-2 Precursor (IL-2)(T-cell growth factor)(TCGF)(Aldeleukin) [Source:UniProtKB/Swiss-Prot; Acc: P60560]

Location: [Chromosome 4: 123,592,080-123,597,100](#) reverse strand

Gene: This transcript is a product of gene [ENSG00000109471](#)

Protein: [ENSP00000226730](#) is the protein product of this transcript

Exons [help](#)

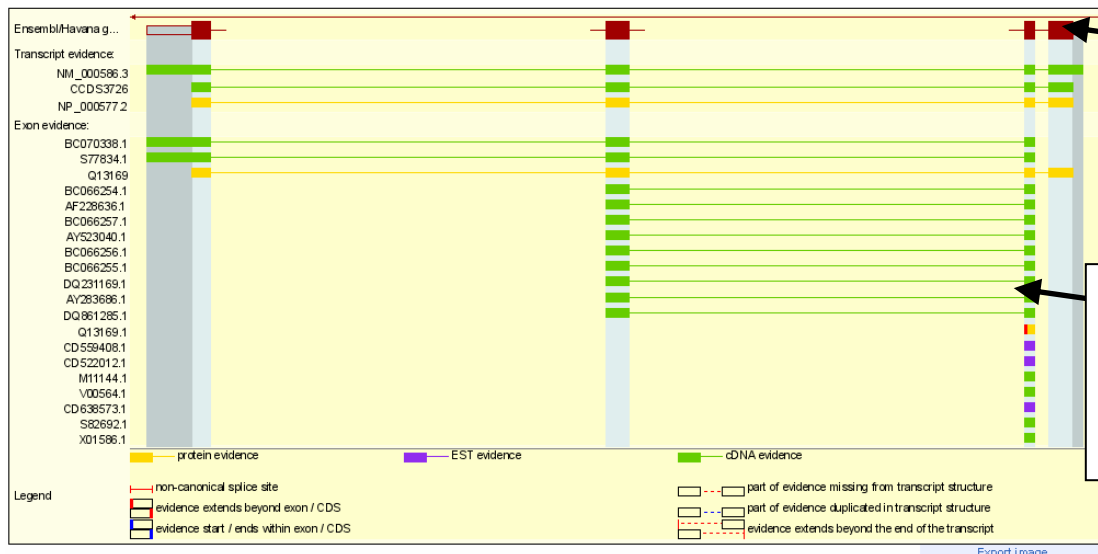
No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
5' upstream sequence							
1	<a href="#">ENSE00001293064</a>	123,596,899	123,597,100	-	0	202	AGTTCCCTATCAGCTCTTTAACTACTACAGTAACCTCAACTCTGCCACAAATGTA CAGGATGCAAGCTCTGCTTTGCAATGGCACTAAGTCTTGGCACTTGTGCAAGAACAGTGCACC TACTGCAAGTTCTACAAAGAAAACACAGCTACAACTGGAGCATTTTACTGCTGATTTACA GATGATTTGAAITGGAAITTAAT
	<a href="#">Intron 1-2</a>	123,596,899	123,596,898			90	gttaagatattctccttctcctaataa.....ataaacaaagcctatactactctccttag
2	<a href="#">ENSE00000986280</a>	123,596,749	123,596,808	0	0	80	AATTCAAGAATCCCAAACTCAACCAGGATGCTCACATTTAAGTTTACATGGCCAAAGAG gttaagtaacaaatcttcttctccta.....gagctgatgataacttcaatctctag
	<a href="#">Intron 2-3</a>	123,594,459	123,596,740			2,290	
3	<a href="#">ENSE00000986278</a>	123,594,315	123,594,458	0	0	144	GCCACAGAACTGAACATCTTCAGTCTTAGAAGAAGAAGCTCAAACTCTGGAGGAAGT CTAAATTTAGCTCAAGCAAACTTTCACCTTAAAGCCAGGACTTAATCAGCAATATC AACGTAATAGTTCGGAACTAAAG
	<a href="#">Intron 3-4</a>	123,592,468	123,594,314			1,847	gttaaggcattacttcttctctctc.....aaaattaacatttctctttatag
4	<a href="#">ENSE00001130295</a>	123,592,080	123,592,467	0	-		GGATCGAAACAACTTCATGTGTGAATATGCTGATGAGACAGCAACCATTGTAGAAATTT CTGAACAGATGGATACCTTTTGTCAAAGCATCATCTCAACACTGACTGTGATAATTAAGT GCTTCCCACTTAAACATATCAGGCCCTCTATTTTAAATTTAAATTTTAAATTTTAAATTTA TTGTGAAATGTATGTTTTCACCTTATGTAACCTATATGTAACCTATATCTTAAATTTAAATTTA TATGGACCTTTTAAATTTCTTTTGTGAAGCCCTAAGGGCTTAAATGTAACCTTAAATTTA TATCCCAAAATTTTATTTATTTATGTTGATGTTAAATATAGTATCTATGTAACCTTAAATTT GTAAACTATTTAAATTTGATAAAT
3' downstream sequence							

Green: Flanking sequence  
 Black: Coding sequence  
 Blue: Introns  
 Purple: UTR

Have you forgotten what the colours mean? No worries- click on the **Help** button (circled in red) and read the help for this page. A link to the glossary is also provided.



Next, follow the *Supporting Evidence* link. The following data display is quite an important one, as it shows which biological evidence has been used for the annotation of this transcript.

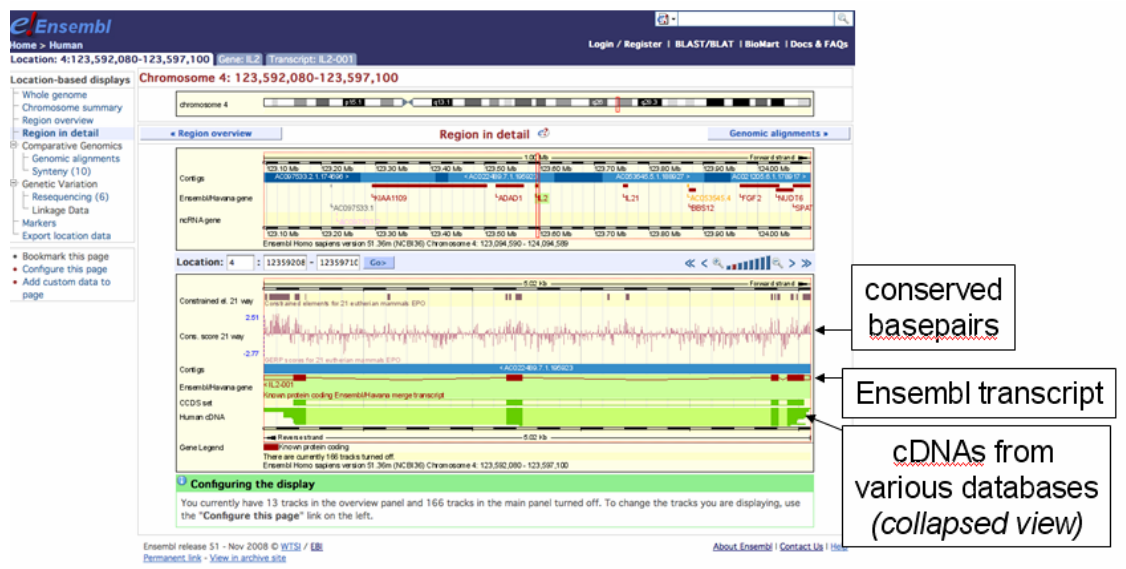


Red boxes:  
 exons in the  
 Ensembl  
 transcript  
 (mRNA)

Alignments of  
 cDNA and  
 protein to the  
 Ensembl  
 exons.

Other transcript-specific displays include the cDNA sequence, general identifiers and gene ontology terms from the GO consortium ([www.geneontology.org](http://www.geneontology.org)).

Let's now view the genomic region in which this gene and its transcript have been annotated by clicking onto the *Location* tab. I've drawn the constrained elements track in the diagram below. How can we view this track?



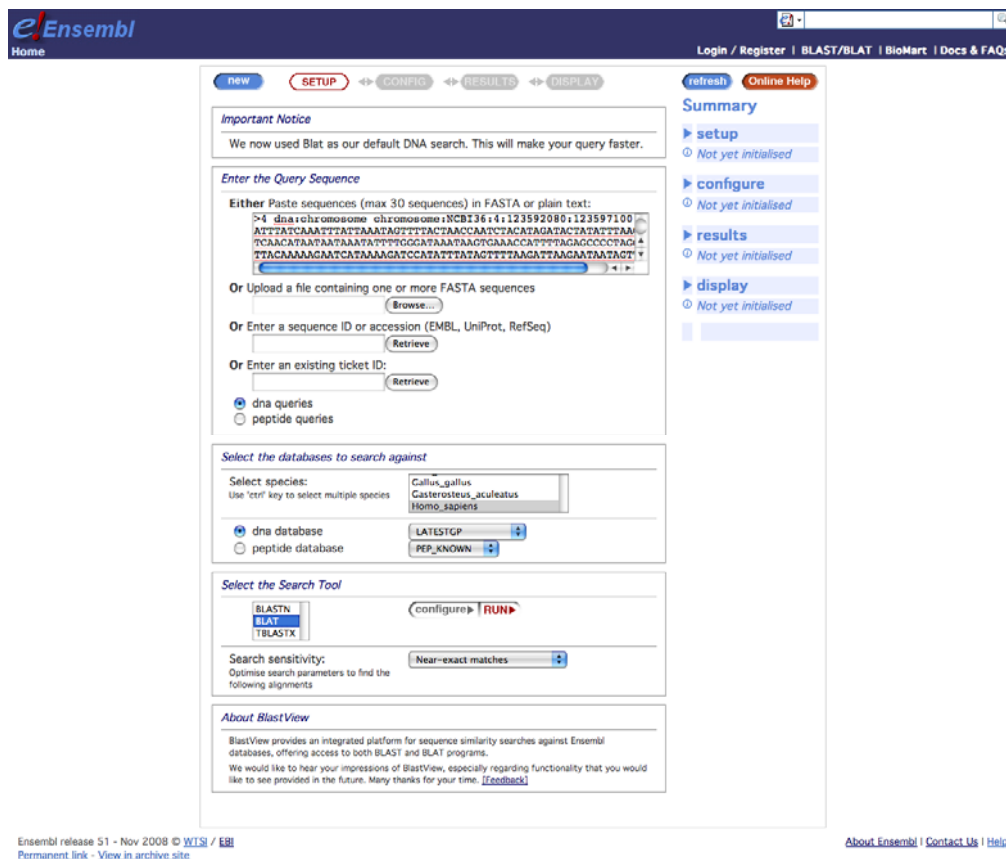
Ensembl *Location* displays are highly configurable. You can switch on additional tracks displaying various feature types that Ensembl annotates in the genome. Use the *Configure this page* link, find the

**Multiple alignments** menu, and then turn on the Conservation score and Constrained elements for 31 eutherian mammals. Also, add *all variations* to the *Region in detail* display and view the *cDNAs* track in normal, expanded form.

After investigating the *Location display*, we would like to export genomic sequence. Click the *Export location* data option and select the FASTA sequence format.

```
>4 dna:chromosome chromosome:NCBI36:4:123592080:123597100:1
ATTTATCAAATTTATTAATAGTTTACTAACCAATCTACATAGATACTATATTTAACAT
TCAACATAATAAATATTTTGGGATAAATAAGTGAACCATTTTAGAGCCCCTAGGGC
TTACAAAAGAATCATAAAAGATCCATATTTATAGTTTTAAGATTAAGAATAATAGTTAC
AATAGGTAGCAAACCATACATTCAACAATAAATATAAAATTTAAATATTTAAATAAATAG
AAGGCCTGATATGTTTTAAGTGGGAAGCACTTAATATCAAGTCAGTGTGAGATGATGC
TTTGACAAAAGGTAATCCATCTGTTTCAGAAAATCTACAATGGTTGCTGTCTCATCAGCAT
ATTCACACATGAATGTTGTTTCAGATCCCTATAAAAAGAAAATGTTAATTTTTTAAAGTA
CAGAGTAGTTTACCTTATATACAGTTATTTCCCAATGAAGTCTTATAGGCCTGTTGCCCTT
TATTTTCAAGCTTACCAACATATTAATTATTCACATTTTCTTGAATGATCAAATGAA
TGCCAAAATTACATATTTGTTATGATACCAATGATAGTAACACAGAAGCTGAGATTTT
CCTCATTTGTTAATTCAGTCAATGAACACAAAATTTTTTCCCTAAATGGGTGCAATA
AAGATTTAATAAAAATAGATCTAGACAAATATTAATGCTGGATGTGAAATAAATTGA
TCGTTTTCTCTGAATGTACACCTATATTTGTGTAAGTAAATGTATGAATTGATATGC
ACATTATACTTTCTTTCTTAAATCTTGATTAGAGTCTCCTATTTTCTACTTAAAGTG
AATCTCCTATATTTTCCCAAAGCAGAACAGAACTACACTAGAGTAAGCTAGGACATGC
...
```

Select the header and a few lines of sequence and then follow the *BLAST/BLAT* link in the bar at the top of the page. Paste the sequence into the appropriate box and select *BLAT* as the search algorithm. Finally, click *Run*.



The screenshot shows the Ensembl BLAST/BLAT interface. At the top, there are navigation tabs: 'new', 'SETUP', 'CONFIG', 'RESULTS', and 'DISPLAY'. The 'SETUP' tab is active. On the right side, there is a 'Summary' panel with links for 'setup', 'configure', 'results', and 'display', each with a 'Not yet initialised' status. The main form area contains several sections: 'Important Notice' (stating that Blat is now the default DNA search), 'Enter the Query Sequence' (with a text area containing the FASTA sequence and a 'Retrieve' button), 'Select the databases to search against' (with 'dna database' selected and 'LATESTGP' chosen), and 'Select the Search Tool' (with 'BLAT' selected and a 'RUN' button). At the bottom, there is an 'About BlastView' section.

Ensembl Home

new SETUP CONFIG RESULTS DISPLAY

Displaying 4 sequence alignments vs Homo\_sapiens LATESTGP database

Showing top 100 alignments of 1, sorted by Raw Score

Alignment Locations vs. Karyotype (click arrow to hide)

Alignment Locations vs. Query (click arrow to hide)

Alignment Summary (click arrow to hide)

Select rows to include in table, and type of sort (Use the "ctrl" key to select multiples)

Query	Subject	Chromosome	Supercontig	Clone	Contig	Chromosome	Stats	Sort By	
off. Name Start	off. Name Start	off. Name Start	off. Name Start	off. Name Start	off. Name Start	off. Name Start	off. Score E-val	>Chromosome <Score >Score	
Links	Query	Chromosome	End	Ori	Chromosome	End	Ori	Score E-val %ID Length	
[A] [S] [G] [C]	1	3060 +	Chr:4	123592080	123595139 +	Chr:4	123592080	123595139 +	15109 0.0e+00 100.00 3060

Ensembl release 51 - Nov 2008 © WTSI / EBI

Permanent link - View in archive site

About Ensembl | Contact Us | Help

Finally, follow links to an alignment [A], the query sequence [S], the genome sequence [G] and the corresponding Location View [C] (for its former name ContigView... or to C (see) the BLAST hit!).

Ensembl Home > Human

Location: 4:123,590,080-123,597,139

Chromosome 4: 123,590,080-123,597,139

Region in detail

Contigs

EnsemblHavana gene

ncRNA gene

Location: 4 : 12359006 - 12359713

Constrained el. 21 way

Cons. score 21 way

BLAT/BLAST hits

EnsemblHavana gene

CCDS set

Human cDNA

Reg. Features

Gene Legend

BLAT hit

Note: you can export the image using the link at the bottom.

END of WORKED EXAMPLE

## EXERCISES and ANSWERS

Note: the answers to these exercises correspond to current version (53) of Ensembl. If you use these exercises at a later date, please use the archive site for version 53.

### III) BROWSING ENSEMBL

These exercises address using the browser to determine a variety of gene-relevant information such as transcript number and size, protein domains, functional classes and sequence.

#### 1. Exploring features related to a gene

*Exercise 1 begins with the TAC1 (tachykinin precursor 1) gene and moves into the browser from the main GeneView page.*

(a) Open the home page of Ensembl ([www.ensembl.org](http://www.ensembl.org)) (or click on the big 'e!' from the top left corner of any Ensembl page. Search for the human TAC1 gene by typing 'human TAC1 gene' in the search window.

(b) How many transcripts are determined for this gene? What is the size of the longest predicted mRNA? How many exons does it have? How many amino acids does it code for?

(c) Follow some of the links in the 'General Identifiers' section of one of the Transcript tabs. What information can be found about the transcript? View any GO (Gene Ontology) terms.

(d) Which protein domains does the protein product contain?

(e) In which chromosomal contig and base pair position in the genomic sequence assembly is the TAC1 gene located?

(f) Is there a putative zebrafish orthologue? If so, where is it in the zebrafish genome?

#### 2. Exploring a region

*Exercise 2 begins with a search for a specific chromosomal region, rather than one gene.*

(a) Click on the large 'e!' at the top left of the screen to start a new search. Find the region:  
zebrafish chromosome 6:58744787-58772982

(b) This gene has two transcripts, both determined by homology evidence (therefore they are novel). Find the supporting evidence for one of the transcripts.

(c) There was no protein or mRNA from z-fish at the time of the genebuild (the determination of the Ensembl geneset). Find when the genebuild was performed.

(d) Go back in the browser to 'Region in Detail'. Let's see if there is any new zebrafish protein or mRNA information that aligns in this region. Turn on the UniProtKB track, and click on the links to see if any zebrafish proteins have been submitted. Also, draw zebrafish Expressed Sequence Tags (ESTs). What do you see?

### **3. Exploring a Region in Dolphin (*Tursiops truncatus*)**

(a) Who sequenced the dolphin assembly?

(b) View the top 40 InterPro hits, showing common protein domains found in dolphin genes. What's the first hit?

(c) Click on 'Gene' under the 'Sample entry points' at the left. How many transcripts are shown, and what sequence was used to determine the gene?

(d) Click on the 'Transcript' tab. View the cDNA. How many exons form this transcript?

(e) Export the cDNA of the transcript you are looking at. Run a BLASTN job against the human genome.

## **Answers (Browsing Ensembl)**

### **1. Exploring features related to a gene**

(a) Click on the identifier ENSG00000006128 from the search results. To ascertain it is indeed the TAC1 gene check that the HGNC symbol (the 'official' gene name given by the HUGO Gene Nomenclature Committee) is 'TAC1'. You should now be in the Gene tab.

(b) The TAC1 gene (ENSG00000006128) has 3 predicted transcripts, ENST00000346867, ENST00000350485 and ENST00000319273. Click on each ENST... identifier for more information about these transcripts. The longest transcript is ENST00000319273. See this information in the heading of each 'Transcript section' of the GeneView page. The length of ENST00000319273 is 1060 bp. It has 7 exons and codes for 129 aa.

(c) The TAC1 gene encodes for Protachykinin 1 precursor. Follow the links to MIM and EntrezGene or UniProt/Swiss-Prot in the 'Similarity Matches' section

to learn more. Also the GO (Gene Ontology) section can give you clues about the biological and molecular function of the TAC1 protein. Tachikininins are neuropeptides. These hormones are thought to function as neurotransmitters which interact with nerve receptors and smooth muscle cells. They are known to induce behavioral responses and function as vasodilators and secretagogues.

(d) Click the 'Domains and Features' link from the transcript tab. The domains include IPR013055 (Tachykinin/Neurokinin like), IPR002040 (Tachykinin/Neurokinin), IPR008215 (Tachykinin) and IPR008216 (Protachykinin). These also relate to domains in the original database such as Prints, Pfam and ProSite.

(e) From the Gene tab, the location is shown as: Chromosome 7:97,199,311-97,207,696bp. In the 'Main panel' you can see that it is located on contig AC004140.2.1.74918. This marks the position in the genomic assembly.

(f) From the Gene tab, the 'Orthologues' link should show ENSDARG00000014490 as an orthologue. Click on it to go to its Gene tab to find that it is located on z-fish chromosome 19.

## 2. Exploring a region

(a) Type zebrafish 6:58744787-58772982 in the search box.

(b) Click on a transcript and follow the link to the ENST Identifier. Click on 'Supporting evidence' at the left to view the proteins and mRNA aligned to the genome to determine the Ensembl transcript.

(c) Click on 'Zebrafish' at the top left of the page, under the big 'Ensembl'. Click on 'Assembly and Genebuild' at the left.

(d) Click 'Configure this page' at the left. Click on 'Protein alignments' at the left. Choose either 'Normal' or 'Stacked unlimited'. ('Normal will only show maximum of 7 tracks').

Click on the yellow bars showing the aligned proteins. The filled boxes show where the UniProt protein has aligned, and the empty boxes are gaps in the alignment. If you click through the different tracks, you might see Q6P4P5.1 aligns to the genome in this region, and corresponds to a *Danio Rerio* protein. It was modified in July, 2008 (this can be found in the UniProt record).

ESTs are turned on using the 'Configure this page', 'EST alignments' track. ESTs are present for z-fish, however Ensembl does not use ESTs in the main gene determination. They are aligned to the genome separately, as alternate evidence, and signify the presence of mRNA.



### 3. Exploring genes in Dolphin

(a) Select the 'Dolphin' under 'All genomes' to see the genome sequence was determined by the Broad Institute, in conjunction with Baylor.

(b) Click on Top 40 InterPro hits at the left. This will show a list of common protein domains. The first match is to a Proline-rich region. Take the IPR000694 link to read more about it in InterPro.

(c) One transcript is shown in the Gene summary tab. The gene is projected from human, and the gene name is HBA\_TURTR. Click on 'External References' or 'Supporting Evidence' at the left to see which human gene was used.

(d) Click the Transcript tab. You might already see there are 3 exons, based on the transcript structure. (Remember, boxes are exons, and lines connecting them are introns). Click on the 'cDNA' link under 'Sequence' at the left. The cDNA sequence is shown (exons only- this is the splice transcript). The three exons are shown in alternating colours (black, blue, and black again). Click on the 'help' button to read more.

*Did you know? Read the Glossary by clicking on 'Help' and following the link at the left.*

(e) Click on 'Export transcript data' at the left, and export the coding sequence. Select the sequence, copy, and then paste it into the BLAT/BLAST viewer. BLASTN against the human genome will result in a hit. Take the 'C' from the top hit in the table to view the hit in the human genome.

## IV) BioMart

### Mining data- worked example

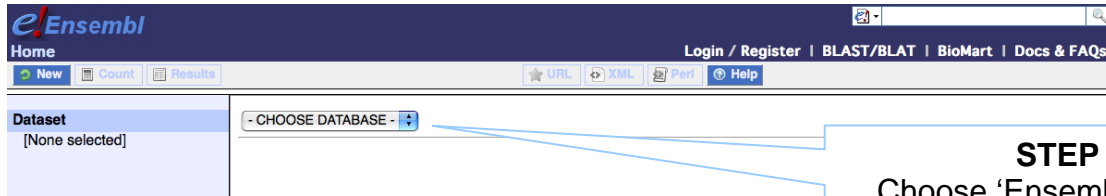
The human gene encoding Glucose-6-phosphate dehydrogenase (G6PD) is located on chromosome X in cytogenetic band q28.

Which other genes related to human diseases locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs?

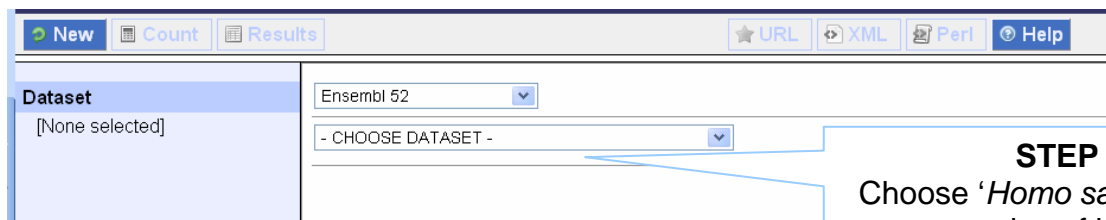
What are their cDNA sequences?

Follow the worked example below to answer these questions.

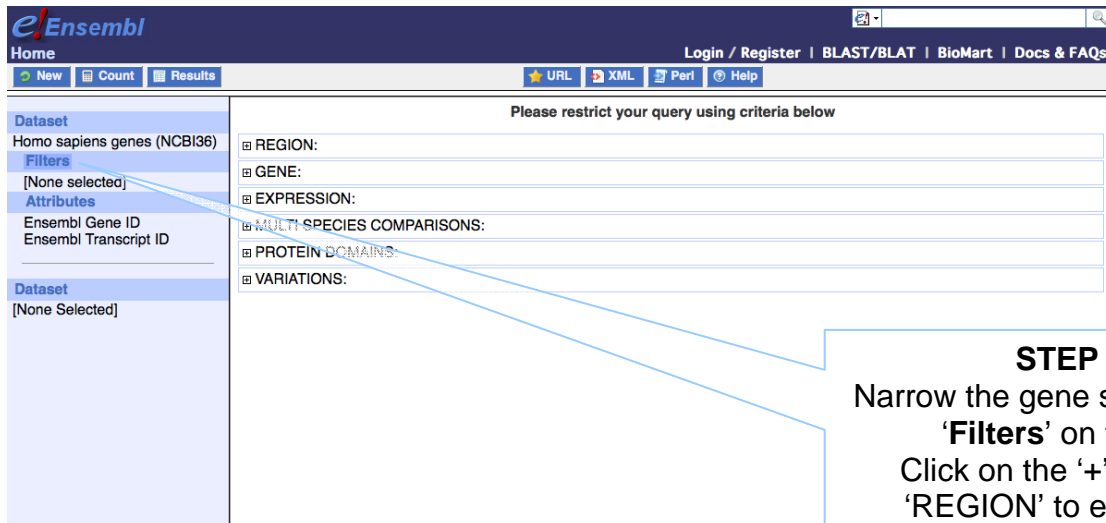
**Step 1:** Either click on 'BioMart' in the top right header bar of the Ensembl home page, or go to <http://www.biomart.org/> and click on the 'MartView' tab.



**STEP 2:**  
Choose 'Ensembl 53' as the primary database.



**STEP 3:**  
Choose 'Homo sapiens' as the species of interest.



**STEP 4:**  
Narrow the gene set by clicking 'Filters' on the left. Click on the '+' in front of 'REGION' to expand the choices.

**STEP 5:**  
Select 'Chromosome X'

**STEP 6:**  
Select 'Band Start q28' and 'End q28'

**STEP 7:**  
Expand the 'GENE' panel.

**STEP 8:**  
Limit to genes **with MIM disease ID**.  
These associations have been determined using MIM (Online Mendelian Inheritance in Man).  
<http://www.ncbi.nlm.nih.gov/omim/>

**STEP 9:**  
The filters have determined our gene set.  
Click 'Count' to see how many genes have passed these filters.

New Count Results URL XML Perl Help

Dataset 24 / 37435 Genes  
 Homo sapiens genes (NCBI36)

Filters  
 Chromosome: X  
 Band Start: q28  
 Band End: q28  
 with MIM disease ID(s): Only

Attributes  
 Ensembl Gene ID  
 Ensembl Transcript ID

Dataset  
 [None Selected]

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

Features  Structures  Sequences  Variations

GENE:  
 EXTERNAL:  
 EXPRESSION:  
 PROTEIN:

The 'Count' results show 24 human genes out of 37,435 total genes passed the filters.

**STEP 10:**  
 Click on 'Attributes' to select output options (i.e. what we would like to know about our gene set).

Home Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

New Count Results URL XML Perl Help

Dataset  
 Homo sapiens genes (NCBI36)

Filters  
 Chromosome: X  
 Band Start: q28  
 Band End: q28  
 with MIM disease ID(s): Only

Attributes  
 Ensembl Gene ID  
 Ensembl Transcript ID

Dataset  
 [None Selected]

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

Features  Homologs  Structures  Sequences  Variations

GENE:  
 EXTERNAL:  
 EXPRESSION:  
 PROTEIN:

**STEP 11:**  
 Expand the 'GENE' panel.

Home Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

New Count Results URL XML Perl Help

Dataset  
 Homo sapiens genes (NCBI36)

Filters  
 Chromosome: X  
 Band Start: q28  
 Band End: q28  
 with MIM disease ID(s): Only

Attributes  
 Ensembl Gene ID  
 Ensembl Transcript ID  
 Associated Gene Name

Dataset  
 [None Selected]

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

Features  Homologs  Structures  Sequences  Variations

GENE:  
**Ensembl Attributes**  
 Ensembl Gene ID  
 Ensembl Transcript ID  
 Ensembl Protein ID  
 Canonical transcript stable ID(s)  
 Description  
 Chromosome Name  
 Gene Start (bp)  
 Gene End (bp)  
 Strand  
 Band  
 Transcript Start (bp)  
 Transcript End (bp)  
 Associated Gene Name  
 Associated Transcript Name  
 Associated Gene DB  
 Associated Transcript DB  
 Transcript count  
 % GC content  
 Biotype  
 Source  
 Status (gene)  
 Status (transcript)

EXTERNAL:  
 EXPRESSION:  
 PROTEIN:

Note the summary of selected options.  
 The order of attributes determines the order of columns in the result table.

**STEP 12:**  
 Select, along with the default options, 'Associated Gene name' (this shows the gene symbol from HGNC).



New Count Results URL XML Perl Help

Dataset 24 / 37435 Genes  
 Homo sapiens genes (NCBI36)

Export all results to: File TSV Unique results only **Go**

Email notification to: \_\_\_\_\_

View: 10 rows as HTML Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name	Entrez Gene ID	MIM Morbid Accession	MIM Morbid Description
ENSG00000185010	ENST00000360256	F8	2157	306700	HEMOPHILIA A
ENSG00000185010	ENST00000360256	F8	2157	134500	FACTOR VIII DEFICIENCY
ENSG00000130826	ENST00000369550	DKC1	1736	300240	HOYERAAL-HREIDARSSON SYNDROME
ENSG00000130826	ENST00000369550	DKC1	1736	305000	DYSKERATOSIS CONGENITA, X-LINKED
ENSG00000073009	ENST00000369609	IKBK G	8517	308300	INCONTINENTIA PIGMENTI
ENSG00000073009	ENST00000369609	IKBK G	8517	300640	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
ENSG00000073009	ENST00000369609	IKBK G	8517	300636	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
ENSG00000073009	ENST00000369609	IKBK G	8517	300584	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
ENSG00000073009	ENST00000369609	IKBK G	8517	300301	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS
ENSG00000073009	ENST00000369609	IKBK G	8517	300291	ECTODERMAL DYSPLASIA, HYPOHIDROTIC, WITH IMMUNE DEFICIENCY

**STEP 16:**  
 Go back and change Filters or Attributes if desired.  
 Or, View ALL rows as HTML...

To save a file of the complete table, click 'Go'.  
 Or, email the results to any address.

Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name	Entrez Gene ID	MIM Morbid Accession	MIM Morbid Description
<a href="#">ENSG00000185010</a>	<a href="#">ENST00000360256</a>	<a href="#">F8</a>	<a href="#">2157</a>	<a href="#">306700</a>	HEMOPHILIA A
<a href="#">ENSG00000185010</a>	<a href="#">ENST00000360256</a>	<a href="#">F8</a>	<a href="#">2157</a>	<a href="#">134500</a>	FACTOR VIII DEFICIENCY
<a href="#">ENSG00000130826</a>	<a href="#">ENST00000369550</a>	<a href="#">DKC1</a>	<a href="#">1736</a>	<a href="#">300240</a>	HOYERAAL-HREIDARSSON SYNDROME
<a href="#">ENSG00000130826</a>	<a href="#">ENST00000369550</a>	<a href="#">DKC1</a>	<a href="#">1736</a>	<a href="#">305000</a>	DYSKERATOSIS CONGENITA, X-LINKED
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369609</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">308300</a>	INCONTINENTIA PIGMENTI
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369609</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300640</a>	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369609</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300636</a>	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369609</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300584</a>	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369609</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300301</a>	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369609</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300291</a>	ECTODERMAL DYSPLASIA, HYPOHIDROTIC, WITH IMMUNE DEFICIENCY
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369601</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">308300</a>	INCONTINENTIA PIGMENTI
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369601</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300640</a>	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369601</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300636</a>	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369601</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300584</a>	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369601</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300301</a>	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369601</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300291</a>	ECTODERMAL DYSPLASIA, HYPOHIDROTIC, WITH IMMUNE DEFICIENCY
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369606</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">308300</a>	INCONTINENTIA PIGMENTI
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369606</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300640</a>	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369606</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300636</a>	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369606</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300584</a>	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369606</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300301</a>	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369606</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300291</a>	ECTODERMAL DYSPLASIA, HYPOHIDROTIC, WITH IMMUNE DEFICIENCY
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369607</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">308300</a>	INCONTINENTIA PIGMENTI
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369607</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300640</a>	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369607</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300636</a>	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369607</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300584</a>	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
<a href="#">ENSG00000073009</a>	<a href="#">ENST00000369607</a>	<a href="#">IKBK G</a>	<a href="#">8517</a>	<a href="#">300301</a>	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS

**Result Table 1**



**STEP 17:**  
To view sequences, go back to 'Attributes'

**STEP 18:**  
Select 'Sequences' and then expand the 'SEQUENCES' section.

**STEP 19:**  
Expand the 'SEQUENCES' panel and select 'cDNA sequences'.

**STEP 20:**  
Expand the 'Header Information' section.

**STEP 21:**  
Choose 'Associated Gene Name' and 'Chromosome Name', in addition to 'Ensembl Gene ID' and 'Ensembl Transcript ID'

**Again, View ALL rows as FASTA for the full list... (make sure pop-up blocker is off).**

## RESULTS

### >Header: Gene ID, Transcript ID, Chromosome and Gene Name

```
>ENSG00000073009|ENST00000369601|X|IKBK
AGCCCGTTCTCTGCTCCGCGCTTCTGGAGCACTGGCCAAGGCGGGCGATTAGGACCCAG
GTTACTTGGGCGGCGAGCTGGACTGTTTCTACTCCTCCCTCCCTCCACTGCGGGGTCT
GACCCCTACTCCTTGTGTGAGGACTCCTCTAGTTTCAGAGACATATTCTGTTACCCAACTI
GACTGCGCTCTATCGAGGTCGTTAAATTTCTCGGAAATGCTCACAATATAGTTTGGCAGC
TAGCCCTTGCCCTGTTGGATGAATAGGCACCTCTGGAAGAGCCAACTGTGTGAGATGGTG
CAGCCCAGTGGTGGCCCGGAGCAGATCAGGACGTACTGGGCGAAGAGTCTCCTCTGGGG
AAGCCAGCCATGCTGCACCTGCCTTCAAGAACAGGGCGCTCTGAGACCCCTCCAGCGCTGC
CTGGAGGAGAATCAAGAGCTCCGAGATGCCATCCGGCAGAGCAACCAGATTCTGCGGGAG
CGCTGCGAGGAGCTTCTGCAATTTCCAAGCCAGCCAGAGGGAGGAGAAGGAGTTCCATG
TGCAAGTTCCAGGAGGCCAGGAACTGGTGGAGAGACTCGGCCTGGAGAAAGCTCGATCTG
AAGAGGCAGAAGGAGCAGGCTCTGCGGGAGGTGGAGCACCTGAAGAGATGCCAGCAGCAG
ATGGCTGAGGACAAGGCCCTCTGTGAAAGCCAGGTGACGTCCTTGTCTCGGGGAGTGCAG
GAGAGCCAGAGTCTGTTGGAGGCTGCCACTAAGGAATGCCAGGCTCTGGAGGGTCCGGCC
CGGGCGGCGAGCAGCAGGCGCGGAGCTGGAGAGTGAGCGCGAGGGCGTGCAGCAGCAG
CACAGCGTGCAGGTGGACCAGCTGCGCATGCAGGGCCAGAGCGTGGAGGGCCGCGCTCCGC
ATGGAGCGCCAGCCCGCTCGGAGGAGAAGAGGAAGCTGGCCAGTTGCAGGTGGCCCTAI
CACCCAGCTCTTCCAAGAATACGACAACCACATCAAGAGCAGCGTGGTGGGCAGTGAGCGG
AAGCGAGGAATGCAGCTGGAAGATCTCAAAACAGCAGCTCCAGCAGGCCGAGGAGGCCCTG
GTGGCCAAACAGGAGGTGATCGATAAGCTGAAGGAGGAGGCCAGCAGCACAAGATTGTG
ATGGAGACCGTTCCGGTGTGTAAGGCCAGGCGGATATCTACAAGGGCGACTTCCAGGCTI
GAGAGCCAGGCCCCGGGAGAAGCTGGCCGAGAAGAAGGAGCTCTGCAGGAGCAGCTGGAG
CAGCTGCAGAGGGAGTACAGCAAACTGAAAGGCCAGCTGTCAAGGATCGGGCAGGATCGAG
GACATGAGGAAGCGGCATGTGAGGTTCTCCAGGCCCTTGTCCCGCCCGCCCTGCCTAC
CTCTCCTCTCCCTGGCCCTGCCAGCCAGAGGAGGAGGCCCGCCCGAGGAGCCACCTGAC
TTCTGCTGTCCCAAGTGCCAGTATCAGGCCCTGATATGGACACCCTGCAGATACATGTC
ATGGAGTGCATTGAGTAGGGCCGGCCAGTGCAAGGCCACTGCCTGCCGAGGACGTGCCCG
GGACCGTGCAGTCTGCGCTTTCTCTCCCGCTGCCAGCCAGGATGAAGGGCTGGGTG
GCCACAACCTGGGATGCCACCTGGAGCCCCACCCAGGAGCTGGCCGCGGCACCTTACGCTI
CAGCTGTTGATCCGCTGGTCCCTCTTTTGGGGTAGATGCGGCCCGGATCAGGCCCTGACTI
CGCTGCTCTTTTTGTTCCCTTCTGTCTGCTCGAACCACTTGCCCTCGGGGCTAATCCCTCCC
TCTTCTCCACCCGGCACTGGGGAAGTCAAGAATGGGGCTGGGGCTCTCAGGGAGAACI
GCTTCCCTGGCAGAGCTGGGTGGCAGCTCTTCTCCACCCGGACACCCGCGCCCGCTI
GCTGTGCCCTGGGAGTGTGCCCTCTTACCATGCACACGGGTGCTCTCCTTTTGGGCTGC
ATGCTATTCCATTTTGCAGCCAGACCGATGTGTATTTAACCCAGTCACTATTGATGGACA
TTGGGTTGTTTCCCATCTTTTGTACCATAAATAATGGCATAGTAAAAATCCTTGTGCA
TT
```

cDNA 1

```
>ENSG000000126895|ENST00000358927|X|AVPR2
TTCACGCCACCGCCAGCTGCCAGGAGCCAGCCAGGACTGGCCATACTGCCACCGACA
CGTGCAACACACGCCAACAGGCATCTGCCATGCTGGCATCTCTATAAGGGCTCCAGTCCAG
AGACCCCTGGGCCATTGAACTTGCTCCTCAGGCAGAGGCTGAGTCCGCACATCACCTCCAG
GCCCTCAGAACACCTGCCCGCCAGCCACCATGCTCATGGCGTCCACCATTCCGCTGTGC
CTGGGCATCCCTCTCTGCCAGCCCTGCCAGCAACAGCAGCCAGGAGAGGGCCACTGGACA
CCCGGGACCCGCTGCTAGCCCGGGCGGAGCTGGCGCTGCTCTCCATAGTCTTTTGGGCTG
TGGCCCTGAGCAATGGCCCTGGTGTCTGGCGGCCCTAGCTCGGGCGGGCCGGCGGGCCACTI
GGGCACCCATACAGTCTTCAATTGGCCACTTGTGCTGGCCGACCTGGCCGTTGGCTCTGI
TCCAAGTGTGCCCCAGCTGGCCCTGGAAGGCCACCGACCCTTCCGTGGGCCAGATGCC
TGTGTCGGGCCGTGAAGTATCTGCAGATGGTGGCCATGTATGCCCTCCTCCTACATGATCC
TGGCCATGACCGCTGGACCGCCACCGTGGCCTCTGCGCGTCCCTGCTGGGCTACCGCCATG
```

cDNA 2

## V) BioMart Exercises and Answers

*These exercises have been designed to familiarise you with different questions you can answer with this tool, and the types of data you can retrieve with BioMart.*

1. Retrieve all SNPs for 'known' human G-protein coupled receptor genes (GPCRs – use the InterPro domain ID: IPR000276) on chromosome 2.

*Note: As this is the first exercise we walk you this time through BioMart step-by-step (but of course you can also try to do this exercise without our help!)*

Start a new BioMart session by clicking 'New', or go back to the Ensembl homepage and click on 'Mine Ensembl with BioMart' under 'Ensembl tools'.

Choose the **database** and the **dataset** for your query as follows:

- Select 'Ensembl 53'
- Select 'Homo sapiens genes (NCBI36)'.

Click on '**Filters**' at the left. Filter this dataset to select your genes of interest as follows:

- Expand the 'REGION' section at the right by clicking on the '+'. Select 'Chromosome 2'. Click [count] at the top of the panel and note the number of Ensembl genes on *Homo sapiens* chromosome 2.
- In the 'GENE' section, select 'Status (gene)' 'KNOWN'.
- In the 'PROTEIN DOMAINS' section, select the 'Limit to genes with these family or domain IDs' option. Select 'InterPro ID(s)' and enter 'IPR000276' in the box. Click [count] again and note that the number of genes is now **28**.

Click on '**Attributes**' (at the left). Select the output for your gene list as follows:

- Select the 'Variations' Attribute Page.
- In the 'GENE' section 'Ensembl Gene ID' and 'Ensembl Transcript ID' are selected by default – also select 'Ensembl Protein ID'.
- In the 'GENE ASSOCIATED VARIATIONS' section 'Reference ID' is selected. Also select 'Allele', 'Protein location (aa)' and 'Protein Allele'.

*Note: Clicking on count now will not show an altered number. Attribute selections should not affect the count (i.e. the number of genes that have passed the filters).*

Click on '**Results**' (at the top) to obtain the first 10 rows of your table. To obtain the entire table select 'View all rows as HTML' or export a file by clicking 'Go'. Check the box 'Unique results only', otherwise you can end up with redundant rows!!

Why are two columns in the preview table blank? These variations are not in the coding sequence.

## Exercise 2

Generate a list of all zebrafish protein-coding genes that are located on chromosome 3. Export gene name, description, Zfin symbol, and InterPro domains.

## Exercise 3

**For this exercise, it's easier to cut and paste the IDs from the online course booklet here:**

[http://www.ebi.ac.uk/~gspudich/workshop\\_presentations/coursebook](http://www.ebi.ac.uk/~gspudich/workshop_presentations/coursebook)

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

NP\_001218, NP\_203125, NP\_203124, NP\_203126, NP\_001007233,  
NP\_150636, NP\_150635, NP\_001214, NP\_150637, NP\_150634,  
NP\_150649, NP\_001216, NP\_116787, NP\_001217, NP\_127463,  
NP\_001220, NP\_004338, NP\_004337, NP\_116786, NP\_036246,  
NP\_116756, NP\_116759, NP\_001221, NP\_203519, NP\_001073594,  
NP\_001219, NP\_001073593, NP\_203520, NP\_203522

Generate a list that shows to which Ensembl Gene IDs and to which HGNC symbols these RefSeq IDs correspond. Not all these IDs will have a match in Ensembl.

## Exercise 4

In a paper from 1995 Ayyagari *et al.* mapped the human 'Usher Syndrome type I C' to the genomic region between the markers D11S1397 and D11S1310 (Mol. Vis. 1:2, 1995).

Confirm this finding by generating a list of the genes located in this region.

## Exercise 5

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

207630\_s\_at, 221840\_at, 219228\_at, 204924\_at, 227613\_at, 223454\_at,  
228962\_at, 214696\_at, 210732\_s\_at, 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

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

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

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

### Exercise 6

Known dolphin genes match to a protein or mRNA in a public database for dolphin (this is in contrast to 'known by projection' which was based on evidence from another species).

**Step 1:** For all known dolphin genes in Ensembl, export human homologues.

**Step 2: *Advanced*:** export a list of the human gene IDs alone (select only one attribute, and then select 'Unique results only'.) Do a second query in BioMart with human genes, upload these gene IDs and export gene names!

### Exercise 7 - CHALLENGE

This a little bit more complicated query in which you have to use two datasets (one from Ensembl and one from Reactome, the biological pathway database (<http://www.reactome.org>)):

Determine in which pathways the gene ENSG00000164305 plays a role.

Note: This query can be very time consuming! Give it a try, but if it takes too long, just focus on the idea.

### Exercise 7

Design your own query!



## Answers: BIOMART

1. You should find **28** known genes on chromosome 2 with this InterPro domain. The result table is quite large; so don't export the entire table if export is going slowly.

2. Click '**NEW**' for a new query.

Start with all the zebrafish Ensembl genes:

Choose the '**ENSEMBL 53**' database.  
Choose the '**Danio rerio genes**' dataset.

Now filter for the genes on the 3 chromosome:

Click on '**Filters**' in the left panel.  
Expand the '**REGION**' section by clicking on the + box.  
Select '**Chromosome 3**'. Make sure the check box in front of the filter is ticked otherwise the filter won't work.

Now filter further for genes that are protein coding:

Expand the '**GENE**' section by clicking on the + box.  
Select '**Gene type**' as '**protein\_coding**'.  
Click the [Count] button on the toolbar.

This should give you 975 / 25546 Genes.

Specify the attributes to be included in the output (note that a number of attributes will already be default selected):

Click on 'Attributes' in the left panel.  
Select the 'Features' attributes page.  
Expand the '**GENE**' section by clicking on the + box.  
Select, in addition to the attributes 'Ensembl Gene ID' and 'Ensembl Transcript ID' that are already default selected, 'Associated Gene Name' and 'Description'.

Expand the '**EXTERNAL**' panel to select ZFIN symbols. These will be equal to the Gene Name, when those are available.

Expand the '**PROTEIN**' section to add 'InterPro ID' 'InterPro Short Description'.

Click the [Results] button on the toolbar.

If you are happy with how the results look in the preview, output all the results:

Select 'View All rows as HTML' or export all results to a file.

**3. Click [New].**

Choose the 'ENSEMBL 53' database.

Choose the '*Homo sapiens genes (NCBI36)*' dataset.

Click on '**Filters**' in the left panel.

Expand the '**GENE**' section by clicking on the + box.

Select '**ID list limit - RefSeq protein ID(s)**' and enter the list of IDs in the text box (either comma separated or as a list).

Click on '**Attributes**' in the left panel.

Select the '**Features**' attributes page.

Expand the '**GENE**' section by clicking on the + box.

**Deselect 'Ensembl Transcript ID'.**

Expand the '**External**' section by clicking on the + box.

Select '**HGNC symbol**' and '**RefSeq Protein ID**' from the '**External References**' section.

Click the [Results] button on the toolbar.

Select 'View All rows as HTML' or export all results to a file. Tick the box 'Unique results only'.

Note: BioMart is 'transcript-centric', which means that it will give a separate row of output for each transcript of a gene, even if you don't include the Ensembl Transcript ID in your output. When you don't want this, use the 'Unique results only' option.

Your results should show that all RefSeq IDs map to **11** Caspase genes.

Some RefSeq IDs map to the same Ensembl Gene ID and HGNC symbol.

**4. Click [New].**

Choose the 'ENSEMBL 53' database.

Choose the '*Homo sapiens genes (NCBI36)*' dataset.

Click on '**Filters**' in the left panel.

Expand the '**REGION**' section by clicking on the + box.

Enter 'Marker Start: **D11S1397**' and 'Marker End: **D11S1310**'.

Click on '**Attributes**' in the left panel.

Select the '**Features**' attributes page.

Expand the '**GENE**' section by clicking on the + box.

**Deselect 'Ensembl Transcript ID'.**

Select '**Associated Gene Name**' and '**Description**'.

Click the [Results] button on the toolbar.

Select 'View All rows as HTML' or export all results to a file. Tick the box 'Unique results only'.

Your results should show **28** genes. Among these there should be of one (ENSG00000006611) with name 'USH1C' and description 'Harmonin (Usher syndrome type-1C protein) (Autoimmune enteropathy-related antigen AIE-75) (Antigen NY-CO-38/NY-CO-37) (PDZ-73 protein) (Renal carcinoma antigen NY-REN-3). [Source:UniprotKB/SWISSPROT;Acc:Q9Y6N9]'. This suggests that Ayyagari et al. correctly mapped Usher Syndrome type I C to this genomic region.

**5. (a) Click [New].**

Choose the '**ENSEMBL 53**' database.

Choose the 'Homo sapiens genes (NCBI36)' dataset.

Click on '**Filters**' in the left panel.

Expand the '**GENE**' section by clicking on the + box.

Select '**ID list limit - Affy hg u133 plus 2 ID(s)**' and enter the list of probe-set IDs in the text box (either comma separated or as a list).

Click on '**Attributes**' in the left panel.

Select the '**Features**' attributes page.

Expand the '**GENE**' section by clicking on the + box.

Select, in addition to the default selected attributes, '**Description**'.

Expand the '**External**' section by clicking on the + box.

Select '**HGNC symbol**' from the '**External References**' section and '**AFFY HG U133-PLUS-2**' from the '**Microarray Attributes**' section.

Click the '**[Results]**' button on the toolbar.

Select 'View All rows as HTML' or export all results to a file. Tick the box 'Unique results only'.

Your results should show that **21** of the 25 probes map to an Ensembl gene.

**(b)** Don't change Dataset and Filters- simply click on '**Attributes**'.

Select the '**Sequences**' attributes page.

Expand the '**SEQUENCES**' section by clicking on the + box.

Select '**Flank (Transcript)**' and enter '**2000**' in the '**Upstream flank**' text box.

Expand the '**Header information**' section by clicking on the + box.

Select, in addition to the default selected attributes, '**Description**' and '**Associated Gene Name**'.

Note: 'Flank (Transcript)' will give the flanks for all transcripts of a gene with multiple transcripts. 'Flank (Gene)' will give the flanks for the transcript with the outermost 5' or 3' end.

Click the '**[Results]**' button on the toolbar.

Select 'View All rows as HTML' or export all results to a file.

(c) You can leave the Dataset and Filters the same, so you can directly go to the '**Attributes**' section:

Click on '**Attributes**' in the left panel.  
Select the '**Homologs**' attributes page.  
Expand the '**GENE**' section by clicking on the + box.  
Select, in addition to the default selected attributes, '**Associated Gene Name**'.

**Deselect 'Ensembl Transcript ID'.**

Expand the '**MOUSE ORTHOLOGS**' section by clicking on the + box.  
Select '**Mouse Ensembl Gene ID**', '**Mouse Chromosome**', '**Mouse Chr Start (bp)**' and '**Mouse Chr End (bp)**'.

Click the [Results] button on the toolbar.  
Check the box 'Unique results only'. Select 'View All rows as HTML' or export all results to a file.

Your results should show that for **18** out of the 21 human genes at least one mouse orthologue has been identified. ENSG00000123130 has two mouse orthologues and ENSG00000172716 has three. Three (ENSG00000186594, ENSG00000130844 and ENSG00000089335) have none.

**6. Step 1:** Choose 'Ensembl 51' and '*Tursiops truncatus*'. Filters: Expand the 'GENE' panel and select Status (gene) as Known.

Attributes: Select Human Ensembl Gene ID under the 'Homologs' page.

**Step 2:** Remove 'Ensembl Gene ID, Ensembl Transcript ID, and Ensembl Protein ID' from the Attributes. Click on 'Unique results only' and export the file.

Click NEW. Start with Ensembl 51, *Homo sapiens*. Filters: Expand the GENE panel, and click browse to upload a file into the 'ID List Limit Box'.

In Attributes, select 'Gene Name'.

Click Results.

**7. This query involves joining of Ensembl and Reactome BioMarts. To do this, Click on the 'MartView' tab from [www.biomart.org](http://www.biomart.org)**

Choose the '**ENSEMBL 51 GENES (SANGER UK)**' database.

Choose the '***Homo sapiens genes (NCBI36)***' dataset.

Click on '**Filters**' in the left panel.

Expand the '**GENE**' section by clicking on the + box.

Select '**ID list limit - Ensembl Gene ID(s)**' and enter '**ENSG00000164305**' in the text box.

Click on '**Attributes**' in the left panel.

Select the '**Features**' attributes page.

Expand the '**GENE**' section by clicking on the + box.

**Deselect 'Ensembl Transcript ID'.**

Click on the **second 'Dataset'** in the left panel.

Choose the '**[REACTOME (CSHL US)] pathway**' dataset.

Click on '**Attributes**' in the left panel.

**Deselect 'Pathway DB\_ ID'.**

**Select 'Pathway Name'.**

Click the **[Results]** button on the toolbar.

Select 'View All rows as HTML' or export your results to a file. Tick the box 'Unique results only'.

Your results should show that ENSG00000164305 plays a role in various processes associated with apoptosis (programmed cell death).

**VI) EXERCISES GENEBUILD****Exercise 1**

- (a) Who supplied the human genome assembly that Ensembl annotated?
- (b) How long did the last gene build take?
- (c) Has the gene set for human been updated since the last gene build was done?

**Exercise 2**

Find the Ensembl GALP (Galanin-like peptide precursor) gene for human.

- (a) From what source did Ensembl get the name for this gene? And from where did it get the description?
- (b) On how many pieces of evidence has the transcript of this gene been built?
- (c) Why do some pieces of evidence not support the first exon of the transcript?

**Exercise 3**

Find the Ensembl Epc1 (enhancer of polycomb homolog 1) gene for mouse.

- (a) How many transcripts has Ensembl annotated for this gene?
- (b) How many transcripts have the manual annotators of Havana annotated for this gene?
- (c) How many transcripts agree between Ensembl and Havana annotation?
- (d) What is the reason that Ensembl hasn't annotated one of the Havana transcripts?

**Exercise 4**

An example of what can go wrong ....

Go to the following page in Ensembl release 46 (of August 2007):

[http://aug2007.archive.ensembl.org/Homo\\_sapiens/geneview?gene=ENSG0000198561](http://aug2007.archive.ensembl.org/Homo_sapiens/geneview?gene=ENSG0000198561)

- (a) What is wrong with this gene? What could be the reason for this?

(b) Has this problem been fixed in Ensembl release 53?

### Exercise 5

Go to the region 104916000-104923000 on human chromosome 14. This should show you a part of the PACS2-201 transcript. Make sure the human cDNA track is switched on in 'Stacked unlimited' mode. Do you see anything unusual when you compare the cDNAs that have been mapped to this region of the genome with the transcript model annotated by Ensembl?

## ANSWERS GENEBUILD

### Exercise 1

Go to <http://www.ensembl.org>.

Click on the human picture or the word 'Human' next to it.

Click on 'Assembly and Genebuild' in the side menu.

(a) NCBI (The National Center for Biotechnology Information) hosts the assembly determined from the IHGP (International Human Genome Project).

(b) About ten months (from December 2006 until October 2007).

(c) Yes, in September 2008.

### Exercise 2

Go to the Ensembl homepage.

Under 'Search Ensembl' type 'human gene GALP' and click [Go].

On the page with search results click on 'Ensembl protein\_coding Gene: ENSG00000197487 (HGNC (automatic): GALP)'.

(a) From the HGNC (HUGO Gene Nomenclature Committee). From UniProtKB/Swiss-Prot.

Click on the 'Transcript: GALP-201' tab.

Click on 'Supporting evidence' in the side menu.

(b) On six pieces of evidence.

(c) The two pieces of protein evidence (NP\_149097.1 and Q9UBC7) as well as the CCDS evidence (CCDS12940) don't support the first exon of the GALP transcript, because this exon is completely untranslated. Thus, protein sequences and coding sequences cannot provide any information for this exon.

### Exercise 3

Go to the Ensembl homepage.

Under 'Search Ensembl' type 'mouse gene Epc1' and click [Go].



On the page with search results click on 'Ensembl protein\_coding Gene: ENSMUSG00000024240 (MGI Symbol: Epc1)'.  
Click on 'Configure this page' in the side menu.  
Click on 'Other genes', select 'Vega gene – Expanded with labels' and click [SAVE and close].

- (a) Three.
- (b) Four.
- (c) Three.

Click on the Epc-004 transcript in the figure.  
Click on 'OTTMUST00000041784' in the pop-up menu.  
Click on 'Supporting evidence' in the side menu.

(d) In this case the reason is that the transcript OTTMUST00000041784 is built on one piece of EST evidence (CF730975.1). As Ensembl doesn't build on just EST evidence, it consequently hasn't annotated this transcript.

#### Exercise 4

Go to

[http://aug2007.archive.ensembl.org/Homo\\_sapiens/geneview?gene=ENSG0000198561](http://aug2007.archive.ensembl.org/Homo_sapiens/geneview?gene=ENSG0000198561)

(a) It seems that two groups of transcripts belonging to different genes have been combined into one gene. This is also reflected by the fact that the gene has two HGNC symbols associated with it, CTNND1 and TXNDC14. The culprit is a cDNA that spans both genes (biological reality or artefact?) and on which the transcript O60716-27(ENST00000360682) has been built.

Go to the Ensembl homepage.

Under 'Search Ensembl' type 'human gene CTNND1' and click [Go].

On the page with search results click on 'Ensembl protein\_coding Gene: ENSG00000198561 (HGNC (automatic): CTNND1)'.

Click on the 'Location' tab.

Zoom out two steps, so both the CTNDD1 transcripts and the TXNDC14 transcripts are shown.

(b) In Ensembl release 52 CTNND1 and TXNDC14 are annotated as separate genes.

#### Exercise 5

Go to the Ensembl homepage.

Under 'Search Ensembl' select 'Human', type '14:104916000-104923000' and click [Go].

Click on 'Configure this page' in the side menu

Click on 'cDNA/mRNA alignments, select 'Human cDNA – Stacked unlimited' and click [SAVE and close].

One of the cDNAs aligned in this region (BC028418.1) seems to contain a retained intron. Is this a real transcript? In this case, Ensembl says no. If the ratio of healthy introns to retained introns is larger than five when comparing multiple cDNAs, like in this specific case, the corresponding cDNA is not used in the gene-build.

## VII) EXERCISES VARIATIONS

### Exercise 1

A non-synonymous SNP, R620W (C1858T), in PTPN22 (Tyrosine-protein phosphatase non-receptor type 22) has been identified as a genetic risk factor for rheumatoid arthritis if an individual is homozygous for the T allele (Van Oene *et al.* Arthritis Rheum. 2005 Jul;52(7):1993-8).

- Find the Ensembl page with information for this SNP.
- Why are the alleles on this page given as A/G and not as C/T?  
(Hint... is PTPN22 a reverse-stranded gene?)
- What is the minor allele of this SNP in Caucasians?
- Have a look at the cDNA of a PTPN22 transcript that contains this SNP. Can you find R620W? Can you find C1858T?

### Exercise 2

Find the Genetic Variation - Comparison image page for human PTPN22 (use transcript PTPN22-001).

- Do all individuals (HuAA, HuCC, HuDD, HuFF, Venter and Watson) have re-sequence coverage at the position of the R620W (C1858T) SNP?
- Do any individuals have a higher risk of getting rheumatoid arthritis based on his/her genotype at this position (remember that A is the minor allele)?
- Is there an individual that is a heterozygote at this position?

### Exercise 3

Use BioMart to generate an Excel spreadsheet that contains the following information on all SNPs in the transcripts of the human PTPN22 gene: reference ID, alleles (both nucleotides and amino acids), location (both in transcript and in protein) and consequence to the transcript.

Note: you can start with the Ensembl gene database, filter for the PTPN22 gene and then select your attributes from the 'Variations' attributes page.

## ANSWERS VARIATIONS

### Answer Exercise 1

Go to the Ensembl homepage.

Under 'Search Ensembl' type 'human gene PTPN22' and click [Go].

On the page with search results click on 'Ensembl protein\_coding Gene: ENSG00000134242 (HGNC (curated): PTPN22)'.

Click on 'Variation Table' in the side menu.

Click on 'Configure this page' in the side menu.

Under 'Select Variation Type', deselect all options except 'Non-synonymous' and click [SAVE and close].

(a) Two of the three transcripts of PTPN22 contain a SNP with AA change W/R and AA co-ordinate 620, so this SNP, rs2476601, must be the one we are looking for.

Click on 'rs2476601'.

(b) The alleles for this SNP are given as A/G, because these are the alleles in the forward strand of the assembly. The SNP is referred to as C1858T because the PTPN22 gene is located on the reverse strand of the assembly.

Click on 'Population genetics'.

(c) In Caucasians (CSHL-HAPMAP:HapMap-CEU population) the minor allele is A.

Click on the 'Gene:PTPN22' tab.

Click on 'ENST00000359785'.

Click on 'cDNA' in the side menu.

(d) The numbering 1858 is relative to the start of the coding sequence (ATG). This is the second sequence shown in the figure. The first sequence is the complete transcript sequence, the third sequence is the protein sequence.

### Answer Exercise 2

Click on the 'Gene: PTPN22' tab.

Click on 'ENST00000359785'.

Click on 'Comparison image' in the side menu.

Click on 'Configure this page' in the side menu.

Under 'Select Variation Type', deselect all options except 'Non-synonymous', under 'Select Individuals' select all individuals and click [SAVE and close].

(a) There is only re-sequencing coverage for Venter and Watson.

- (b) Neither Venter nor Watson is homozygous for the minor allele (A) of rs2476601, that predisposes one for rheumatoid arthritis.  
(c) Watson is heterozygous for rs2476601.

### **Answer Exercise 3**

Go to the Ensembl homepage  
Click the BioMart link on the toolbar.

Choose the 'Ensembl 52' database.  
Choose the 'Homo sapiens genes (NCBI36)' dataset.

Click on 'Filters' in the left panel.  
Expand the 'GENE' section by clicking on the + box.  
Select 'ID list limit – HGNC symbol' and enter 'PTPN22' in the text box.

Click on 'Attributes' in the left panel.  
Select the 'Variations' attributes page.  
Expand the 'GENE ASSOCIATED VARIATIONS' section by clicking on the + box.  
Select, in addition to the attribute 'Reference ID' that is already default selected, 'Allele', 'Protein Allele', 'Transcript location (bp)', 'Protein location (aa)' and 'Consequence Type (Transcript Variation)'.

Click the [Results] button on the toolbar.  
Select 'Export all results to file – XLS' and click [Go].

This should give you an Excel spreadsheet with 785 rows.

## **VIII) EXERCISES COMPARATIVE GENOMICS**

### **Exercise 1**

Find the Ensembl CASP5 (Caspase-5) gene of human.

- (a) How many within-species paralogues are predicted for this gene? Note the Target %id and Query %id. Which paralogue has the most sequence similarity with CASP5? Retrieve an alignment between CASP5 and one of its paralogues.  
(b) Is there an orthologue predicted for this gene in gorilla?  
(c) Have a look at the genetree for this gene. Which of the paralogues of CASP5 is due to the most recent duplication event? Is this reflected in the sequence similarity between CASP5 and this paralogue when compared with the other paralogues (see also question a)?  
(d) Retrieve an alignment between members of any node using Jalview.

## Exercise 2

Find the Ensembl CYCS (Cytochrome c) gene for human.

- Between which genes is CYCS located in human?
- Is there an orthologue predicted for this gene in horse?
- Does the Cytochrome c protein family contain a horse protein?
- What does the annotation (e.g. similarity matches, GO terms, protein domains) suggest with regard to the function of the gene encoding this horse protein?
- Does the location of this gene in horse confirm your finding in (d)?

## Exercise 3

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

- Turn on some of the BLASTZ alignment tracks and some of the Translated BLAT alignment tracks. Does the degree of conservation between human and the various other species reflect their evolutionary relationship? Which parts of the BRCA2 gene seem to be the most conserved? Did you expect this?
- Turn on the 'Constrained elements 31 way' and 'Conservation score 31 way' tracks. Do these tracks confirm what you already saw in the tracks with pairwise alignment data?

## ANSWERS COMPARATIVE GENOMICS

### Answer Exercise 1

Under 'Search Ensembl' type 'human gene CASP5' and click [Go]. On the page with search results click on 'Ensembl protein\_coding Gene: ENSG00000137757 (HGNC (curated): CASP5)'.

- Click on 'Paralogues' in the side menu.

There are eleven within-species paralogues predicted for human CASP5. The first one has the highest Target %id and Query %id.

Click on [Align] next to the paralogue.

- Click on 'Orthologues' in the side menu.

Yes, there is an orthologue predicted for human CASP5 in gorilla: ENSGGOG00000015759 (CASP5).

- Click on 'Gene Tree' in the side menu.  
Click on 'View paralogs of current gene' under the figure.

Click on the nodes (red squares) for the duplication events that have given rise to the various paralogues.

CASP5 and CASP4 are related by a duplication event on the level of the level of the Eutheria, while CASP5 and CASP12 are related by is a duplication event on the level of the Theria. CASP5 and CASP4 are closer in the gene tree. This agrees with the fact that CASP4 shows the most sequence similarity with CASP5 (see question a).

(d) Click on the duplication node (red square) or speciation node (blue square) of the sub-tree that you are interested in.

In the pop-up menu click on [Start Jalview].

To edit the alignment display, you can remove sequences using the option Edit > Delete in the menu bar. Note the other available edit options, e.g. Remove Empty Columns.

## Answer Exercise 2

Under 'Search Ensembl' type 'human gene CYCS' and click [Go].

On the page with search results click on 'Ensembl protein\_coding Gene: ENSG00000172115 (HGNC (curated): CYCS)'.

Click on the Location tab.

(a) The human CYCS gene is located between the OSBPL3 and C7orf31 genes.

(b) There is no orthologue predicted for human CYCS in horse.

Go back to the Gene tab and click on 'Orthologues' in the side menu to see this.

(c)

Click on the 'Protein families' in the side menu.

Click on 'all proteins in family'.

The Cytochrome c protein family does contain an Ensembl protein from horse, ENSECAP00000012031.

Click on 'ENSECAP00000012031'.

(d) The gene encoding ENSECAP00000012031 is named CYC\_HORSE, and its annotation strongly suggests that it encodes for the Cytochrome c protein.

Click on the Location tab.

(e) The CYC\_HORSE gene is, just like the human CYCS gene, located between the OSBPL3 and C7orf31 genes, which is additional proof that we seem to be dealing here with the horse Cytochrome c gene.



### Answer Exercise 3

Under 'Search Ensembl' type 'human gene BRCA2' and click [Go].

On the page with search results click on 'Ensembl protein\_coding Gene: ENSG00000139618 (HGNC (curated): BRCA2)'.

Click on the Location tab.

Click on 'Configure this page' in the side menu

Click on 'BLASTZ alignments', select some tracks, click on 'Translated BLAT' alignments, select some tracks and click [SAVE and close].

(a) Species that are closer to human in evolution show a larger extent of conservation. Especially the exon sequences of BRCA2 seem to be highly conserved between the various species, which is what you would expect because these are expected to be under higher selection pressure.

(b) The 'Conservation score' and 'Constrained elements 31 way' tracks largely correspond with the data in the pairwise alignment tracks; the exons of the BRCA2 gene seem to show high conservation. A highly conserved region in a BRCA2 intron corresponds to the position of a pseudogene.