

# Access to genes and genomes with Ensembl



**Introduction and Worked Example**

Spring 2008



## 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

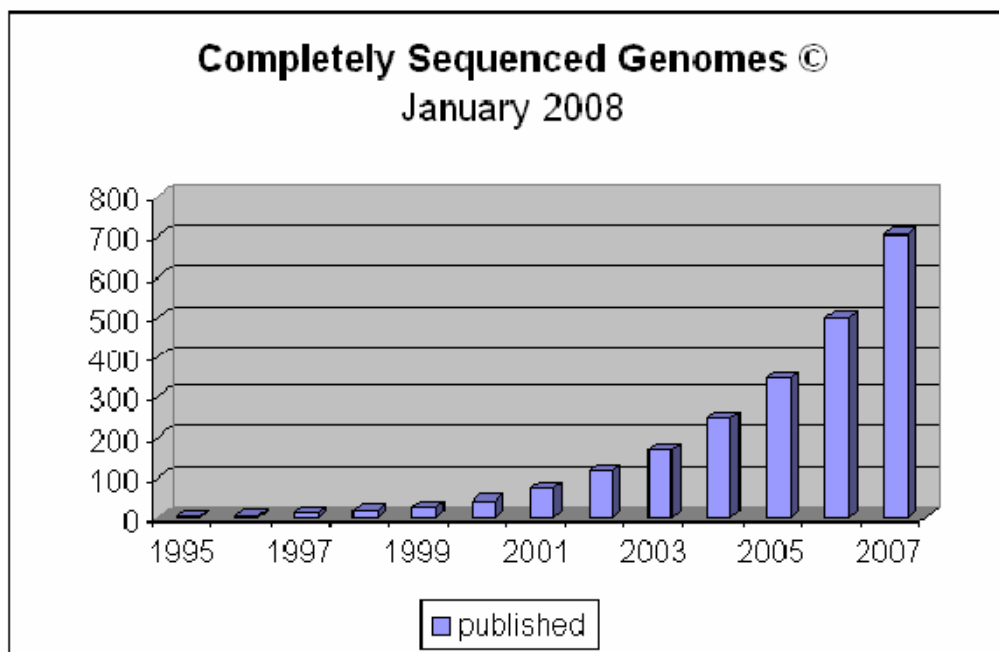


Figure 1. Completely sequenced genomes as of Jan, 2008 (figure taken from <http://www.genomesonline.org>).

laboratory biologist when provided along with quality annotation of the genomic sequence. 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 easy access to genomic information with a number of visualisation tools. The Ensembl website gives you for example the possibility to directly download data, whether it is a 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 called Views (e.g. GeneView, ContigView and SNPView), and will all be introduced appropriately later on. An updated version of the website is released bimonthly. Old

versions are for at least two years accessible on the 'Archive!' website. 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. Finally, Ensembl BLAST offers the possibility to perform sequence searches against genomes and Ensembl gene and peptide sets.

### Further reading

Flicek, P. *et al.*

#### **Ensembl 2008**

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

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.*

#### **Ensembl 2006.**

Nucleic Acids Res. 2006 Jan 34:D556-D561 (2006)

Hubbard, T. *et al.*

#### **Ensembl 2005.**

Nucleic Acids Res. 2005 33 D447-D453 (2005)

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

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

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

Kasprzyk, A. *et al.*

#### **EnSMart: a generic system for fast and flexible access to biological data.**

Genome Research (2004) 14:1, 160-9.

Ashurst, J. L. *et al.*

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

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

\* Additional references can be found here:

<http://www.ensembl.org/info/about/publications.html>

---

<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

<b>SPECIES: Mammals</b>		<b>ASSEMBLY</b>		<b>GENEBUILD</b>	
Armadillo	<i>Dasybus novemcinctus</i>	ARMA	May-05	Ensembl	Aug-06
Bushbaby	<i>Otolemur garnettii</i>	otoGar1	May-06	Ensembl	Feb-07
Cat	<i>Felis catus</i>	CAT	Feb-07	Ensembl	Jun-06
Chimpanzee	<i>Pan troglodytes</i>	PanTro 2.1	Mar-06	Ensembl	May-07
Cow	<i>Bos taurus</i>	Btau 3.1	Feb-07	Ensembl	Sep-06
Dog	<i>Canis familiaris</i>	CanFam 2.0	May-06	Ensembl	Dec-06
Elephant	<i>Loxodonta africana</i>	BROAD E1	May-05	Ensembl	Aug-06
Guinea pig	<i>Cavia porcellus</i>	cavPor2	Feb-07	Ensembl	Oct-06
Hedgehog	<i>Erinaceus europaeus</i>	eriEur1	Feb-07	Ensembl	Oct-06
Horse	<i>Equus caballus</i>	Equus1	Sep-07	Ensembl	Sep-07
Human	<i>Homo sapiens</i>	NCBI 36	Oct-05	Ensembl	Sep-07
Lesser hedgehog tenrec	<i>Echinops telfairi</i>	TENREC	May-05	Ensembl	Aug-06
Microbat	<i>Myotis lugifugus</i>	myoLuc1	Mar-06	Ensembl	Jan-07
Mouse	<i>Mus musculus</i>	NCBI m37	Apr-07	Ensembl	Sep-07
Mouse Lemur	<i>Microcebus murinus</i>	micMur1	Jun-07	Ensembl	Jul-07
Opossum	<i>Monodelphis domestica</i>	MonDom 5.0	Oct-06	Ensembl	Feb-07
Orangutan	<i>Pongo pygmaeus</i>	PPYG2	Sep-07	Ensembl	Oct-07
Pig*	<i>Sus scrofa</i>	Sscrofa1			
Pika	<i>Ochotona princeps</i>	OchPri2.0	Jun-07	Ensembl	Jul-07
Platypus	<i>Ornithorhynchus anatinus</i>	OANA 5	Dec-05	Ensembl	Jan-07
Rabbit	<i>Oryctolagus cuniculus</i>	RABBIT	May-05	Ensembl	Aug-06
Rat	<i>Rattus norvegicus</i>	RGSC 3.4	Dec-04	Ensembl	Feb-06
Rhesus macaque	<i>Macaca mulatta</i>	MMUL 1	Feb-06	Ensembl	Aug-06
Shrew	<i>Sorex araneus</i>	sorAra1	Oct-05	Ensembl	Apr-07
Squirrel	<i>Spermophilus tridecemlineatus</i>	speTri1	Jun-06	Ensembl	Oct-06
Tree shrew	<i>Tupaia belangeri</i>	tupBel1	Feb-07	Ensembl	Oct-06
<b>Species: Other</b>					
Aedes	<i>Aedes aegypti</i>	AaegL 1	Oct-05	VectorBase	Jun-06
Anole Lizard*	<i>Anolis carolinensis</i>	AnoCar1.0	Feb-07		
Anopheles	<i>Anopheles gambiae</i>	AgamP 3	Feb-06	VectorBase	Jun-07
<i>C. elegans</i>	<i>Caenorhabditis elegans</i>	WS 180	Sep-07	WormBase	Sep-07
Chicken	<i>Gallus gallus</i>	WASHUC 2	May-06	Ensembl	Aug-06
<i>C. intestinalis</i>	<i>Ciona intestinalis</i>	JGI 2	Mar-05	Ensembl	Feb-06
<i>C. savignyi</i>	<i>Ciona savignyi</i>	CSAV 2.0	Oct-05	Ensembl	Apr-06
Fruitfly	<i>Drosophila melanogaster</i>	BDGP 5.4	Nov-07	FlyBase	Oct-07
Lamprey*	<i>Petromyzon marinus</i>	PMAL3			
Medaka	<i>Oryzias latipes</i>	HdrR 1	Oct-05	Ensembl	May-06
<i>S. cerevisiae</i>	<i>Saccharomyces cerevisiae</i>	SGD 1.01	Dec-06	SGD	Dec-06
Stickleback	<i>Gasterosteus aculeatus</i>	BROAD S1	Feb-06	Ensembl	Jun-06
Takifugu	<i>Takifugu rubripes</i>	FUGU 4.0	Jun-05	Ensembl	Nov-07
Tetraodon	<i>Tetraodon nigroviridis</i>	TETRAODON 7	Apr-03	Genoscope	Sep-04
<i>X. tropicalis</i>	<i>Xenopus tropicalis</i>	JGI 4.1	Aug-05	Ensembl	Nov-05
Zebrafish	<i>Danio rerio</i>	Zv 7	Apr-07	Ensembl	Jun-07

Table 1 – Species, assemblies and gene sets in Ensembl (\*currently available in Pre! website [pre.ensembl.org](http://pre.ensembl.org))

**II) WORKED EXAMPLE – A walk through the main pages of the Ensembl browser, using the IL2 (Interleukin-2 precursor) gene as an example.**

**STEP 1:**  
Load Ensembl  
www.ensembl.org

The screenshot shows the Ensembl website interface with several callouts:

- Navigation column:** A green callout pointing to the left sidebar containing links like 'Your Ensembl', 'Login or Register', 'Help & Documentation', and 'Ensembl Archive'.
- Search:** A green callout pointing to the search bar at the top of the page.
- Help:** A green callout pointing to the 'HELP' link in the top right corner.
- Help pages and Documents:** A green callout pointing to the 'Help & Documentation' section in the sidebar.
- STEP 2: Click on "Human":** A yellow callout pointing to the 'Human' link in the 'Popular genomes' section.
- What's new:** A green callout pointing to the 'Ensembl headlines' section at the bottom of the page.

**STEP 3:**  
Type in 'IL2 Gene'.  
Click 'Go'.

**Karyotype**

**Source and version of assembly and genebuild**

**e!Ensembl Human**

Ensembl release 44 - Apr 2007

Search Ensembl *Homo sapiens*

Search:  **Go**

e.g. chromosome X or 14:10000..20000 or BRCA2

**Karyotype**

Click on a chromosome for a closer view

Jump directly to sequence position

Chromosome:  or region

From (bp):

To (bp):  **Go**

**What's New in Ensembl 44**

**Homo sapiens News**

- Patch for Ensembl Human database**  
Ensembl *Homo sapiens* has been patched with a few extra transcripts and some new CCDS IDs, and the corresponding xrefs and variation databases have been updated.
- cDNA Updates**  
Ensembl human and mouse databases have received their usual cDNA updates.
- Vega updates**  
Vega human and mouse have both been updated since the last release of Ensembl. See [VEGA](#) for more information.
- Variation updates**  
All species with variation data now have a failed\_variation table. Also, in species that have duplicate variations (with the same mapping but different IDs), these have been put into the variation\_synonym table.

**General News**

- ncRNAs for Ensembl chordates**  
All Ensembl genebuild databases (i.e. excluding the imported invertebrate databases for mosquitoes, fruit fly, worm and yeast) have been updated with new ncRNA data.

[More news...](#)

**About the Human genome**

**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**

The human genome sequence is now considered sufficiently stable that the three major genome browsers have come together to produce a common set of gene IDs for their annotations. This Consensus CDS ID set has been incorporated into the Ensembl database alongside the existing identifiers.

- [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 project](#)

**Statistics**

<b>Assembly:</b>	NCBI 36, Oct 2005
<b>Genebuild:</b>	Ensembl, Aug 2006
<b>Database version:</b>	44.36f
<b>Known genes:</b>	21,724
<b>Novel genes:</b>	1,017
<b>Pseudogenes:</b>	1,040
<b>RNA genes:</b>	4,113
<b>Immunoglobulin/T-cell receptor gene segments:</b>	388
<b>GenScan gene predictions:</b>	69,185
<b>Gene exons:</b>	270,214
<b>Gene transcripts:</b>	44,567
<b>SNPs:</b>	11,561,833
<b>Base Pairs*:</b>	3,253,037,807
<b>Golden Path Length**:</b>	3,093,120,360

\* Total number of base pairs = sum of lengths of DNA table  
\*\* Reference assembly (Golden path) length = sum of non-redundant top level seq regions

© 2007 [WTG](#) / [EBI](#). Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.



# The Gene View Page

Gene Model

Whole Genome Alignments

Orthologues in other species

External Information (DAS)

IDs in other databases

GO (Gene Ontology) terms

Protein domains

**STEP 5:**  
Click on 'Transcript Information'



**Result of STEP 6:**

```

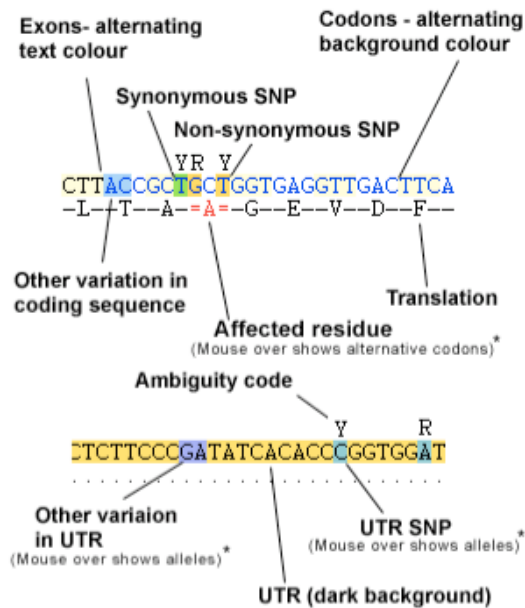
121 AAGGTAAATGTTTTTTCAGACAGGTAAAGTCTTTGAAAATATGTGTAAATATGTAAAACATT
.....

181 TTGACACCCCCATAAATATTTTTCCAGAATT AACAGT AT AAATTGCATCTCTTGTTCAAGA
.....

241 GTTCCCTATCACTCTCTTTAATCACTACTCACAGTAACCTCAACTCCTGCCACAATGTAC
.....-M--Y-

301 AAGGATGCAACTCCTGTCTTGCATTGCACTAAGTCTTGCACCTTGTCACAAACAGTGCACCT
  3 -R--M--Q--L--L--S--C--I--A--L--S--L--A--L--U--T--N--S--A--P-

351 ACTTCAAAGTTCTACAAAAGAAAACACAGCTACAACCTGGAGCATTACTGGCTGGATTTACAG
      KK
 23 -T--S--S--S--T--K--K--T--Q--L--Q--L--E--H--L--L--L--D--L--Q-
  
```



# Result of STEP 7:

**STEP 9:**  
Click on 'Graphical view'

**STEP 8:**  
Choose 'Flanking sequence at either end of transcript – 500', tick 'Show full intronic sequence' and click on [Go]

Human ExonView

ENST00000226730

- Gene
- Gene image
- Gene name
- Gene info.
- ID history
- Reset alignment
- Transcript information
- Exon information
- Protein information
- Export transcript data

Chromosome 4  
123,592,000 - 123,597,339

- View of chromosome 4
- Graphical view
- Graphical overview
- Export from region...

Ensembl Archive

- View previous release of page in Archive!
- Stable Archive! link for this page

Sanger EMBL EBI

**Horse**  
*Equus caballus*

Now in Ensembl

Ensembl Exon Report

Transcript: IL2 (HGNC Symbol) Synonyms: IL-2 To view this transcript is a member of the Human CCDS set: CCDS3726 Havana transcript having same CDS: OTTHUMT0000256715

Ensembl Transcript ID: ENST00000226730

Transcript information: Exons: 4 Transcript length: 1,033 bps Translation length: 153 residues This transcript is a product of gene: ENSG00000109471

Genomic Location: This transcript can be found on Chromosome 4 at location 123,592,080-123,597,339. The start of this transcript is located in Contig AC022489.7.1.195923.

Description: Interleukin-2 precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin) Source: UniProt/SWISSPROT P60568

Rendering options:
 

- Flanking sequence at either end of transcript:  50
- Intron base pairs to show at splice sites:  25
- Show full intronic sequence:
- Show exons only:

[Go]

Flank (green)

UTR (purple)

Intron (blue)

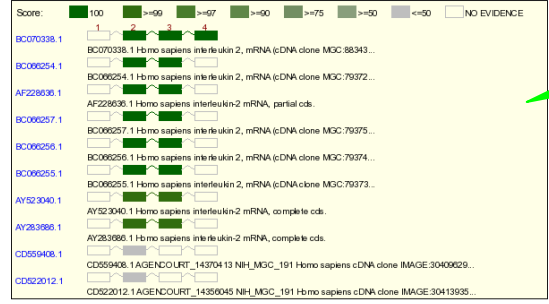
Coding sequence (black)

Exon Information

No.	Exon / Intron	Chr	Strand	Start	End	Start Phase	End Phase	Length	Sequence
1	5' upstream sequence	4	-1	123,596,899	123,597,339	-	0	441	.....ccttaagaagaaggaggaaaaaactgtttccatacagaaggccttaattgcat
1	ENSE00001293064	4	-1	123,596,899	123,597,339	-	0	441	GAATTAGAGCTATCACTAAGTGTGGGCTAATGTAACAAGAGGGATTTCACCTACATCC ATTCAAGTCAAGTCTTTGGGGGTTTAAAGAAATTCCAAAGAGTCATCAGAGGAGGAAAAATG AAGTAATGTTTTTTCAGACAGTAAAGTCTTTGAAATATGTGTAATATGTAAGAAATG TTGACACCCCCATAATATTTTCCAGAATTAACAGTATAAATGCAATCTCTTGTTCAGAA GTTCCCATCACTCTTTAATCACTACTCACAGTAACCTCAACTCCTGCCACAATGTAC AGGATGCAACTCCTGTCTTGCATTGCACTAAGTCTTGCACCTTGCAAAAACAGTGCACCT ACTTCAAGTTCTCAAAAAGAAAACACAGCTACAACCTGGAGCAITTAAGTCTGCAITTAAG ATGATTTTGAATGGAATTAAT
	Intron 1-2	4	-1	123,596,899	123,597,339	-	0	90	gtaagtattttcttcttactaa.....atacaaatgattatattctttcttag
2	ENSE00000935280	4	-1	123,596,749	123,596,808	0	0	60	AATTACAAGAATCCAAACTCACCAGGATGCTCACATTTAAGTTTACATGCCAAGAAG gtaagtacaatattttgttcaat.....gagctgatgataattattcttag
	Intron 2-3	4	-1	123,596,749	123,596,808	0	0	2,280	gtaagtacaatattttgttcaat.....gagctgatgataattattcttag
3	ENSE00000935280	4	-1	123,596,808	123,597,339	0	0	44	GCCACAGAAGTAAACATCTTCAGTGTCTAGAAGAAAGAACTCAACCTGTGAGGGAAGT CTAAATTTAGCTCAAAAGCAAAAACCTTCACTTAAGACCAGGGACTTAATCAGCAATATC TACCTAATAGTTCTGGAACTAAG
	Intron 3-4	4	-1	123,597,339	123,597,339	0	0	388	gtaaggcattactttttgtctc.....aaaaataacatttttttttag
4	ENSE00001138280	4	-1	123,597,339	123,597,339	0	0	388	GGATCTGAAACAACATTCATGTGTAATATGCTGATGAGACAGCAACATTTGAGAATTT CTGAAACAGATGGATTACCTTTTGGCAAGCATCATCTCAACACTGACTTGAATATTAAGT GCTTCCCACTTAAAAATATCAGGCTCTTATTTATTTAAATTAATTAATTTATTTATTTA TTGTTGAATGTATGTTTGTACCTATTTGAACATTAATCTTAATCTTAAACTATATAA TATGGATCTTTATGATCTTTTGTAAAGCCTAGGGGCTCAAAATGGTTTCACTTATT TATCCCAAAATATTTATTTATGTTGAATGTTAAATATAGTATCTATGATGATTTGGTTA GTAAACTATTTAATAAATTTGATAAAT
	3' downstream sequence								ataaacagcctggatatttgttattttggaaacagcagcagtaagcat.....

Supporting Evidence

The supporting evidence below consists of the sequence matches on which the exon predictions were based and are sorted by alignment score. There are a large number of supporting evidence hits for this transcript. Only the top ten 10 hits have been shown. [Click to view all 22 supporting evidence hits.](#)



Supporting evidence

**Result of STEP 8:**

Gene ID	Strand	Start	End	Feature	Length	Sequence				
1	ENSE00001293064	4	-1	123,596,899	123,597,339	-	0	441	GAATTAGAGCTACCTAAGTGTGGGCTAATGTAACAAGAGGGATTTCACCTACATCC AATCAGTCAGTCTTTGGGGGTTTAAAGAAATTCCAAAGAGTCATCAGAAGAGGAAAAATC AAGGTAATGTTTTTCAGACAGGTAAAGTCTTTGAAAATATGTGAATATGTAAAAATTT TTGACACCCCATTAATATTTTTCCAGAATTAACAGTATAAATTCATCTCTTTGTTCAAGA GTCCCTATCACTCTCTTTAATCACTACTCACAGTAACTCAACTCCTGCCAACATGTAC AGGATGCAACTCCTGTCTTTGCATTGCACATAAGTCTTGCACCTTGCACAAACAGTGCACCT ACTTCAAGTCTACAAAAGAAAACACAGCTACAACCTGGAGCATTTACTGTGGATTTACAG ATGATTTTGAATGGAATTAAT	
	Intron 1-2	4	-1	123,596,809	123,596,898			90	gtaagtataatttccttcttactaaaattattacatttagtaactagctggagatcatt tcttaataacaatgcattatactttcttag	
2	ENSE00000935280	4	-1	123,596,749	123,596,808			0	60	AATTACAAGAAATCCCAAACCTACCAGGATGCTCACATTTAAGITTTACATGCCCAAGAAG
	Intron 2-3	4	-1	123,594,459	123,596,748			2,290		gtaagtacaatttttattgttcaatttctgttttaataaaaattcaaaagtaaatgaaaaat ttgcacagatgggactaatagcagctcatctgggtaaaagtaactttaattgtttttt ttgaaaacccaagtttgataatgaagcctctataaaacagttttacctatatttttaaat atatatttgggttgggtggggggggaagaaacataaaaaataatttctactttatc gataagacaattcctaacaaaaatgttcatttaggtttcattaaaaatgtaaaactct aaaatattgattatgtcatttagtatgtaaaatcacaatactattccaaggagccc acttttaaaaacttttcttggtttaggaaggtttctaagtgaaggcagcataaacact aatagcacagagctctggggccagatactgaagtgaatctcagctctgccaatgctctag ctttcatgatctttggcaaatcactactctgtttgtgattcagtttcatgtctacttaa atgaataactgtatatacttaatatggctttgtgagaattagtaagtaaatgtaaaagcac tcagaaccgtgtctggcataaggtaaataccatacaagcattagctattatttagtagtat taaagataaaaatttctactgagaaatcacaagtaaaaatttggactttatctttttacca atagaacttgagatttataatgctatatgacttaatttccaagataaaaagcttcatag gttgttttggattcagatagagcataagcataatcatccaagctcctagctacattag ntgtgtaaaactacctagtagctgtgcccagttaagagagaaatgaacaaaaatctggtgcca gaaagagcttggccagggtgaatccaagcccagaaaaataggatataagggggacaca gatgcaatcccatgactcaaatcttataattcaagagaaaatctgcttcaactaccct tctgaaagatgtaaaaggagacagcttacagatgttactctagtttaactcagagccacata atgcaactccagaaacataaaagatactagatgctgttttctgaaagaaaattctccacat

Flank (green)

UTR (purple)

Coding sequence (black)

Intron (blue)

## Result of STEP 9: ContigView

**Chromosome** (pointing to 'Chromosome 4')

**Markers** (pointing to 'Markers' in the left sidebar)

**1 Mb region** (pointing to the 'Chr. 4' overview map)

**Assembly** (pointing to the 'Vega assembly' track)

**IL2 transcript** (pointing to the 'Ensembl trans.' track showing the IL2 transcript)

**STEP 10:**  
 In the 'Features' drop-down menu select 'SNPs' and 'Ensembl genes' (deselect other options). Close the menu to view new tracks.

© 2008 WTSI / EBI. Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

**(STEP 10):** Deselect all other options from the features menu and select 'SNPs' and 'Ensembl genes'.

Features ▾ Comparative ▾ DA

- SNPs
- Genotyped SNPs
- Affy 100k SNP
- Affy 500k SNP
- Regulatory features
- CTCF
- tRNA
- CpG islands
- Eponine regions
- First EF
- Markers
- cisRED/miRANDA
- cisRED search regions
- Vega assembly
- Ensembl genes
- Ig segments
- Vega Havana gene
- Vega External gene
- ncRNA genes
- EST genes
- Genscan
- UniProtKB
- UniProtKB (mammal)
- UniProtKB (non-mammal)
- Human proteins

## RESULT OF STEP 10

The screenshot displays a genomic browser interface for a region on Chromosome 4. The top navigation bar includes menu items: Features, Comparative, DAS Sources, Repeats, Decorations, Export, Image size, and Help. Below this, there are input fields for 'Jump to region' (set to 4) and coordinates (123592080 to 123597339), with a 'Refresh' button. A 'Band' field is also present with another 'Refresh' button. Navigation controls include zoom in (+) and zoom out (-) buttons, a 'Zoom' slider, and window size buttons (<< 5MB, < 2MB, < 1MB, < Window, Window >, 1MB >, 2MB >, 5MB >>).

The main content area shows several tracks for the region from 123,593,000 to 123,597,000 (5.26 Kb):

- Chr. 4 Length:** Shows the forward and reverse strands.
- Conservation:** Displays 'Constrained elements' as grey bars.
- DNA(cortigs):** Shows a blue bar for contig <AC022489.7.1.195923.
- Human tilepath clones:** Shows a yellow bar for clone RP11-671P3.
- Ensembl trans.:** Shows gene models with exons and introns.
- SNPs:** Displays vertical lines representing single nucleotide polymorphisms. A yellow callout box points to one of these lines.
- Length:** Shows the physical length of the region.

A legend at the bottom left identifies track colors: brown for 'Common Known Protein coding', green for 'Synonymous coding', and blue for 'Intro'. A note states: 'There are currently 129 tracks switched off, use the Ensembl Homo sapiens version 49.36k (NCBI 36)'.

**STEP 11:** Click on a SNP (vertical line) and subsequently on 'SNP properties'

Your Ensembl report

- Login or Register
- About Users

dbSNP: rs1051753

- GeneSNP info
- rs1051753 - SNP info
- rs1051753 - LD info

Chromosome 4  
123,592,363

- View of Chromosome 4
- Graphical view
- Graphical overview
- Export from region...

Ensembl Archive

- View previous release of page in Archive!
- Stable Archive! link for this page

Sanger EMBL-EBI

dbSNP identifier

Alleles

dbSNP: rs1051753 (dbSNP128)	
HGVbase SNP000818875	
C/T (ambiguity code: Y) Ancestral allele: C	
Validation status	Unknown
Linkage disequilibrium data	No linkage data for this SNP
Flanking sequence	TAAATAGAAAGGCTGATATGTTTAAATGGGAAGCACTTAATTATCAAGTAACTGTTGAG ATGATGCTTTGACAAAAGGTAATCCATCTGTCAGAAATTC (SNP highlighted)

SNP rs1051753 is located in the following transcripts

Genomic location (strand)	Gene	Transcript relative SNP position	Translation relative SNP position	AA Type	GeneSNPView
4:123592363-123592363 (1)	ENSG00000109471	ENST00000226730: 750-750	ENSP00000226730: 152-152	L. SYNONYMOUS_CODING	SNPs in gene context

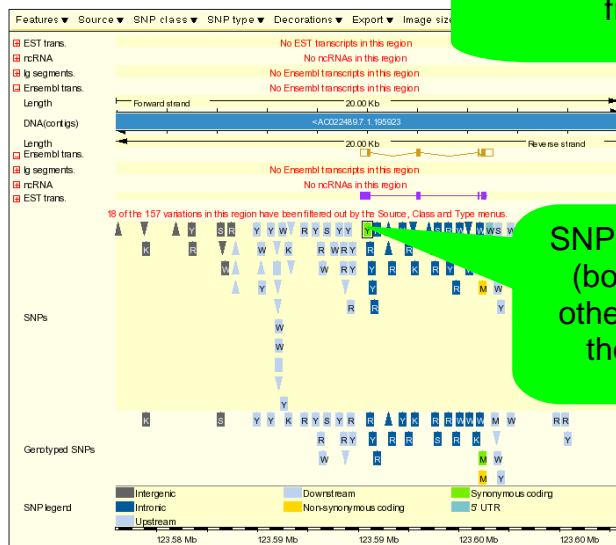
Population genotypes and allele frequencies

Population	Alleles	Alleles	Description
	C	T	
SEQUENOM/CEPH 1,000	unknown	unknown	composed of UTAH (93%), French (4%), and Venezuelan (3%) samples were purchased from, Genomic DNA samples were obtained for a panel of 92 unrelated individuals, Coriell Cell Repository and pooled in equimolar amounts.

Allele and genotype frequencies

Individual genotypes for SNP rs1051753

SNP Context - 4 123592363



SNP of interest (boxed) and other SNPs in the region

Orangutan  
*Pongo pygmaeus abelii*

Now in Ensembl!

**STEP 12:**  
Go back to ContigView with the back button of the internet browser.

**STEP 13:**  
 To see the same chromosomal region in the UCSC genome browser, click on 'View region at UCSC' on the left of the page. A new window will open.

The screenshot shows the UCSC Genome Browser interface for the IL2 gene region on chromosome 4. A green callout points to the 'IL2 gene' label. The interface includes navigation tools, track lists, and a 'refresh' button.

**STEP 14:**  
 Turn on 'Ensembl genes' by changing 'hide' to 'full' and clicking 'refresh' at the bottom of the page.  
 Close the window to return to ContigView.

Your Ensembl

- Login or Register
- About User Accounts

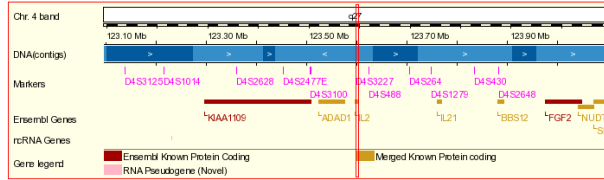
Chromosome 4  
123,592,080 - 123,597,339

- View of Chromosome 4
- Graphical view
- Graphical overview
- Resequencing alignment
- View alignment with ...
- View alongside ...
- View Syntenic regions ...
- View alignment at UCSC
- View alignment at NCBI

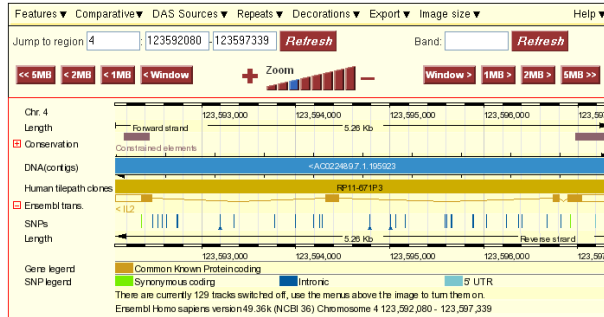
Chromosome 4



Overview



Detailed view



Basepair view

**STEP 15:**  
Click on 'View Syntenic regions ... with *Mus musculus*'



**Ensembl Human SyntenyView**  
 Ensembl release 49 - Mar 2008

**Your Ensembl**

- Login or Register
- About User Accounts

**Chromosome 4**

- View Chromosome 4
- View Chr 4 Synteny
- Map your data onto this chromosome

**Ensembl Archive**

- View previous release of page in Archive!
- Stable Archive! link for this page

**Orangutan**  
*Pongo pygmaeus abelii*  
 Now in Ensembl

**Mouse chromosomes**

**Human genes**

**Human chromosome**

**Syntenic block**

**Mouse homologues**

**STEP 16: Click on [MultiContigView]**

**Homology Matches**

Homo sapiens Genes	Mus musculus Homologues
<a href="#">IL2</a> (0.12 Gb) [ContigView]	-> <a href="#">IL2</a> (3: 37.02 Mb) [ContigView] [MultiContigView]
<a href="#">IL21</a> (0.12 Gb) [ContigView]	-> <a href="#">IL21</a> (3: 37.12 Mb) [ContigView] [MultiContigView]
<a href="#">BBS12</a> (0.12 Gb) [ContigView]	-> <a href="#">Bbs12</a> (3: 37.21 Mb) [ContigView] [MultiContigView]
<a href="#">FGF2</a> (0.12 Gb) [ContigView]	-> <a href="#">Fgf2</a> (3: 37.25 Mb) [ContigView] [MultiContigView]
<a href="#">NUDT6</a> (0.12 Gb) [ContigView]	-> <a href="#">Nudt6</a> (3: 37.30 Mb) [ContigView] [MultiContigView]
<a href="#">SPATA5</a> (0.12 Gb) [ContigView]	-> <a href="#">Spata5</a> (3: 37.37 Mb) [ContigView] [MultiContigView]
<a href="#">SPRY1</a> (0.12 Gb) [ContigView]	-> No homologues
<a href="#">ANKRD50</a> (0.13 Gb) [ContigView]	-> No homologues
<a href="#">BLYM_HUMAN</a> (0.13 Gb) [ContigView]	-> No homologues
<a href="#">ENSG00000214786</a> (0.13 Gb) [ContigView]	-> No homologues
<a href="#">FAT4</a> (0.13 Gb) [ContigView]	-> <a href="#">Fat4</a> (3: 38.79 Mb) [ContigView] [MultiContigView]
<a href="#">INTU</a> (0.13 Gb) [ContigView]	-> <a href="#">Intu</a> (3: 40.43 Mb) [ContigView] [MultiContigView]
<a href="#">SLC25A31</a> (0.13 Gb) [ContigView]	-> <a href="#">Slc25a31</a> (3: 40.51 Mb) [ContigView] [MultiContigView]
<a href="#">HSPA4L</a> (0.13 Gb) [ContigView]	-> <a href="#">Hspa4l</a> (3: 40.55 Mb) [ContigView] [MultiContigView]
<a href="#">PLK4</a> (0.13 Gb) [ContigView]	-> <a href="#">Plk4</a> (3: 40.60 Mb) [ContigView] [MultiContigView]

**Navigate Homology**

[Upstream \(<0.12 Gb\)](#) [Downstream \(>0.13 Gb\)](#)

**Change Chromosome**

Chromosome

Fields marked with \* are required

Human and mouse side-by-side.

**Your Ensembl**

- Login or Register
- About User Accounts

**Chromosome 4**  
123,591,080 - 123,598,339

- View of Chromosome 4
- Graphical view of...
- Graphical overview
- Resequencing alignment
- View alignment with ...
- View alongside ...
- View Syntenic regions ...
- View region at UCSC
- View region at NCBI

**Export data**

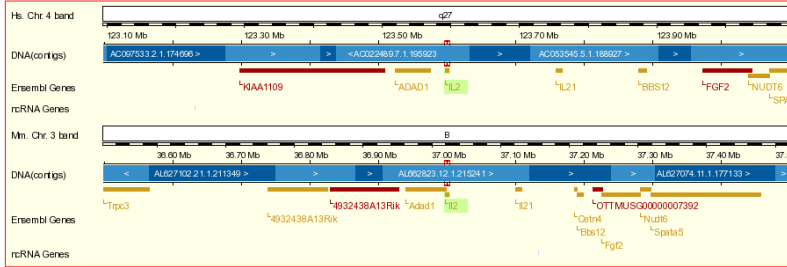
- Export from region...

Ensembl Archive

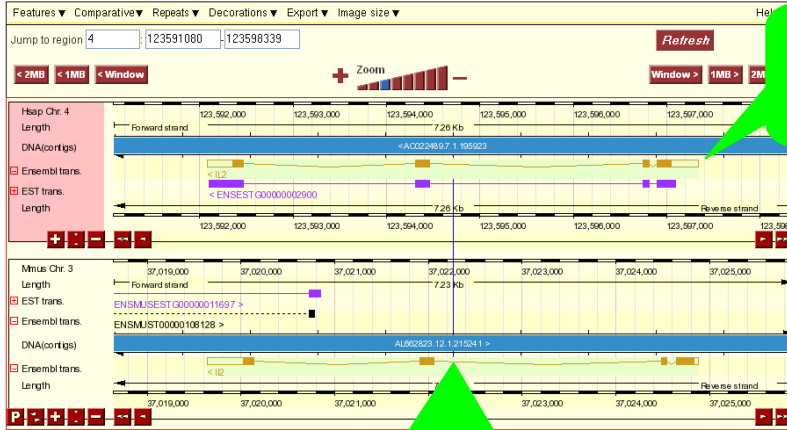
Top level



Navigational overview



Detailed View



**STEP 17:**  
Click on 'Export from region'

IL2 gene (Human)

IL2 homologue in mouse

## Select region/feature to Export

Choose at least one feature to export. Features must map to the current Ensembl Golden tile path. *Please note we will not export more than 5Mb.*

### Region

Chromosome name/fragment

From (type):

To (type):

### Context

Bp upstream (to the left)

Bp downstream (to the right)

### Output

Output

[Continue >>](#)

**STEP 18:**  
Click on  
[Continue>>]

### Configure FASTA File output for FASTA sequence

You are exporting Chromosome 4 123,591,080 - 123,598,339.

This region is defined by: Chromosome 4, Bp 123591080, Bp 123598339

Output format  HTML

Text

Compressed text (.gz)

[Continue >>](#)

**STEP 19:**  
Click on  
[Continue>>]

© 2008 [WTSI](#) / [EBI](#). Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

**STEP 20:**  
Select and copy a part of the sequence

**STEP 21:**  
Click on 'BLAST'

Ensembl release 49

Your Ensembl

- Login or Register
- About User Accounts

Chromosome 4  
123,591,080 - 123,598,339

- View of Chromosome 4
- Graphical view
- Graphical overview
- Resequencing alignment
- View alignment with ...
- View alongside ...
- View Syntenic regions ...
- View region at UCSC
- View region at NCBI

Export data

- Export from region...

Ensembl Archive

- View previous release of page in Archive!
- Stable Archive! link for this page

sanger EMBL-EBI

Horse  
*Equus caballus*

Now in Ensembl!

Ensembl Human Blast View

Search: #Human EBI

e.g. AL138722.15.1.44776, ENSG00000139818

HOME · BLAST · BIOMART · SITEMAP · HELP

**STEP 22:**  
Paste the copied sequence

**STEP 23:**  
Select 'Homo\_sapiens' and 'BLASTN' and click on [RUN>]

Ensembl release 49

Your Ensembl

- Login or Register
- About User Accounts

sanger EMBL-EBI

Microbat  
*Myotis lucifugus*

2X assembly and genebuild

Ensembl Human Blast View

Search: #Human EBI

e.g. AL138722.15.1.44776, ENSG00000139818

HOME · BLAST · BIOMART · SITEMAP · HELP

SETUP CONFIG RESULTS DISPLAY refresh Online Help

Sequence (30 sequences) in FASTA or plain text

Or Upload a file containing one or more FASTA sequences

Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)

Or Enter an existing ticket ID:

Select the databases to search against

Select species:

Use 'ctrl' key to select multiple species

dna database

peptide database

Select the Search Tool

BLASTN

SSAHA2

TBLASTX

Search sensitivity: Near-exact matches

About BlastView

BlastView provides an integrated platform for sequence similarity searches against Ensembl databases, offering access to both BLAST and SSAHA programs. [More]

We would like to hear your impressions of BlastView, especially regarding functionality that you would like to see provided in the future. Many thanks for your time. [Feedback]

© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

Your Ensembl

- Login or Register
- About User Accounts

sanger



new SETUP CONFIG RESULTS DISPLAY refresh Online Help

Retrieve result for ID:  
[input: BLA\_APLFannCO] [button: Retrieve]

Retrieving Results

Job pending' results can be retrieved by clicking on the button above. Alternatively, this page can be bookmarked for later, or the ID noted and entered on the BLAST page. Results are retained for 7 days. After this, they must be re-submitted.

1: unnamed (3300 letters) Vs. LATESTGP

Hom sapiens	Job Queued
-------------	------------

Summary of BLAST search

STEP 24: Click on 'Retrieve'

new SETUP CONFIG RESULTS DISPLAY refresh Online Help

Retrieve result for ID:  
[input: BLA\_APLFannCO] [button: Retrieve]

Alignment Display Options:

- Locations vs. Karyotype
- Locations vs. Query
- Summary Table

1: unnamed (3300 letters) Vs. LATESTGP

Hom sapiens	147 alignments, 52 hits	[RawResult]	[button: view ▶]
-------------	-------------------------	-------------	------------------

STEP 25: Click on [VIEW]

Your Ensembl

- Login or Register
- About User Accounts

Ensembl BLOG

Keep up-to-date with news, views & announcements from the Ensembl team

new SETUP CONFIG RESULTS DISPLAY

refresh Online Help

Summary

- setup
  - Homo\_sapiens
  - Genomic sequence
  - BLASTN
  - Low sensitivity
- configure
  - E: 10
  - B: 100
  - filter: dust
  - RepeatMasker
  - W: 15
  - M: 1
  - N: 3
  - Q: 3
  - R: 3
- results
- display

Not yet initialised

Displaying 4 sequence alignments vs Homo\_sapiens LATESTGP database

Showing top 100 alignments of 5257, sorted by Raw Score

Genotype (click arrow to hide)

Best hit (boxed)

Location of hits on the genome

Alignment of hits to query sequence

Alignment Locations vs. Query (click arrow to hide)

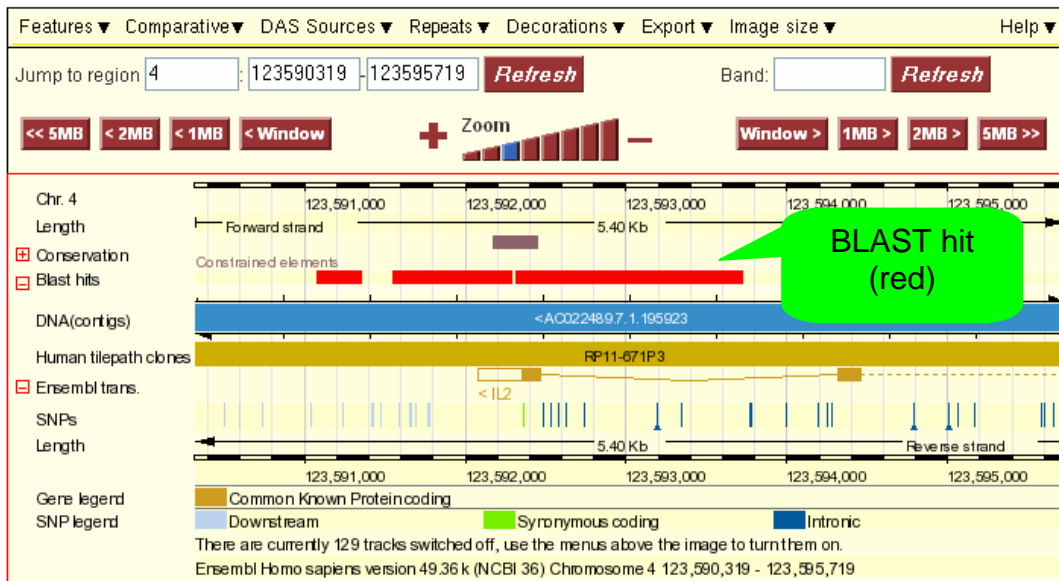
Alignment Summary (click arrow to hide)

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	off_ Name	off_ Name	off_ Name	off_ Name	off_ Name	off_ Name	off_ Score	>Chromosome
Start	Start	Start	Start	Start	Start	Start	E-val	<Score
[A] [S] [G] [C]	1240 2640	+	Chr.4	123692319	123693719	+	1401 0.	100.00 1401
[A] [S] [G] [C]	470 1206	+	Chr.4	123691549	123692286	+	737 0.	100.00 737
[A] [S] [G] [C]	1 272	+	Chr.4	123691080	123691351	+	272 0.	100.00 272
[A] [S] [G] [C]	193 233	+	Chr.8	124375530	124375570	+	30 1.4e-05	92.86 42
[A] [S] [G] [C]	185 224	+	Chr.2	71524316	71524355	+	28 0.024	92.50 40
[A] [S] [G] [C]	1124 1152	+	Chr.15	95492489	95492528	+	28 1.2	96.88 32
[A] [S] [G] [C]	1395 1424	+	Chr.6	130524130	130524160	+	23 1.1e-05	93.55 31
[A] [S] [G] [C]	2301 2323	+	Chr.2	240016107	240016129	+	23 0.0036	100.00 23
[A] [S] [G] [C]	47 73	-	Chr.2	237422040	237422068	+	23 0.0017	96.30 27
[A] [S] [G] [C]	980 1002	+	Chr.1	110989000	110989622	+	23 0.73	100.00 23
[A] [S] [G] [C]	1058 1083	+	Chr.6	127562106	127562131	+	22 1.1e-05	96.15 28
[A] [S] [G] [C]	916 948	+	Chr.6	49496575	49496607	+	22 1.1e-05	91.18 34
[A] [S] [G] [C]	2294 2324	+	Chr.5	89471051	89471096	+	22 0.10	96.96 46
[A] [S] [G] [C]	532 553	+	Chr.5	71467086	71467087	+	22 3.1	100.00 22
[A] [S] [G] [C]	896 1021	-	Chr.11	100577751	100577775	+	22 0.24	96.15 26
[A] [S] [G] [C]	2294 2314	+	Chr.4	172121719	172121739	+	21 0.	100.00 21
[A] [S] [G] [C]	1398 1419	+	Chr.4	147602722	147602742	+	21 0.	100.00 21
[A] [S] [G] [C]	1407 1431	-	Chr.4	156363178	156363202	+	21 0.0012	96.00 25
[A] [S] [G] [C]	1391 1415	-	Chr.4	31837404	31837428	+	21 5.4	96.00 25
[A] [S] [G] [C]	1397 1417	+	Chr.6	94510145	94510165	+	21 1.2	100.00 21
[A] [S] [G] [C]	176 195	+	Chr.8	19840422	19840442	+	21 0.00047	100.00 21
[A] [S] [G] [C]	1936 2020	-	Chr.8	97852565	97852588	+	21 0.00012	96.00 25
[A] [S] [G] [C]	1 25	+	Chr.2	32023441	32023484	+	21 0.0036	96.00 25
[A] [S] [G] [C]	1398 1418	+	Chr.2	181715913	181715933	+	21 0.0017	100.00 21
[A] [S] [G] [C]	2332 2356	+	Chr.5	142264541	142264564	+	21 0.030	96.00 25
[A] [S] [G] [C]	1635 1659	+	Chr.5	57984528	57984552	+	21 0.0060	96.00 25
[A] [S] [G] [C]	1614 1637	+	Chr.3	67261928	67261952	+	21 0.0081	96.00 25
[A] [S] [G] [C]	1973 1997	+	Chr.13	68769192	68769205	+	21 0.0100	96.00 25
[A] [S] [G] [C]	1393 1413	+	Chr.13	66043919	66043939	+	21 4.2	100.00 21
[A] [S] [G] [C]	1633 1657	-	Chr.13	35117385	35117409	+	21 4.2	96.00 25

**STEP 26:**  
Click on [C] in front of best (top) hit

Back in the contigview page...



**END of the  
 Worked Example**

## EXERCISES and ANSWERS

Note: the answers to these exercises correspond to current version (49) of Ensembl. If you use these exercises at a later date, please use the archive site for version 49.

### 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)). This is the current version. Search for the human TAC1 gene by typing 'human TAC1 gene' in the search window.

(b) How many transcripts are predicted 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 'Similarity Matches' section of GeneView. What is a possible function of TAC1?

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

(e) Find the GO section of GeneView and follow some of the links to explore the 'Gene ontology' terms (describing gene and protein function) in Ensembl GOView.

(f) In which chromosomal band and on which clone and contig in the genomic sequence assembly is the TAC1 gene located?

(g) Go back to GeneView by clicking on 'TAC1' in the Overview panel and following the link for the gene. Is there a putative mouse orthologue? If so, where is it in the mouse genome?

(h) Go to the Ensembl main page. Look up 'Ensembl genes' in the glossary by following the 'Using Ensembl' link in the 'Help and Documentation' at the left. Follow the link: 'Browse the provided data' on the website.

[http://www.ensembl.org/Homo\\_sapiens/glossaryview](http://www.ensembl.org/Homo_sapiens/glossaryview)

## 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. Go to the human homepage, and click on chromosome 12. From 'MapView', choose to display the region between markers D12S764 and D12S1871 (in the 'Jump to ContigView' section).

(b) How many contigs are used to make this portion of the assembly? View the human tile path clones. Do they correspond to the assembly?

(c) Click on a marker. What are other names for the marker? Is there an expected product size (what does this mean?)

(d) In ContigView, zoom in three steps on the zoom triangle/ladder of 'Detailed view' (towards the '+') and turn on the SNP track. Identify an intronic SNP and look at the corresponding SNPView page.

## 3. Exploring the zebrafish (*Danio rerio*) genome with Ensembl

*Exercise 3 explores Ensembl genes for a specific region, along with comparisons to mRNA and protein in other databases.*

(a) What assembly is currently used for zebrafish?

(b) Bring up a ContigView display of zebrafish (*Danio rerio*) chromosome 1 between base pairs 3900000 and 3910000. There are many 'known' genes. What is the difference between known and novel genes in Ensembl? For any gene, find some information about its function by taking the link to the GeneView page.

(c) View homologues in other species for the Zebrafish gene you have selected. Take the link to FamilyView. Go back to GeneView and look for any InterPro domains listed at the bottom, or in the 'Protein features' diagram.

## 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 GeneView page.

(b) The TAC1 gene (ENSG00000006128) has 3 predicted transcripts, ENST00000319273, ENST00000346867 and ENST00000350485. Scroll down to the 'Transcript' sections 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 is Protachykinin 1 precursor. Follow the links to MIM and EntrezGene or UniProt/Swiss-Prot in the 'Similarity Matches' section to learn more. Choose 'UniProt' under 'DAS Sources' to see references in the literature (click 'Update' after making the selection). Also the GO (Gene Ontology) and InterPro sections can give you clues about the biological and molecular function of the TAC1 protein. Tachikinins 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) Check the 'InterPro' section in GeneView. The domains include IPR013055 (Tachykinin/Neurokinin like), IPR002040 (Tachykinin/Neurokinin), IPR008215 (Tachykinin) and IPR008216 (Protachykinin).

(e) Clicking on a GO identifier gives you a GOView page showing the position of that term in the hierarchical 'GO tree' (note the number of Ensembl genes mapped to each term). Click [Help] to find out more about GOView.

(f) Go back to GeneView and click the 'Graphical View' link in the side menu to go to ContigView. In the 'Overview' panel you can see that TAC1 is located on band 7q21.3 ('Chr.7 band' track). In the 'Detailed view' panel you can see that it is located on contig AC004140.2.1.74918 ('DNA(contigs)' track). If you click on the contig and follow the link to the EMBL source (or if you turn on the 'Human tile path clones' track from the 'Decorations' menu of ContigView) you can see that this sequence is derived from clone RP5-841B21.

(g) In GeneView, ENSMUSG00000061762 (Tac1) is named in the 'Orthologue Prediction' section. Click on it to go to its GeneView page to find that it is located on mouse chromosome 6 (band A1).

## 2. Exploring a region

(a) In the 'Jump to ContigView' section choose 'From (type): Marker D12S764 To (type): Marker D12S1871' and click [Go]. This leads you to ContigView.

(b) The region between the two markers will be displayed in 'Detailed View'. This region includes sequence from 4 different contigs (one is quite small), displayed in light blue and dark blue in the 'DNA(contigs)' track. To see also the 4 clones that make up this region, select the 'Human tilepath clones' track from the 'Decorations' menu. Clones are shown in gold and pink. Portions of the 'Tile path clones' were used to form the assembly and correspond to 'contigs'. The clones overlap each other whereas the contigs don't.

(c) Click on a marker name (shown in pink at the top) and follow the link 'Marker info' to the MarkerView page. Other names in the UniSTS database will be at the top. Synonyms (or names in other databases) are listed further

down. The expected product size is the calculated size of the fragment (in base pairs) if both primers are used against the genome.

(d) SNPs can be turned on using the 'Features' menu. Coding SNPs are shown in yellow (non-synonymous) and green (synonymous). Intronic SNPs are dark blue. Click on a SNP. Be careful to click exactly on the vertical bar representing the SNP, otherwise you will get the wrong pop-up menu. Follow the link 'SNP properties' to the SNPView page. Note the 'SNP Context' display in SNPView.

### **3. Exploring the zebrafish (*Danio rerio*) genome with Ensembl**

(a) Zv7 is the current assembly (see this in the main page or by clicking on z-fish to go to the species-specific index page).

(b) Start on the homepage for zebrafish (*Danio rerio*). In the 'Karyotype' section choose 'Chromosome: 1', 'From (type): Base pair: 3900000 To (type) Base pair: 3910000' and click [Go].

On the 'Overview' panel of ContigView Ensembl known and novel genes are displayed in the 'Ensembl Genes' track in reddish brown and black, respectively. Known genes are Ensembl gene predictions that match species specific entries in the UniProt and/or RefSeq database, while novel genes map back to entries from other species. In this region, they are all known genes. Click on one of the known genes to go to its GeneView page and explore the links in the 'Similarity Matches' section for information about function.

(c) In the GeneView page: if the gene has orthologues in other species (shown in the 'Orthologue Prediction' section) and these orthologues are well-understood, possible functions of this gene could be hypothesised. If the gene belongs to a family (shown in the 'Protein Family' section) other family members may provide information. InterPro domains (shown in the 'InterPro' section at the bottom of GeneView) may also provide clues.

## IV) Data mining in Ensembl with BioMart 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:**  
Go to the Ensembl main page  
[www.ensembl.org](http://www.ensembl.org)

The screenshot shows the Ensembl website interface. At the top, the Ensembl logo is on the left, and navigation links (HOME, BLAST, BIOMART, MITEMAP, HELP) are on the right. The main content area is divided into several sections:

- Your Ensembl:** Includes links for Login or Register and About User Accounts.
- Help & Documentation:** Includes links for About Ensembl, Genomic Data, Help & Information, and Software.
- Ensembl Archive:** Includes links for View previous release of page in Archive! and Stable Archive! link for this page.
- Search Ensembl:** A search bar with a dropdown menu set to "All species" and a "Go" button. Below the search bar, there is a text input field with the example text "e.g. mouse chromosome 2 or rat X:10000..20000 or human gene BRCA2".
- Ensembl tools:** A section with four icons and text boxes:
  - Start a sequence search:** Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.
  - Mine Ensembl with BioMart:** Extract information from the Ensembl database and export sequences or tables in text, html, or Excel format with BioMart.
  - Customise Your Ensembl:** Register your Ensembl to bookmark your favourite pages, customise your homepage and much more!
  - Help:** Get help with this site.
- About Ensembl:** A section with text describing the project and providing contact information for the Ensembl HelpDesk.
- Other Ensembl websites:** A list of links to other Ensembl-related resources, including archive, Vega, Ensembl Pre!, EBI Genome Reviews, and Trace server.
- Ensembl 47:** A section with a "Pre! species" button and a list of popular genomes (Human, Mouse, Zebrafish) with links to customize the view.
- Ensembl headlines:** A section with a "Release 47 (October 2007)" header and a list of news items, including "New Mouse 37 assembly and genebuild", "New genebuild on human assembly NCBI 36", and "WormBase 180".

A blue box with a white background and a blue border is overlaid on the "Ensembl tools" section, containing the text "STEP 2: Click on 'BioMart'". A blue arrow points from the top of this box to the "Mine Ensembl with BioMart" icon.

At the bottom of the page, there is a copyright notice: "© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details."

New Count Results XML Perl Help

Dataset  
[None selected]

Ensembl 47

- CHOOSE DATASET -

**STEP 3:**  
Select the database:  
Ensembl genes (version 49)  
and the species of interest  
under 'Choose Dataset'.  
(*Homo sapiens*)

New Count Results XML Perl Help

Please restrict your query using criteria below

Dataset  
Homo sapiens genes (NCBI36)

Filters  
[None selected]

Attributes  
ensembl Gene ID  
ensembl Transcript ID

REGION:  
GENE:  
GENE ONTOLOGY:  
EXPRESSION:  
MULTI SPECIES COMPARISONS:  
PROTEIN:

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

New Count Results XML Perl Help

Please restrict your query using criteria below

Dataset  
Homo sapiens genes (NCBI36)

Filters  
Chromosome: X  
Start: q28  
End: q28

Attributes  
ensembl Gene ID  
ensembl Transcript ID

Dataset  
[None Selected]

REGION:  
 Chromosome X  
 Base pair  
Gene Start (bp) 1  
Gene End (bp) 10000000  
 Band  
Start q28  
End q28  
 Marker  
Start  
End  
 Encode type manual\_picks  
 Encode region

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

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

The screenshot shows a web interface with a top navigation bar containing 'New', 'Count', and 'Results' tabs. The 'Count' tab is circled in red. Below the navigation bar, there is a left sidebar with 'Dataset' (Homo sapiens genes (NCBI36)), 'Filters' (Chromosome: X, Start: q28, End: q28, with Disease association: Only), and 'Attributes' (Ensembl Gene ID, Ensembl Transcript ID). The main area shows a search results page with a 'Browse...' button and a 'GENE' panel. The 'GENE' panel is expanded, and the 'ID LIST FILTERS' section is checked. The 'with Disease association' filter is selected, and the 'Only' radio button is chosen. The 'Ensembl Gene' dropdown is also visible.

**STEP 7:**  
 Expand the 'GENE' panel and choose 'with Disease Association only'. These associations have been determined using MIM (Online Mendelian Inheritance in Man). [www.ncbi.nlm.nih.gov/omim/](http://www.ncbi.nlm.nih.gov/omim/)

The filters have determined our gene set. Click 'Count' (at the top) to see how many genes have passed these filters.

The screenshot shows the same web interface, but now the 'Attributes' panel is expanded, showing options for 'Gene ID' and 'Transcript ID'. The 'GENE' panel is also expanded, showing options for 'EXTERNAL:', 'EXPRESSION:', 'PROTEIN:', and 'GENOMIC REGION:'. The 'Features' radio button is selected. The 'Count' tab is still highlighted in red.

**STEP 8:**  
 Click on 'Attributes' to select output options (i.e. what we would like to know about our geneset).

**STEP 9:**  
 Expand the 'GENE' panel.

New Count Results XML Perl Help

Dataset 25 / 32584 Genes

**Filters**

Chromosome: X  
Start : q28  
End : q28  
with Disease association: Only

**Attributes**

Ensembl Gene ID  
Ensembl Transcript ID  
Gene name

**Dataset**

[None Selected]

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

Features  Homologs  
 Structures  Sequences  
 SNPs

GENE:

**Ensembl Attributes**

Ensembl Gene ID  
 Ensembl Transcript ID  
 Ensembl Peptide ID  
 Description  
 Chromosome Name  
 Gene Start (bp)  
 Gene End (bp)  
 Strand  
 Band  
 Transcript Start (bp)  
 Transcript End (bp)  
 Gene name

Gene DB  
 External Transcript ID  
 External Transcript DB  
 Ensembl CDS length  
 Ensembl cDNA length  
 Ensembl Peptide length  
 Transcript count  
 % GC content  
 Biotype  
 Source  
 Status (gene)  
 Status (transcript)

Note the summary of selected options.

The order of attributes determines the order of columns in the result table.

**STEP 10:**  
Select, along with the default options, 'Gene name' (this shows the gene symbol from HGNC).

New Count Results XML Perl Help

Dataset 25 / 32584 Genes

**Filters**

Chromosome: X  
Start : q28  
End : q28  
with Disease association: Only

**Attributes**

Ensembl Gene ID  
Ensembl Transcript ID  
Gene name

**Dataset**

[None Selected]

EXTERNAL:

**GO Attributes**

GO ID  
 GO description  
 GO evidence code

**External References (max 3)**

CCDS ID  
 EMBL (Genbank) ID  
 EntrezGene ID  
 Havana ID  
 Havana transcripts (Shared) ID  
 Havana transcripts (Identical) ID  
 HGNC ID  
 HGNC symbol  
 IPI ID  
 Imgt gene db  
 Imgt ligm db  
 Mim Gene Accession  
 Mim Morbid accession  
 Mirbase  
 OTTP

**Microarray Attributes (max 2)**

Affy UGC 110  
 Affy UGC 1055

**STEP 12:**  
Click 'RESULTS' at the top to preview the output.

**STEP 11:**  
Expand the 'EXTERNAL' panel to select External References (IDs outside of Ensembl). Select 'EntrezGene ID' and 'Mim Gene Accession' and 'Mim Morbid Accession'. These are MIM phenotypes and diseases, respectively.  
[www.ncbi.nlm.nih.gov/omim/](http://www.ncbi.nlm.nih.gov/omim/)

New Count Results XML Perl Help

Dataset 25 / 32584 Genes

Export all results to: File TSV Unique results only Go

Email notification to: [input]

View: 10 rows as HTML Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Gene name	EntrezGene ID	Mim Gene Accession
<a href="#">ENSG00000102081</a>	<a href="#">ENST00000370475</a>	FMR1	2332	<a href="#">309550</a>
<a href="#">ENSG00000102081</a>	<a href="#">ENST00000370475</a>	FMR1	2332	<a href="#">309550</a>
<a href="#">ENSG00000155966</a>	<a href="#">ENST00000370460</a>	AFF2	2334	<a href="#">309548</a>
<a href="#">ENSG00000010404</a>	<a href="#">ENST00000340855</a>	IDS	3423	<a href="#">309900</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000370401</a>	CXorf6	10046	<a href="#">300120</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000370401</a>	CXorf6	728030	<a href="#">300120</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000370401</a>	CXorf6	730818	<a href="#">300120</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000262858</a>	CXorf6	728030	<a href="#">300120</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000262858</a>	CXorf6	730818	<a href="#">300120</a>
<a href="#">ENSG00000171100</a>	<a href="#">ENST00000306167</a>	MTM1	4534	<a href="#">300415</a>

Attributes: Ensembl Gene ID, Ensembl Transcript ID, Gene name, EntrezGene ID, Mim Gene Accession, Mim Morbid accession

Dataset: [None Selected]

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

**STEP 13:**  
Go back and change Filters or Attributes if desired. Or, View 'ALL' as HTML...

**Result Table 1**

Ensembl Gene ID	Ensembl Transcript ID	Gene name	EntrezGene ID	Mim Gene Accession	Mim Morbid accession
<a href="#">ENSG00000102081</a>	<a href="#">ENST00000370475</a>	FMR1	2332	<a href="#">309550</a>	<a href="#">300623</a>
<a href="#">ENSG00000102081</a>	<a href="#">ENST00000370475</a>	FMR1	2332	<a href="#">309550</a>	<a href="#">300624</a>
<a href="#">ENSG00000155966</a>	<a href="#">ENST00000370460</a>	AFF2	2334	<a href="#">309548</a>	<a href="#">309548</a>
<a href="#">ENSG00000010404</a>	<a href="#">ENST00000340855</a>	IDS	3423	<a href="#">309900</a>	<a href="#">309900</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000370401</a>	CXorf6	10046	<a href="#">300120</a>	<a href="#">300633</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000370401</a>	CXorf6	728030	<a href="#">300120</a>	<a href="#">300633</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000370401</a>	CXorf6	730818	<a href="#">300120</a>	<a href="#">300633</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000262858</a>	CXorf6	728030	<a href="#">300120</a>	<a href="#">300633</a>
<a href="#">ENSG00000013619</a>	<a href="#">ENST00000262858</a>	CXorf6	730818	<a href="#">300120</a>	<a href="#">300633</a>
<a href="#">ENSG00000171100</a>	<a href="#">ENST00000306167</a>	MTM1	4534	<a href="#">300415</a>	<a href="#">310400</a>
<a href="#">ENSG00000147383</a>	<a href="#">ENST00000370274</a>	NSDHL	50814	<a href="#">300275</a>	<a href="#">308050</a>
<a href="#">ENSG00000130821</a>	<a href="#">ENST00000253122</a>	SLC6A8	6535	<a href="#">300036</a>	<a href="#">300352</a>
<a href="#">ENSG00000130821</a>	<a href="#">ENST00000253122</a>	SLC6A8	731026	<a href="#">300036</a>	<a href="#">300352</a>
<a href="#">ENSG00000185825</a>	<a href="#">ENST00000345046</a>	BCAP31	10134	<a href="#">300398</a>	<a href="#">300475</a>
<a href="#">ENSG00000185825</a>	<a href="#">ENST00000370133</a>	BCAP31	10134	<a href="#">300398</a>	<a href="#">300475</a>
<a href="#">ENSG00000101986</a>	<a href="#">ENST00000218104</a>	ABCD1	215	<a href="#">300371</a>	<a href="#">300100</a>
<a href="#">ENSG00000101986</a>	<a href="#">ENST00000218104</a>	ABCD1	642762	<a href="#">300371</a>	<a href="#">300100</a>
<a href="#">ENSG00000101986</a>	<a href="#">ENST00000218104</a>	ABCD1	215	<a href="#">300371</a>	<a href="#">300475</a>
<a href="#">ENSG00000101986</a>	<a href="#">ENST00000218104</a>	ABCD1	642762	<a href="#">300371</a>	<a href="#">300475</a>
<a href="#">ENSG00000198910</a>	<a href="#">ENST00000370058</a>	L1CAM	3897	<a href="#">308840</a>	<a href="#">142623</a>
<a href="#">ENSG00000198910</a>	<a href="#">ENST00000370058</a>	L1CAM	3897	<a href="#">308840</a>	<a href="#">303350</a>
<a href="#">ENSG00000198910</a>	<a href="#">ENST00000370058</a>	L1CAM	3897	<a href="#">308840</a>	<a href="#">304100</a>
<a href="#">ENSG00000198910</a>	<a href="#">ENST00000370058</a>	L1CAM	3897	<a href="#">308840</a>	<a href="#">307000</a>
<a href="#">ENSG00000198910</a>	<a href="#">ENST00000370060</a>	L1CAM	3897	<a href="#">308840</a>	<a href="#">142623</a>

**STEP 14:**  
To view sequences, go  
back to 'Attributes'

Chromo  
Start : q  
End : q2  
with Disease association: Only

**Attributes**

Ensembl Gene ID  
Ensembl Transcript ID  
External Gene ID  
EntrezGene ID  
Mim Gene Accession

---

**Dataset**  
[None Selected]

XML Perl Help

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

Features  Homologs  
 Structures  Sequences  
 SNPs

**GENE:**

**Ensembl Attributes**

Ensembl Gene ID  
 Ensembl Transcript ID  
 Ensembl Peptide ID  
 Description  
 Chromosome Name  
 Gene Start (bp)  
 Gene End (bp)  
 Strand  
 Band  
 Transcript Start (bp)  
 Transcript End (bp)  
 External Gene ID

Ensembl Peptide length  
 Transcript count  
 % GC content  
 Biotype  
 Source  
 Status (gene)  
 Status (transcript)

**EXTERNAL:**

**GO Attributes**

GO ID  
 GO description  
 GO evidence code

**External References (max 3)**

CCDS ID  
 Codelink ID  
 EMBL ID  
 EntrezGene ID  
 Havana ID  
 HGNC Symbol  
 Illumina v1  
 Illumina v2  
 IPI ID  
 Imgt gene db  
 Imgt ligm db  
 Mim Gene Accession  
 Mim Morbid accession

Protein ID  
 RefSeq DNA ID  
 RefSeq Predicted DNA ID  
 RefSeq Peptide ID  
 Rfam ID  
 Unigene ID  
 Shares cds with enst  
 Shares cds with ott  
 UniProt/SPTREMBL ID  
 UniProt/Swiss-Prot ID  
 UniProt/Swiss-Prot Accession  
 Unified UniProt ID  
 Unified UniProt Accession

**STEP 15:**  
Select 'Sequences'

New Count Results XML Perl Help

**Dataset** 25 / 32584 Genes

**Filters**

Chromosome: X  
Start : q28  
End : q28  
with Disease association: Only

**Attributes**

Ensembl Gene ID  
Chromosome  
Biotype  
cDNA sequences

---

**Dataset**  
[None Selected]

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

Features  Homologs  
 Structures  Sequences  
 SNPs

**SEQUENCES:**

**Sequences (max 1)**

-----[ ]-----  
-----[ ]-----

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

5' UTR  
 3' UTR  
 Exon sequences  
 cDNA sequences  
 Coding sequence  
 Peptide

**Upstream flank**  
 Upstream flank [ ]

**Downstream flank** [ ]

**STEP 16:**  
Expand the 'SEQUENCES' panel and  
select 'cDNA sequences'.

**STEP 17:**  
Expand the 'Header Information' to view the default options: 'Ensembl Gene ID', 'Chromosome', and 'Biotype' are selected.

The screenshot shows the Ensembl interface with the 'Header Information' section expanded. The 'Gene Attributes' section has the following options checked: Ensembl Gene ID, Chromosome, and Biotype. The 'Transcript Attributes' section has the following options checked: Ensembl Transcript ID, Ensembl Peptide ID, Transcript Start (bp), and Transcript End (bp). The 'Exon Attributes' section has the following options checked: Exon Start (bp), Exon End (bp), and Exon Strand. The 'UTR Attributes' section has the following options checked: 5 UTR Start (bp), 5 UTR End (bp), 3 UTR Start (bp), and 3 UTR End (bp). The 'Coding Attributes' section has the following options checked: Ensembl CDNA Start (bp), Ensembl CDNA End (bp), Ensembl CDS Start (bp), Ensembl CDS End (bp), Coding Start (bp), and Coding End (bp). The 'Dataset' section shows 25 / 32584 Genes.

**STEP 18:**  
Click 'Results'

The screenshot shows the Ensembl interface with the 'Export' section. The 'Export' section has the following options: 'File' (selected), 'FASTA' (selected), and 'Unique results only' (unchecked). The 'Email notification to' field is empty. The 'View' section has the following options: '10' (selected), 'rows as' (selected), 'FASTA' (selected), and 'Unique results only' (unchecked). The 'Dataset' section shows 25 / 32584 Genes.

**STEP 19:**  
'View all rows as FASTA'

## RESULTS

### Header: chromosome, Ensembl Gene ID, Biotype

```
>ENSG00000196924|X|protein_coding
GCGCGTCGCGCGCAGCGGACGCCGACAGAATCCCCGAGGCGCCTGGCGCGGGCGCGGGCG
CGAAGGCGATCCGGGCGCCACCCCGGGTCATCGGTCACCGGTCGCTCTCAGGAACAGCA
GCGCAACCTCTGCTCCCTGCCTCGCTCCCGCGCGCTAGGTGCCTGCGACTTTAATTAA
AGGGCCGTCCCCCTCGCCGAGGCTGCAGCACCGCCCCCGGCTTCTCGCGCCTCAAAATG
AGTAGCTCCCACTCTCGGGCGGGCCAGAGCGCAGCAGGCGCGGCTCCGGGCGGCGGCGTC
GACACGCGGGGACGCCGAGATGCCGGCCACCGAGAAGGACCTGGCGGAGGACGCGCCGTGG
AAGAAGATCCAGCAGAACAACCTTTACGCGCTGGTGCAACGAGCACCTGAAGTGCGTGAGC
AAGCGCATCGCCAACCTGCAGACGGACCTGAGCGACGGGCTGCGGCTTATCGCGCTGTTG
GAGGTGCTCAGCCAGAAGAAGATGCACCGCAAGCACAACCAGCGGCCCACTTTCCGCCAA
ATGCAGCTTGAGAACGTGTGCGGTGGCGCTCGAGTTCCTGGACCGCGAGA
GTGTCCATCGACAGCAAGGCCATCGTGACGGGAACCTGAAGCTGATCCT
TGGACCCTGATCCTGCACTACTCCATCTCCATGCCCATGTGGGACGAGGA
GAGGCCAAGAAGCAGACCCCCAAGCAGAGGCTCCTGGGCTGGATCCAGAACAAGCTGCCG
CAGCTGCCCATCACCAACTTCAGCCGGGACTGGCAGAGCGGCCGGGCCCTGGGCGCCCTG
GTGGACAGCTGTGCCCCGGGCTGTGTCTGCTGACTGGGACTCTTGGGACGCCAGCAAGCCC
GTTACCAATGCGCGAGAGGCCATGCAGCAGGCGGATGACTGGCTGGGCATCCCCAGGTG
```

cDNA 1

```
...
GACAAGGGGGAGTACACACTGGTGGTCAAATGGGGGGACGAGCACATCCCAGGCAGCCCC
TACCGGTTGTGGTGCCCTGAGTCTGGGGCCCGTGCCAGCCGGCAGCCCCCAAGCTGCC
CCGCTACCCAAGCAGCCCCGCCCTTTCCTTCAACCCCGGCCAGGCCGCCCTGGCCGC
CCGCTGTCACTGCAGCCGCCCTGCCCTGTGCCGTGCTGCGCTCACCTGCCTCCCCAGC
CAGCCGCTGACCTCTCGGCTTTCACTTGGGCAGAGGGAGCCA
```

```
>ENSG0000013619|X|protein_coding
GGCGCGGAGCCGGGCGGTGGGAATGGAGCGAGCAGATTGAGGCCGCCACTGCAGCGCCGC
CAGCATGAACTTGGCCGCGAGCTGAAGCGGCCGGCGGGCGGGCGGGCGGGCGGGCGGG
CGCTAGCCAGGGGGTGATCTGCAACAGGGCTGGGTTTCTTGGCGGCCGCC
CTGCCCCGCGCCGGCGGCCGCCCTCGGACACTGCCCCCGCCCGCCG
CAGCACGCCCTGTGTCTAGGTCTTTGGGAAACGCCTTGGAGAGTCAAGAATAAAATTTGC
AGGTCAAACAATGGATGACTGGAAAAGTCGGCTTGTAAATCAAGAGCATGCTTCCCCATTT
CGCCATGGTGGGAAATCGTCAGGAGCCCAGAAAGCTCCAGGAATCGGGAAAGAAGCCCTC
GTGGATGGAGGAAGAAGATTTATCTTTTCTCTACAAGAGCAGCCCAGGAAGAAAGCATCA
GGGAAGTGTAAAGAGGAGACAAGAAGAAGACCCTTCCAGTTTCCAGACATGGCTGATGG
GGGCTACCCTAATAAAAATTAAGAGGCCCTTGCCCTGAAGATGTCACCCCTTGCAATGGGCC
AGGTGCTCATCCTAGTACTGCTTGTGTCAGAACTGCAGGTCCCTCCATTGACAATAAATCC
TAGCCCTGCGGCTATGGGAGTGGCTGGCCAGTCATTAAGTCTGGAGATAAACCCTATGAA
TGGCAACATCATGGGCTCACCATTTGTAGTACCACAGACTACAGAAGTGGGACTGAAAGG
GCCCACTGTTTCTTACTATGAGAAAATCAACAGCGTGCCGGCTGTAGACCAGGAGCTTCA
AGAGCTGCTAGAGGAGCTCACCAAAATCAAGACCCTTCTCCAAATGAGCTAGATCTTGA
GAAGATACTGGGGACGAAGCCAGAAGAGCCACTGGTTTTAGATCATCCCAGGCAACCCT
AAGCACAACCTCCCAAGCCTTCGGTTCAGATGTCACACTTGGAGAGCCTGGCTTCCAGCAA
GGAGTTTGCTTCTAGTTGCAGCCAAGTTACTGGCATGTCACCTCAGATCCCATCCTCCTC
CACAGGGATCAGCTATTTCGATTCCCTTCCACCAGTAAGCAGATAGTGTACCCGAGTTCTTC
AATGGCACAGTCCAAGAGCCAGGTCCAGGCCATGCTCCCTGTGCTCTGCCCCCTTACC
AGTGCCTCAGTGGCATCACGCCACCAGCTGAAGGCGTTGGCAGCCAGCAAGCAGGGGTC
TGCTACAAAGCAGCAAGGGCCACCCCCAGTTGGTCTGGTCTGCCCTCCTCCAGGACTCTC
TCCACCTTACCGCCCAGTGCCATCACACACCCACCACCGCTGCCACTGCCACCACCACC
```

cDNA 2

## V) BIOMART - Exercises

*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 'novel' 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 49'
- 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)' 'NOVEL'.
- In the 'PROTEIN' section, select the second '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 1.

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

- Select the 'SNPs' Attribute Page.
- In the 'GENE' section 'Ensembl Gene ID' and 'Ensembl Transcript ID' are selected by default – also select 'Ensembl Peptide ID and 'Ensembl Peptide length'.
- In the 'GENE ASSOCIATED SNPs' section 'Reference ID' is selected. Also select 'Allele', 'Peptide location (aa)', 'Location in Gene (coding etc)', 'Synonymous Status' and 'Peptide Shift'.

*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'.

Note that the output for this query gives you one row for each SNP, and if there are alternative transcripts then SNP data is given for each. This means that a particular SNP may appear more than once.

Find the coding SNPs, and note that you have information about the effect of the SNP, and its location within the protein. Synonymous status is 'yes' for silent mutations. Two amino acids will be shown in the 'Peptide Shift' column if there are two alleles on the protein level. The Peptide location (aa), Synonymous Status and Peptide Shift will all be blank if the SNP is not in a coding region.

2. Click 'New' to start a new query. Retrieve the gene structure (i.e. start and end coordinates of exons) of the mouse gene ENSMUSG00000042351.
3. Retrieve all human disease genes located between p11.2 and q22 (these are bands on chromosome 1).
4. The file at [http://www.ebi.ac.uk/~xose/Affy\\_exercise.txt](http://www.ebi.ac.uk/~xose/Affy_exercise.txt) contains a list of probeset IDs from a microarray experiment using the Affymetrix array HG-U133 Plus 2.0 (human). Enter the probe list at this url into BioMart to retrieve the 500 bp upstream of the transcripts matching these probeset IDs.
5. Retrieve the sequences 5kb upstream of all human 'known' genes between D1S2806 and D1S464.
6. Retrieve sequence (including reference ID in the header) of all human SNPs on chromosome 6, band p25.3, that have an ID from Watson's genome. Export both the SNP sequence and 200 bases flanking (adjacent to) it.
7. Retrieve the mouse homologues of *Homo sapiens* genes CASP1, CASP2, CASP3, and CASP4. (These are HGNC symbols for the genes).
8. Design your own query!

## Answers (BioMart)

1. You should find **one** novel gene on chromosome 2 with this InterPro domain. (Note: there can be more than one gene with one InterPro domain). The result set has one transcript and a total of 264 rows of output (to see this, change the option from TSV to XLS under 'Export all results' and click 'Go', then open in Excel so you don't have to count the rows manually). The transcript has 9 coding SNPs ('Location in Gene' is 'coding'), most of which are non-synonymous ('Synonymous status' is 'no') and thus affect the amino acid sequence of the encoded peptide. One allele is a stop codon (\*)- can you find it?

2. **Click New. Select: Database and dataset:** 'Ensembl 49' and '*Mus musculus* genes (NCBIM36)'.

**Filters: GENE** 'ID list limit Ensembl Gene ID(s)': enter the mouse gene ID.

**Attributes 'Structures'**: select in the **EXON** panel: 'Ensembl Exon ID', 'Exon Start' and 'Exon End'.

Click '**Results**'.

You should find **8 exons**. Follow the link from the Ensembl Gene ID in your output back to the **GeneView** page to confirm the BioMart data with the gene structure displayed on this page.

**3. Database and dataset:** 'Ensembl 49' and '*Homo sapiens* genes (NCBI36)'.

**Filters: REGION** 'Chromosome 1', 'Band Start p11.2, 'Band End q22' ,  
**GENE:** 'with Disease Association Only' (look under 'ID LIST FILTERS')

**Attributes: Features: EXTERNAL:** select 'GO ID' and 'GO description' along with the default options ('Ensembl Gene ID' and 'Transcript ID').

**Results** should show **17 Ensembl genes** (multiple transcripts and GO terms).

**4. Database and dataset:** 'Ensembl 49' and '*Homo sapiens* genes (NCBI36)'.

**Filters: GENE:** 'ID list limit': Affy hg u133 plus 2 ID(s) and enter the list of probe IDs.

**Attributes: 'Sequences'** select 'Flank (Transcript)', 'Upstream flank 500'. In the header, apart from the already default selected options, select 'Ensembl Transcript ID'.

You should find upstream sequences for the transcripts of **31 genes** (Hint: click 'count' to see the number of genes!)

**5. Database and dataset:** 'Ensembl 49' and '*Homo sapiens* genes (NCBI36)'.

**Filters: REGION** 'Marker' Start D1S2806 End D1S464  
**GENE:** 'Status: KNOWN'.

**Attributes 'Sequences'** and select, apart from the already default selected options, 'Flank (Gene)' and 'Upstream flank 5000'.

You should find sequences for **26 genes**.

When you choose the option 'Flank (Gene)' you will see only one upstream sequence per gene in the output. In the case where a gene has multiple transcripts, the upstream sequence of the transcript that extends the furthest at the 5' end is shown. If you want to export the upstream sequences for each transcript you should choose the option 'Flank (Transcript)'.

'Known' genes are Ensembl gene predictions that could be matched to same-species external database entries (e.g. UniProt/SwissProt) with a high similarity score (i.e. with BLAST or a similar sequence identity-matching program)

**6. Database:** 'SNP' and **dataset:** '*Homo sapiens* SNPs (dbSNP128;HGVbase 15; TSC 1; ENSEMBL, Affy 100K and 500K arrays)

**Filters: REGION:** 'Chromosome 6', 'Band p25.3'

**GENERAL SNP FILTERS:** SNP source: 'with Ensembl Watson ID(s)' Only

**Attributes 'Sequences': SEQUENCES :** 'SNP sequences', 'Upstream flank 200', 'Downstream flank 200'.

**SNP: SNP attributes,** default is 'Reference ID'.

You should find **209,168 SNPs**.... Beware this is a large dataset. To export it all, it is best to use the 'compressed web file (notify by email)' option.

**7. Database:** 'Ensembl 49' **Dataset:** '*Homo sapiens* genes (NCBI36)

**Filters: GENE:** 'ID list limit HGNC Symbol(s)'. Enter the human HGNC (HUGO) symbols in the box: CASP1, CASP2, CASP3, and CASP4.

**Attributes:** Under '**Homologs**', select in the '**MOUSE ORTHOLOGS**' panel 'Mouse Ensembl Gene ID' and 'Mouse External ID'. Also, scroll back up to the **GENE** panel. Along with the default options 'Ensembl gene ID' and 'Transcript ID' select 'Description' (these attributes are for the starting dataset... i.e. Human.)

**Results** displays the mouse orthologues of the 4 human CASP genes.

## VI) EVALUATING GENES AND TRANSCRIPTS

### (The 'GeneBuild')

#### **Main Exercise: Examine the evidence for the FOXH1 gene.**

*These exercises focus on the FOXH1 gene to demonstrate how the underlying protein and mRNA used to build an Ensembl gene can be seen. Is it a well-determined gene...?*

#### **1. Display the human FOXH1 gene in GeneView.**

Enter FOXH1 into the text search box at the top of any human Ensembl page. Take the link to the **GeneView** page for the Ensembl Gene: ENSG00000160973.

**Q1:** What are the external database sources for the gene name and for the description?

#### **2. Examine the supporting evidence for the FOXH1 gene in ExonView.**

Scroll down the **GeneView** page and have a look at the predicted exon structures. Note the 5' and 3'UTRs (untranslated regions). Compare the two transcripts using this view.

Click on 'Exon information' in the left-hand menu to go to **ExonView** for one of the transcripts. The bottom section of **ExonView** shows the supporting evidence that was used during the Ensembl transcript building process.

**Q2:** Which databases did the entries come from?

Click on a green box of a supporting evidence entry to see the alignments of the Ensembl predicted transcript against the supporting evidence.

*Optional: Click on the 'Gene information' link in the left-hand menu to return to the **GeneView** 'Gene Report' for FOXH1.*

#### **3. Examine other evidence for the FOXH1 gene in ContigView.**

Click on the link 'Graphical View' in the left-hand menu. This takes you to **ContigView** displaying only the region encompassing the gene. Zoom out by clicking on the '-' button next to the Zoom triangle.

Look at other protein/mRNA tracks, e.g. Human proteins, Unigene (for all species), Human cDNAs, and EMBL mRNAs (again across species). These are proteins/mRNAs that align to the genome in this area. Note that the track labels are links to the help page. Clicking on a block drawn in the new tracks brings up a pop-up menu with a link to the database entry. Try some links.

Condense the Unigene and Protein track as well as the Overview section by clicking on the '-' boxes.

Under 'Decorations' select 'Show empty tracks'. This will help you remember which tracks you have selected. Examine the evidence for the FOXH1 gene in the new evidence tracks. Are there proteins and mRNAs that align to the Ensembl prediction?

#### **4. Compare transcript predictions made by other methods.**

In the 'Features' menu, turn off most of the evidence and make sure Ensembl genes, Vega Havana and Genscans are turned on.

Look at the Genscan track. Zoom out further.

**Q3:** How does the Genscan prediction differ from the Ensembl prediction? Note that this track shows *ab initio* Genscan predictions, not relying on supporting evidence.

**Q4:** What is a Vega Havana transcript?

Use the 'DAS sources' menu to turn on tracks showing transcript predictions from other groups (e.g. NCBI Gnomon). Compare and contrast!

#### **5. Look at the 'Similarity Matches' section.**

Go to **TransView** by clicking on the Ensembl transcript FOXH1 and taking the direct 'Transcr.' link from the pop-up menu.

'Known' Ensembl transcripts like this one (shown in red in **ContigView**) have been successfully mapped to external database entries such as UniProt or NCBI EntrezGene entries *for the same species*. (Note that this mapping is done *after* the genes have been built). Novel transcripts may match to mRNA and protein information in other databases, for an alternate species. Thus a novel transcript is novel for that species.

Matches to IDs in other databases are shown in the 'Similarity Matches' section of **TransView** (repeated in the 'Transcript' section of **GeneView**). Have a look at the types of databases linked out to.

**Q5:** What do the Target and Query % ids indicate? Check the online Help pages.

#### **Answers (Evaluating Genes and Transcripts).**

**A1:** Name: HUGO Gene Nomenclature Committee (HGNC). Description: Uniprot/Swiss-Prot.

**A2:** Click on the ID number to go to the original database entry. You may have to scroll down to find the correct entry. The boxes represent the exons,

the darker green they are, the better the supporting evidence is. Click on a green box for the alignments with the transcript.

**A3:** The Genscan transcript prediction shows exons not present in the Ensembl transcripts for FOXH1 and doesn't predict UTRs. Genscan is an *ab initio* predictor, a program run on the sequence alone, without using protein and mRNA evidence. It has the tendency to overpredict exons.

**A4:** Havana is a subgroup of VEGA, the Vertebrate Genome Annotation consortium. Havana transcripts are manually curated (determined on a case-by-case basis) and are merged with Ensembl predictions if they match up exactly (leading to golden transcripts). For more, go to these links:

<http://www.sanger.ac.uk/HGP/havana/>  
<http://vega.sanger.ac.uk/index.html>

**A5:** In **TransView**, under Similarity Matches, Target %ID indicates the percentage of the Ensembl prediction matching the external sequence database and Query %ID is the percentage of the external database sequence matching the Ensembl prediction! Can you find this in the help pages?

## VII) COMPARATIVE GENOMICS

### 1. Investigating a protein family.

*This exercise focuses on protein families and orthologies, using AIM1, a cancer-related gene, as an example.*

Find the **GeneView** page for human AIM1, a gene absent in a type of melanoma. Beware... there is a VEGA gene as well as an Ensembl gene! Let's work with the Ensembl gene for now.

**Q1:** What is the MIM ID for the AIM1 transcript? Hint... find this in 'Similarity Matches' for one of the transcripts. Click on the MIM ID...

Examine the protein family by taking the link to the associated Protein Family (the **FamilyView** page).

**Q2:** How many human Ensembl genes produce peptides in this family?

**Q3:** Are they all 'known' genes?

**Q4:** Are there peptides in the same family for mouse (*Mus musculus*), rat (*Rattus norvegicus*) and zebrafish (*Danio rerio*)?

**Q5:** Families are calculated using Ensembl peptides along with UniProt peptides. Can you find on this page where the UniProt peptides in this family are listed?

### 2. Genomic alignments and conserved regions

Find the **ContigView** page for human SNX5.

**Q6:** In which band is the human SNX5 gene located?

The 'constrained elements' track shows the most highly conserved regions calculated from the 10-way multi-species alignment.

**Q7:** Do any of these regions fall outside of exons?

Expand the 'Conservation' track by clicking on the '+' in front of it. This shows a score for each basepair based on how conserved it is across the alignments. Clusters of peaks correspond to the 'Constrained elements blocks' shown above.

Collapse the track if desired. Choose the pairwise alignment with mouse from the Comparative menu. Selecting the '+' in front of the track will expand the track so that all alignments on the same mouse chromosome and strand will be clustered together.

**Q8:** On which chromosome(s) and strand(s) are the mouse alignments for this region?

Turn on the frog (*X. tropicalis*) pairwise alignment with human, using the Comparative roll-down menu.

**Q9:** Are there fewer or more conserved regions between human and frog when compared with human and mouse? Would you expect an SNX5 frog gene in this region?

Expand the frog/human alignment track by clicking on the '+' in front of it. There are several overlapping regions; this region in human aligns to several scaffolds in the frog assembly. This could be a result of the frog sequence not being completely finished (i.e. overlapping sequence blocks, or 'scaffolds' have not yet been aligned into one continuous frog genome sequence). Click on an exon in the second, long cluster of frog alignments. 'Jump to *Xenopus tropicalis*' to view ContigView for frog. Can you see the snx5 gene there?

### 3. Examine the syntenic regions

Syntenic regions in Ensembl are calculated from the pairwise alignments (whole genome alignments between two species) and correspond to long regions (100 kb or more) that have high sequence similarity. Gene order is expected to be conserved in these regions. Two pages are available to view these regions: **SyntenView** and **CytoView**.

To go to SyntenView, where two species may be compared, start by finding the **ContigView** page for human Myosin VI.

Click on 'View syntenic regions with... *Mus musculus*', a link at the left hand side of the page.

**Q10:** How many mouse chromosomes show syntenic regions with human chromosome 6?

**Q11:** Is there a mouse homologue to human Myosin VI? Can you see this without leaving the **SyntenView** page?

Gene order is conserved in human and mouse for this syntenic region. Comparing homologous genes across both species can show where gene loss may have occurred.

**Q12:** Is there a human gene in this list without a mouse homologue?

Click on '**MultiContigView**' next to 'Myo6' in mouse to reach a comparison of the human and mouse chromosomes in this region.

**Q13:** What does the vertical blue line connecting the lowest panels show?

**Q14:** The Myosin VI genes are golden in human and mouse. What does the colour signify?

Syntenic regions may also be viewed in the **CytoView** page. Take the 'Graphical overview' link at the left to go to **CytoView**. Turn on syntenic regions with mouse and opossum using the 'Comparative' roll-down menu. Click on the red and green lines in the 'Mouse' and 'Opossum' tracks to see which base pairs and chromosome strand are syntenic to human in this region.

The 'View alongside...' link at the left will take you back to **MultiContigView**. Instead, take the 'View alignment with...' link and choose '*Mus musculus*'. This will bring you to the '**AlignSliceView**' page, where again, human and mouse chromosomes can be compared side-by-side. However, as **MultiContigView** shows the chromosomes as they are, **AlignSliceView** shows the alignment, and the genomic assembly may be altered (gaps introduced) to fit the alignment in this view. Read the 'Help' page to understand the view.

Select the '10 amniota vertebrates' in the 'Comparative' roll-down menu.

**Q15:** Is there a gene in this aligned region across the ten species?

## Answers (Comparative Genomics)

### Main Exercise (1)

**A1:** Find this in the GeneView page. Make sure the ID at the top is ENSG00000112297. The MIM ID (corresponding to the entry on the 'Online Mendelian Inheritance in Man database) is 601797, found in the 'Similarity Matches' section of the **GeneView** page. (See the report for the second, AIM1, transcript).

**A2:** The Family ID is ENSF00000000036 ('Unknown', as there is no consensus name for all peptides involved). Under the link to FamilyView (this ID) it indicates that 80 human Ensembl genes belong to the family.

**A3:** They are all "Known" protein-encoding Ensembl genes (all genes have a name; none are 'novel'). Take the link to 'FamilyView' to see this.

**A4:** Yes. Scroll down the 'FamilyView' page to find the peptide lists for other species.

**A5:** In the 'Other peptides in Family ENSF...' section, just under the gene names and karyotype, the UniProt identifiers clustered into this family are listed.

**A6:** Band p11.23, chromosome 20 (*hint: go to **ContigView***)

**A7:** Yes. This track is turned on by default- if not, switch it on in the 'Comparative' roll-down menu under 'Constrained elements'. Looking at the visual display in Detailed view, most of the conserved blocks line up with exons (if you are unsure, zoom in.) Clicking on a conserved element shows its precise location across the genomes aligned. There are some conserved elements near the 3' end of the gene that fall outside of exons (especially in the long intron of the SNX5 transcript). Could these be important regulatory regions? *Hint... turn on ncRNA (non-coding RNAs) in the Features menu.*

**A8:** All are on chromosome 2, forward strand. In the expanded format, i.e. a (-) is found next to the track, all alignments appear grouped together on one line, indicating they are on the same chromosome and strand. Click on a block to find this is chromosome 2, forward strand.

**A9:** Fewer. You would expect a gene, because the conserved regions with frog line up to the human SNX5 exons, supporting a hypothesis in which those exons have been conserved across time in frog as well as in human. Thus, they are probably part of an expressed protein.

**A10:** 7 chromosomes in mouse are shown in **SyntenyView**. These are the chromosomes that share syntenic blocks with human chromosome 6.

**A11:** The homologues are listed at the right of the page. Mouse Myo6 is a homologue of human Myo6. It is on chromosome 9 at position 80.01 Mb in mouse.)

**A12:** Q9H3A8 is a human gene without a partner in the mouse homologue list. Perhaps the mouse gene has simply not yet been annotated by Ensembl, or perhaps, over time, gene loss occurred at that locus in the mouse. Or, perhaps the human gene prediction is faulty! More investigation would be needed to determine the cause of the missing homologue.

**A13:** The vertical blue line in the lowest, 'Detailed View' panel connects homologous genes across, in this case, human and mouse. Remember homologues are determined using the longest translation of a gene only.

**A14:** A golden gene has at least one transcript that is identical to (and merged with) a transcript determined by manual curation (the Havana team).

**A15:** Yes, a gene is present in these aligned regions for all species. Some exons align well, indicating they could be homologues.

## VIII) VARIATIONS

These exercises focus on SNPs (Single Nucleotide Polymorphisms) in Ensembl. Most of these are downloaded from dbSNP and placed along the sequence assembly using flanking sequence to the SNP to match it.

### 1. Display all SNPs for a region.

*This exercises focuses on SNPs in the **ContigView** page.*

Go to the **ContigView** page displaying Human chromosome 7: bp 116124438-116129149. Turn on SNPs.

**Q1:** Are there non-synonymous coding SNPs in the gene? (*Remember, non-synonymous SNPs change the amino acid sequence*).

You can see information about a SNP by clicking on it. Click on a non-synonymous SNP and follow the link 'SNP properties' to its **SNPView** page.

In the **SNPView** page, click on 'SNPs in gene context' at the right of the second panel. You will be taken to the **GeneSNPView** page, where you can see all the SNPs for the MET gene. Under 'SNP type' roll-down menu in the **GeneSNPView** page, select only 'non-synonymous'.

### 2. Display SNPs linked to a disease within the transcript sequence.

*Starting with disease-causing SNPs, how can we view them within the sequence? Hint... focus on the TransView page.*

In the article "Screening of the delta-F508 mutation and analysis of two single nucleotide polymorphisms of the CFTR gene in a sample of the general population of Valparaiso, Chile" by L.A. Vera et al." (Rev Med Chil 2005, 133:767-775.) the SNPs leading to peptide alleles M470V and T854T are studied.

**Q2:** Find these two SNPs in the **TransView** and **GeneSNPView** pages. Note, the positions refer to numbers in the amino acid sequence.

### 3. Display all SNPs for one gene.

Retrieve all the validated SNPs associated with the human CFTR gene as follows. Display only validated SNPs in the '**GeneSNPView**' page used in question 2.

Now use BioMart to ask how many of these validated SNPs are coding.

**Q3:** How many validated SNPs are in the coding region of CFTR?

## Answers (Variations)

**A1:** Click on the 'human' icon on the Ensembl homepage. Enter the chromosome number and region in base pairs under the karyotype. This should take you to **ContigView**. You may have to turn on the 'SNPs' track from the 'Features' drop-down menu on top of 'Detailed View'.

SNPs should now be shown as vertical lines of different colours along the chromosome. The SNP legend is shown at the bottom of the panel. This region contains one exon from the 'MET' gene. There are a few non-synonymous coding SNPs (shown in yellow) in this region.

**A2:** Perform a text search for 'CFTR' in human. This should lead you to ENSG0000001626. Go to the **GeneView** page and then the **TransView** page by clicking on 'transcript information' on the left hand navigation column. At the bottom of the transcript sequence are panels for customisation of the display. By selecting the options 'Exons, Codons, Translations and SNPs' and 'Number residues: yes' you can display the SNPs in the transcript sequence. Alleles and alternative codons are shown by pointing your mouse over the nucleotide and amino acid residues, respectively. M470V is shown as a red 'V' at position 470 in the peptide sequence (the second line of numbering in the sequence).

Continue on to the **GeneSNPView** page for this gene by clicking 'Gene variation info.' in the side menu.

The two SNPs are displayed in the 'SNPs and variations' figure and the 'Variations and consequences' table. In the figure, M470V is shown as 'V/M' in yellow (as it is a non-synonymous coding SNPs). T854T is shown as 'T' in green, as it is a synonymous coding SNP. The diagram only shows bp positions, so it could be easier to find the two allelic changes in the table. Note that you can use the 'SNP class' and 'SNP type' drop-down menus in the figure to configure (simplify) both figure and table by choosing only to see coding SNPs.

**A3:** Validated SNPs have been checked after submission to determine if they are SNPs or mutations. The validation status comes from NCBI, and explanation to the validation terms can be found in the 'Help' to SNP-based pages.

To view only validated SNPs, configure the 'Validation' roll-down menu in the **GeneSNPView** page by deselecting 'no information' and closing the menu. Now only validated SNPs are shown. How many are coding? Deselect, under the 'SNP type' roll-down menu in this page, upstream, downstream, UTR, intronic and intergenic SNPs. Close the menu. The display and table will be simplified. Count **14** coding SNPs that have been validated.

Go to BioMart. Select the SNP database and *Homo sapiens* SNPs. In **Filters**, expand the 'GENERAL SNP FILTERS' section. Select 'SNPs that have been validated' (Only). In the 'GENE ASSOCIATED SNP FILTERS' section, enter the Ensembl Gene ID ENSG00000001626. Select 'Ensembl Gene Location' as 'Coding Only'.

'Count' gives you **14** validated, coding SNPs.

## IX) TYING IT TOGETHER – CASE STUDIES (using Ensembl comprehensively)

Consider the following case studies. The answers to these questions tie together pages and tools on the Ensembl site.

- 1) I work on non-coding RNA genes such as snoRNA (small nucleolar RNA) and miRNA. Are there any on mouse chromosome 1 (of any type), and where are they? How many are there (use BioMart). Could I view them graphically?
- 2) Many regulatory factors appear in the 5'UTR, and I am interested in searching for them using motif-scanning tools. How would I obtain the 5'UTR or upstream region in the IRX4 gene in human? What can I learn about the UTR and upstream regions using **ContigView**?
- 3) (*Using DAS*): We have determined two SNPs in the lab not found in Ensembl. We have not yet submitted to dbSNP, but can we just display these SNPs on our computer in Ensembl? Here is some information about the SNPs:

Names: 1 and 2

Positions: both on chromosome 3 (mouse)

base pairs: 75305500 and 75450001

Type and subtype are both 'a' (internal lab naming convention)

Phase is 0, Score is 100.

Hint: start from ContigView for mouse chromosome 3: 75300000 - 75500000

- 4) I would like to export the gene structure of zfish 'pp2ca2' in GFF format (I will use a program that requires data in this format).  
(Find out more about the GFF format here)  
[http://www.sanger.ac.uk/Software/formats/GFF/GFF\\_Spec.shtml](http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml)

### Case Studies: Answers

- 1) Using BioMart to find all the RNA genes will take 2 queries as the filters only allow a subset of ncRNA types to be selected. There are **192** total RNA genes on chromosome 1. You need only select the filters, then click 'count' to see how many genes there are.

From the result table, follow the link to one of these genes into the Ensembl browser (for example, ENSMUSG00000070103).

Alternatively, go to '**ContigView**' for a region on chromosome 1 and select, in the 'features' menu, 'ncRNA' to visualise the Ensembl ncRNAs in the chromosome. Also, select 'rFAM' to select RNAs predicted by this program. You can also select tRNA, miRNA, and other options.

- 2) **Step 1: Exporting UTR sequence.** Export the UTR information by clicking 'Export gene data' at the left of the **GeneView**, **TransView**, or other page and then selecting '5' or 3' UTR' in the next window to export the sequence.

Alternatively, use **BioMart** to export the 5' and 3' UTRs.

**Step 2: Visual display and comparison of promoter elements.**  
*What can you learn about the upstream regions of this gene in ContigView? Use the features menu and conserved tracks.*

In **ContigView** (the 'graphical view' link) select 'CTCF' for CTCF binding sites (click the blue 'Help' if you are unsure of what this is), CpG islands, and motifs in the CisRED/miRANDA database' for sequences associated with regulatory regions. For genes that have no UTR annotated, it can be helpful to select "Eponine" (a program that predicts transcription start sites) under the 'Features' menu. Also, in the Comparative menu, select alignments such as '10 amniota vertebrates'. Constrained elements show the highest sequence similarity for the alignment.

- 3) **HINT: Upload as a DAS source from ContigView.**

Go to **ContigView** encompassing this region in mouse, then select '**DAS Sources**' from the roll-down menu above the detailed-view panel and '**Manage Sources**'. Next, click on '**Upload your data**' in the new window (at the left). Read the instructions by clicking on the link at the top of the page (especially the 'formatted correctly' link.) Enter in your email and a NON-SECURE password, then paste your data according to the format described in the link. (Note: columns must be separated by tabs, NOT SPACES! This will require making the data columns in another program such as Notepad).

Once the data is entered, click 'Next' and select **ContigView** (note you can select more than one page). Click 'Next' again, name the track if you'd like, and click 'Finish'. Once you reach the DAS sources list (end page) 'Close window' to refresh **ContigView**... a new track should be displayed, along with SNPs 1 and 2!

Here is the correct format for the SNP information given in the question:

[http://www.ebi.ac.uk/~gspudich/workshop\\_presentations/snp\\_example.txt](http://www.ebi.ac.uk/~gspudich/workshop_presentations/snp_example.txt)

Note: you can deselect your track under the 'DAS sources' menu of ContigView.

- 4) To export in GFF format: Select 'export gene data' at the left of the **GeneView** page and select GFF under 'Output format'. Or, export the chromosome and base pair start and stop of the exons using the 'structures' attribute page in **BioMart**. Select GFF format under 'Display rows as GFF' and click 'Go' to export the file.

**Focus**

**1** Genome sequence assemblies determined by sequencing institutes are incorporated into **Ensembl**. View the assembly used for a species on the main page.



**2** The **Ensembl** gene set is based on mRNA and protein evidence and is aligned to the genomic sequence assembly in the 'genebuild'. Use 'ExonView' to view this information.

**3** **Ensembl** focuses on vertebrates however other model organisms are included such as yeast, fly, and soil worm. View newly available genomic sequences in the **Pre!** site.



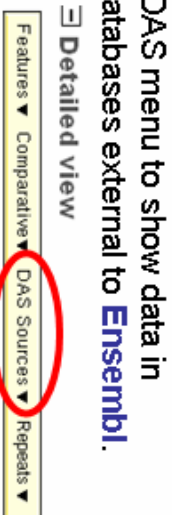
**Browsing**

**4** A new version of **Ensembl** is released every two months. View older versions in the **Archive!** site. New gene sets are determined when a new sequence assembly is available.



**ContigView** and **GeneView** can be customised using the 'features' menu.

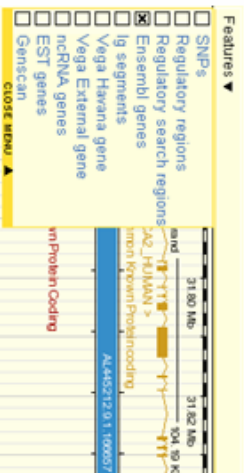
Select options to visualise transcripts, variations, regulatory regions, etc. Use the DAS menu to show data in current databases external to **Ensembl**.



View a short video about **ContigView** here!

[http://www.ensembl.org/CommonWorkshops\\_Online](http://www.ensembl.org/CommonWorkshops_Online)

**5**



**Tools**

**6** **Export SNP info in region**  
Use **BioMart** to get tables of genes and annotation in Microsoft Excel, HTML, or txt format. You can also export sequences with this Web-based tool!  
[www.biomart.org](http://www.biomart.org)

**7** **BLAST** and **SSAHA** are alignment programs- use them to match a sequence on any genome.



**8**

View variations for a gene or sequence in **GenesNPView** or **ContigView**.

We calculate phylogenetic gene trees, alignments, and homologies.

**HELP** **10**  
Send any questions or comments to

[helpdesk@ensembl.org](mailto:helpdesk@ensembl.org)