

# Getting started with Ensembl

## [www.ensembl.org](http://www.ensembl.org)

Ensembl provides genes and other **annotation** such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on protein and mRNA evidence in **UniProtKB** and **NCBI RefSeq** databases, along with manual annotation from the **VEGA/Havana** group. All the data are freely available and can be accessed via the web browser at [www.ensembl.org](http://www.ensembl.org). Perl programmers can directly access Ensembl databases through an Application Programming Interface (**Perl API**). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the **BioMart** web interface, which can extract information from the Ensembl databases without the need for programming knowledge by the user!

The screenshot shows the Ensembl website homepage. At the top, the Ensembl logo is on the left, and navigation links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors are on the right. Below the navigation bar is a search box with a dropdown menu set to 'All species' and a 'Go' button. Below the search box are two columns of content. The left column has a 'Browse a Genome' section with 'Favourite genomes' (Human GRCh37, Mouse NCBI37) and 'All genomes' (a dropdown menu). The right column has a 'Search any organism' section. Annotations with orange boxes and lines point to various elements: 'Mine data with BioMart' and 'Try our variant effect predictor' point to the search box; 'Change favourites' points to the 'Favourite genomes' section; 'Choose your favourite vertebrate' points to the 'Human' and 'Mouse' options; 'Browse plants, bacteria, fungi, protists, and metazoa at EnsemblGenomes' points to the 'All genomes' section; and 'Search any organism for a gene, location, variation, clone, probeset, or phenotype' points to the search box.

You will learn about

- Why do we need genome browsers?
- An introduction to Ensembl
- How information can be obtained from the site
- An overview of Ensembl tools

### Tired of reading?

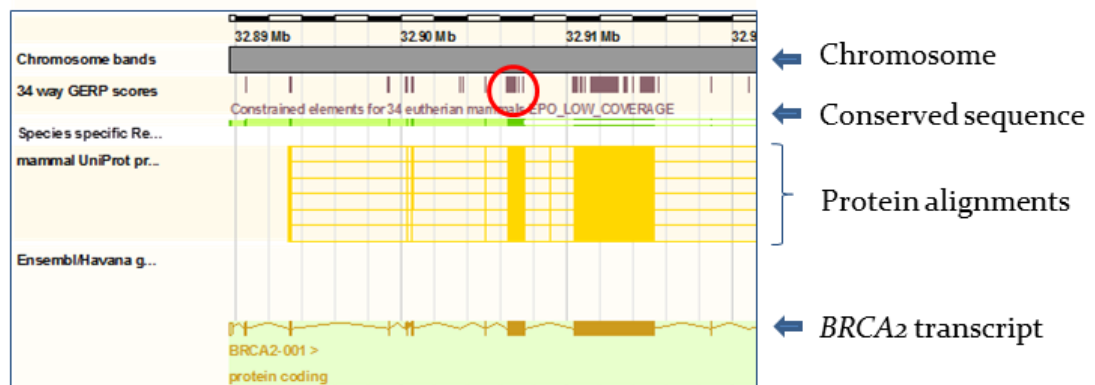
Check our video tutorial instead!

<http://www.youtube.com/user/EnsemblHelpdesk>

[The Ensembl Genome browser](#)  
[Introduction to BioMart](#)

## Introduction to Ensembl

Ensembl is a joint project between the EBI ([European Bioinformatics Institute](#)) and the [Wellcome Trust Sanger Institute](#) that annotates **chordate** genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl 'compara' team. Most annotation is updated every two months, leading to increasing Ensembl versions (such as version 62), however the gene sets are determined less frequently. A sister browser at [www.ensemblgenomes.org](http://www.ensemblgenomes.org) is set up to access non-chordates, namely bacteria, plants, fungi, metazoa, and protists.



*The Region in Detail view*

The vast amount of information associated with the genomic sequence demands a way to organise and access that information. This is where genome browsers come in. Ensembl strives to display many layers of genome annotation into a simplified view for the ease of the user. The picture above shows the **'Region in Detail'** page for the BRCA2 gene in human. The example shows blocks of conserved sequence reflecting conservation scores of sequence identity on a base pair level across 34 species. Conserved regions are displayed as dark blocks that represent local regions of alignment. One of the blocks is circled in red. You would only have to click on this block to see more details.

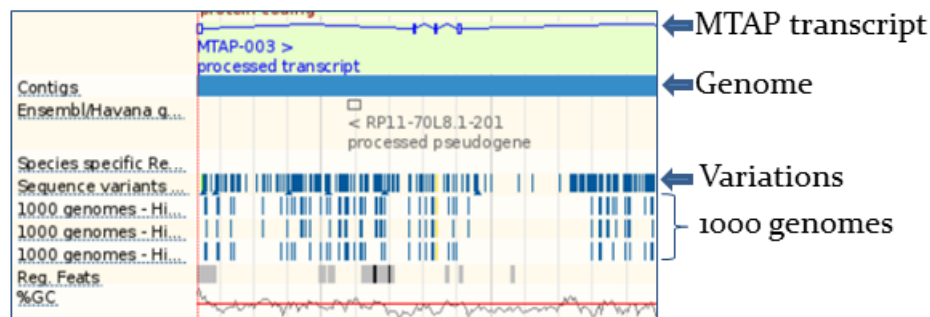
### Can't see Uniprot?

To add more tracks just click:

And select mammals: Uniprot  
under protein alignments

Also in this figure are proteins from the UniProtKB aligned to the same genomic region. Filled yellow blocks show where these UniProtKB proteins align to the genome, and gaps in the alignment are shown as empty yellow blocks. Note, in this case, the UniProtKB proteins support most of the exons shown in the Ensembl BRCA2-001 transcript (in gold).

Both [Ensembl](#) and [Vega \(Havana\)](#) transcripts are portrayed as exons (boxes) and introns (connecting lines). In fact, filled boxes show coding sequence, and empty boxes reflect UnTranslated Regions (UTRs). This 'Region in Detail' view is useful for comparing Ensembl gene models with current proteins and mRNAs in other databases like **NCBI RefSeq**, **EMBL-Bank**, and, in the example above, UniProtKB. Everything in this view is aligned to the genome.



*The Region in Detail view: 1000 genomes tracks*

The region in detail view can be configured (using the *Configure this page* tool button) to show regulatory features, sequence variation, and more! Click on any vertical line in the variation track for a menu about the SNP (single nucleotide polymorphism) or InDel (insertion deletion mutation). Clicking on 'Variation properties' in the pop-up box will bring you to [an information page](#) for the genetic variation, including links to population frequencies, if known. You can do the same for any regulatory feature.

An [index page](#) is provided for each species with information about the source of the genomic sequence assembly, a [karyotype](#) (if available), and a link to past or archive sites. The picture below shows the Ensembl homepage for human. Links to the human karyotype, a summary of gene and genome information, and the most common [InterPro](#) domains in the genome are found at the left of this index page.

**Description**

**Human (*Homo sapiens*)**

**Assembly**

**About the genome sequence**

This site provides a data set based on the February 2009 *Homo sapiens* high coverage assembly (Hg19) from the [Genome Reference Consortium](#). The data set consists of gene models built from the *genewise* alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

It also includes nine [haplotypic regions](#), mainly in the MHC region of chromosome 6.

As the GRC maintains and improves the assembly, patches are being introduced. [Patch release two \(GRCh37.p2\)](#) was included in Ensembl release 60. Currently, assembly patches are of two types:

- Novel patch: new sequences that add alternative sequence at a loci and will remain as haplotypes in the next major assembly release by GRC
- Fix patch: sequences that correct the reference sequence and will replace the given region of the reference assembly at the next major assembly release by GRC

The addition of the patches allows the annotation of some genes that can not be annotated correctly on the reference genome, such as the [ABO blood group gene](#), which can now be annotated as a [protein coding gene](#).

To convert your old data from Human assembly NCBI36 to GRCh37, click on 'Manage your data' on any human page and select 'Assembly converter' from the left-hand menu.

A preliminary assembly of the Neanderthal (*Homo sapiens neanderthalensis*) genome is available via the [Neanderthal Genome Browser](#), an Ensembl-powered project based at the Max Planck Institute.

**Previous assemblies**

NCBI36 (May2009) [Go to archive](#)

**Annotation**

In release 61 (January 2011), we continue to display a joint gene set based on the merge between the automatic annotation from Ensembl and the manually curated annotation from Havana. This refined gene set corresponds to GENCODE release 6. The Consensus Coding Sequence (CCDS) identifiers have also been mapped to the annotations. [More information about the CCDS project](#)

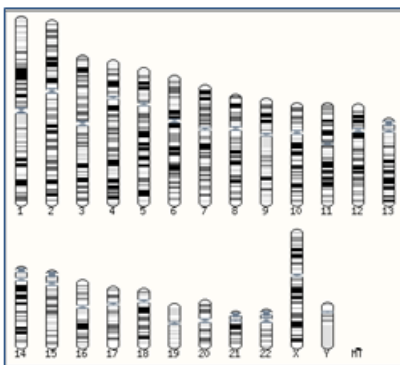
**Vega\***  
Additional manual annotation of this genome can be found in Vega

Ensembl release 61 - Feb 2011 © WTSI / EBI

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

**Links to older versions of Ensembl**

[Permanent link](#) [View in archive site](#)



Summary	
Assembly:	GRCh37.p2, Feb 2009
Database version:	61.371
Base Pairs:	3,279,005,676
Golden Path Length:	3,101,804,739
Genebuild by:	Ensembl
Genebuild method:	Full genebuild
Genebuild started:	Mar 2009
Genebuild released:	May 2009
Genebuild last updated/patched:	Jan 2011
Gene counts	
Known protein-coding genes:	20,935
Novel protein-coding genes:	615
Pseudogenes:	13,483
RNA genes:	8,383
Immunoglobulin/T-cell receptor gene segments:	553

Ensembl devotes separate pages and views in the browser to display a variety of information types, using a tabbed structure.

Human (GRCh37) Location: 13:32,889,611-32,973,347 Gene: BRCA2 Transcript: BRCA2-001 Variation: rs80358836

Variation displays

Summary Variation: rs80358836

View genotype information in the variation tab, gene trees in the gene tab, a chromosomal region in the location tab, and cDNA sequence alongside the protein translation in the [transcript](#) pages. Compare conserved regions with the position of genes and population variation in the **Region in Detail** view. See homology relationships in the [gene](#) page, or perform a **BLAST** or **BLAT** search against any species in Ensembl.

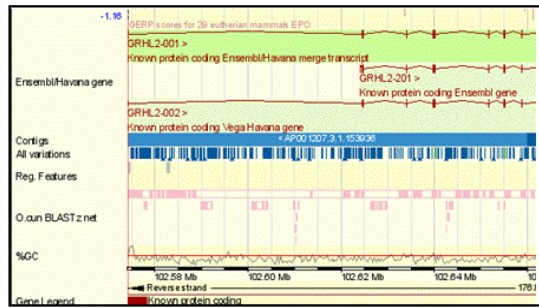
## Transcript Sequence w/Variations

```

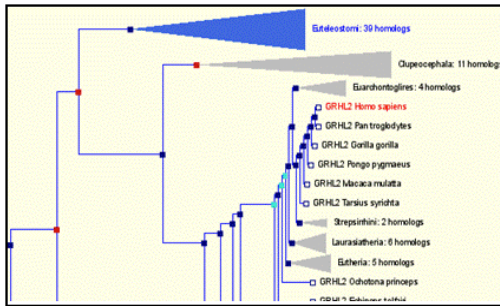
1  ATTGGATCAACATGTCACAAGAGTCGGACAATAAAGACTAGTGGCCCTAGTGCCC
   .....ATGTCACAAGAGTCGGACAATAAAGACTAGTGGCCCTAGTGCCC
   .....-M--S--Q--E--S--D--N--N--K--R--L--V--A--L--V--P--
61  ATGCCCAAGTGACCCCTCCATTCAATACCCGAAAGGCCTACACCGTGAAGGATGAAAGCCTGG
109  ATGCCCAAGTGACCCCTCCATTCAATACCCGAAAGGCCTACACCGTGAAGGATGAAAGCCTGG
   -M--S--S--D--P--P--F--N--T--R--R--A--Y--T--S--E--D--A--A--W--
121  AAGTCATACTTGGAGAATCCCTGACAGCAGCCACCAAGGCCATGATGAGCATTAAATGTT
159  AAGTCATACTTGGAGAATCCCTGACAGCAGCCACCAAGGCCATGATGAGCATTAAATGTT
   -R--S--Y--L--E--N--P--L--T--A--A--T--K--A--N--H--S--I--N--G--
181  GATGAGGACAGTGTCTGCTGGCCCTGGCCCTGCTCTATGACTACTACAAGTTCTCTGGAGAC
169  GATGAGGACAGTGTCTGCTGGCCCTGGCCCTGCTCTATGACTACTACAAGTTCTCTGGAGAC
   -D--E--D--S--A--A--L--G--L--L--Y--D--Y--Y--K--V--P--R--D--
57  -D--E--D--S--A--A--L--G--L--L--Y--D--Y--Y--K--V--P--R--D--

```

## Genes, SNPs, and Conserved Regions



## Homologues in Gene Trees



## BLAST and BLAT aligners

## Retrieving Data from Ensembl

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

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

## Synopsis- What can I do with Ensembl?

- View genes along with other annotation along the chromosome
- View alternative transcripts (including splice variants) for a gene
- Explore homologues and phylogenetic trees across more than 50 species for any gene
- Compare whole genome alignments and conserved regions across species
- View microarray sequences that match to Ensembl genes
- View ESTs, clones, mRNA and proteins for any chromosomal region
- Examine single nucleotide polymorphisms (SNPs) for a gene or chromosomal region
- View SNPs across strains (rat, mouse), populations (human), or even breeds (dog)
- View positions and sequence of mRNA and protein that align with an Ensembl gene
- Upload your own data
- Use BLAST, or BLAT, a similar sequence alignment search tool, against any Ensembl genome
- Export sequence, or create a table of gene information with BioMart
- Use the Variant Effect Predictor

## Need more help?

- Check Ensembl [documentation](#)
- Watch [video tutorials](#) on YouTube
- View the [FAQs](#)
- Try some [exercises](#)
- Read some [publications](#)

## Stay in touch!

- [Email](#) the team with comments or questions at [helpdesk@ensembl.org](mailto:helpdesk@ensembl.org)
- Follow the Ensembl [blog](#)
- Sign up to a [mailing list](#)



## Further reading

Flicek, P. *et al.*

### **Ensembl 2011**

Nucleic Acids Res. Advanced Access (*Database Issue*)

<http://nar.oxfordjournals.org/content/early/2010/11/02/nar.gkq1064.full>

### **Ensembl Methods Series**

<http://www.biomedcentral.com/series/ENSEMBL2010>

Giulietta M Spudich and Xosé M Fernández-Suárez

**Disease and Phenotype Data at Ensembl** UNIT 6.11 in *Current Protocols in Human Genetics*, Apr 2011.

Xosé M. Fernández-Suárez and Michael K. Schuster

**Using the Ensembl Genome Server to Browse Genomic Sequence Data.** UNIT 1.15 in *Current Protocols in Bioinformatics*, Jun 2010.

Giulietta M Spudich and Xosé M Fernández-Suárez

**Touring Ensembl: A practical guide to genome browsing**  
*BMC Genomics* 2010, 11:295 (11 May 2010)

Vilella, A.J. *et al.*

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

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

Smedley, D. *et al.*

**BioMart – biological queries made easy**

*BMC Genomics* 2009 Jan 14;10:22