

Need help?

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

Support helpdesk:
helpdesk@ensembl.org

Blog: View upcoming projects and find information about new species at <http://ensembl.blogspot.com>

Video: www.youtube.com/user/EnsemblHelpdesk

Facebook: www.facebook.com/Ensembl.org

Ensembl User Support
EMBL-EBI
Wellcome Trust Genome Campus
Cambridge
CB10 1SD, UK

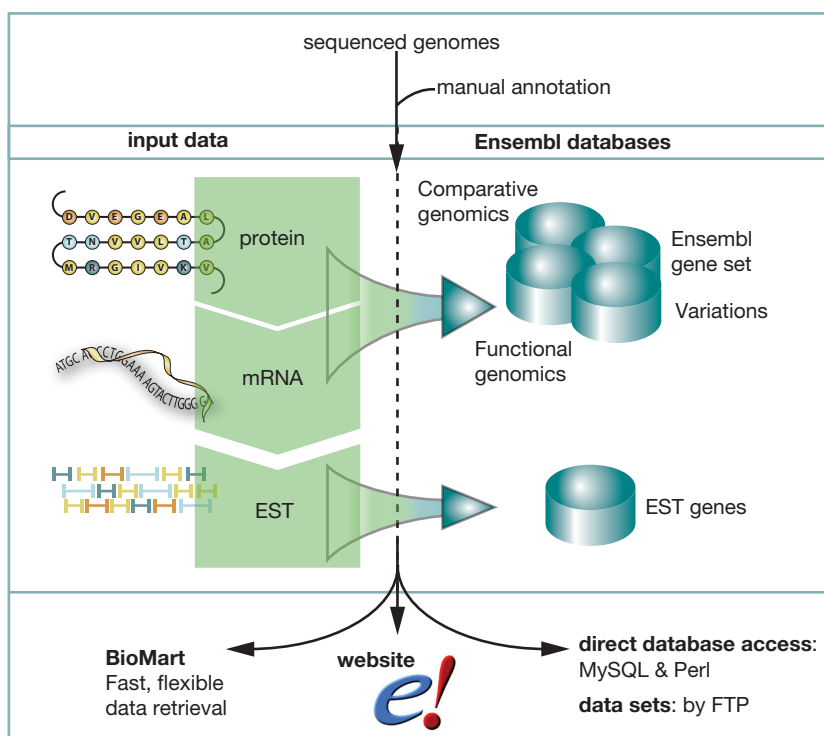
Ensembl

www.ensembl.org

The vast amount of information that comes with annotating a genomic sequence demands a way of organising and accessing that information. This need is met by Ensembl – a genome browser providing free access to the complete sequences of higher and model organisms. The Ensembl resource acts as both a source of genome sequence related data and an open source software system that can be used to organise such data. Ensembl provides a window to not only gene sequence, structure and position, but further annotation at both the gene and protein level. This includes information on protein domains, genetic variation, homology, syntenic regions and regulatory elements. Genomic sequence can be downloaded from the Ensembl browser itself, or by using the BioMart web interface or Perl API to extract information from Ensembl databases.

What is Ensembl?

Ensembl provides a comprehensive resource for the scientific community which allows analysis of genetic information within and between species. Hence, the resource is of use in a wide range of research fields from evolutionary biology to clinical research. Ensembl annotates chordate genomes (i.e. vertebrates and closely related invertebrates such as the sea squirt). Gene sets from model organisms such as yeast and fly are also imported for comparative analysis. All Ensembl genes are placed according to the experimental evidence of protein and mRNA sequences obtained from UniProt/Swiss-Prot, UniProt/TrEMBL and RefSeq. Sequence data is obtained from relevant genome sequencing centres and consortia. Manual annotation from the VEGA/Havana project is included for human, mouse and other supported vertebrates.



Schematic representation of the integration and assembly of genomic information to determine the Ensembl protein coding gene set and separate EST genes.



In addition to human, mouse, rat and zebrafish (Ensembl's central genomes), Ensembl provides annotation for chicken, cow, dog, chimpanzee, platypus, yeast, *C. elegans* and more. Ensembl includes a variety of annotation-specific pages in its genome browser that are accessible via the World Wide Web and annotation is updated every two months. Alignments and homology are calculated anew for each updated Ensembl version in order to include new genomic data in these predictions. Invertebrate genomes are now supported by a sister project Ensembl Genomes (www.ensemblgenomes.org).

What can I do with Ensembl?

With Ensembl you can:

- Retrieve all or part of a genome sequence.
- Use the sequence alignment search tools BLAST and BLAT against any Ensembl genome.
- Link to genome annotation from microarray results.
- View expressed sequence tags (ESTs), clones, mRNA and proteins for any chromosomal region.
- Predict consequences of sequence variants using the Variant Effect Predictor.
- Examine genes, markers and single nucleotide polymorphisms (SNPs) in a chromosomal region.
- View variations such as SNPs across strains (rat, mouse) or populations (human).
- View all alternative transcripts (splice variants) for a gene.
- View positions and sequence of mRNA and protein that align with an Ensembl gene.
- Explore homologues and phylogenetic trees across more than 30 species for any gene.
- View sequence alignments and conserved regions across species.
- Find possible promoters or gene regulatory regions.

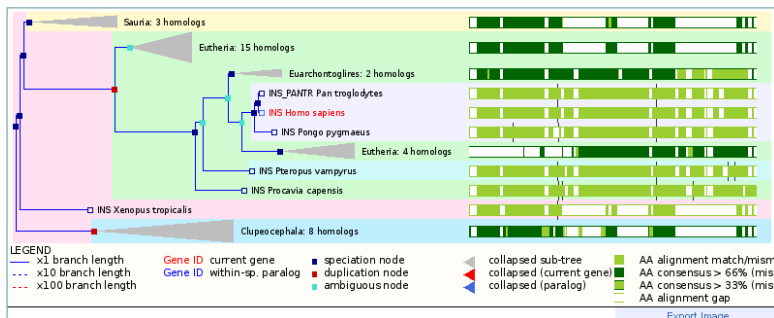
The image displays a screenshot of the Ensembl genome browser interface. At the top, a genomic track shows the SMAD2 gene structure with three alternative transcripts labeled SMAD2-201, SMAD2-202, and SMAD2-203. The transcripts are represented by red lines with exons as boxes and introns as lines with arrows. The genomic sequence is shown as a blue bar. Below the gene structure, a track for 'Sequence variants' shows a list of SNPs. A pop-up window titled 'Variation: rs35641887' is shown on the right, displaying the following information:

Variation Properties	
bp:	between 132848744 & 132848745
status:	-
class:	in-del
ambiguity code:	
mapweight:	1
alleles:	-/C
source:	dbSNP
type:	INTRONIC

On the bottom left, there are four buttons: 'Configure this page', 'Manage your data', 'Export data', and 'Bookmark this page'. A central window shows a list of 'Alternative variants' with columns for variant ID, position, and other details. Arrows indicate the flow of information from the 'Configure this page' button to the variant list, and from a specific variant in the list to the 'Variation: rs35641887' pop-up window.

Three alternative transcripts for the human SMAD2 gene are shown on the Gene Summary page. Views and displayed information can be customised using the 'Configure this page' link (shown bottom left). Transcripts are drawn along the genomic sequence (blue bar) along with variations represented as vertical lines. Gold transcripts represent agreement between automatic and manual annotation. Clicking on any variation creates a pop-up window (bottom right) containing more information and links.

Alignments between genomes in Ensembl can be displayed in several pages. In addition, protein homology is predicted across every species, and gene trees are created based on protein information. These trees can be viewed in the 'GeneTree', a link from the gene summary page.

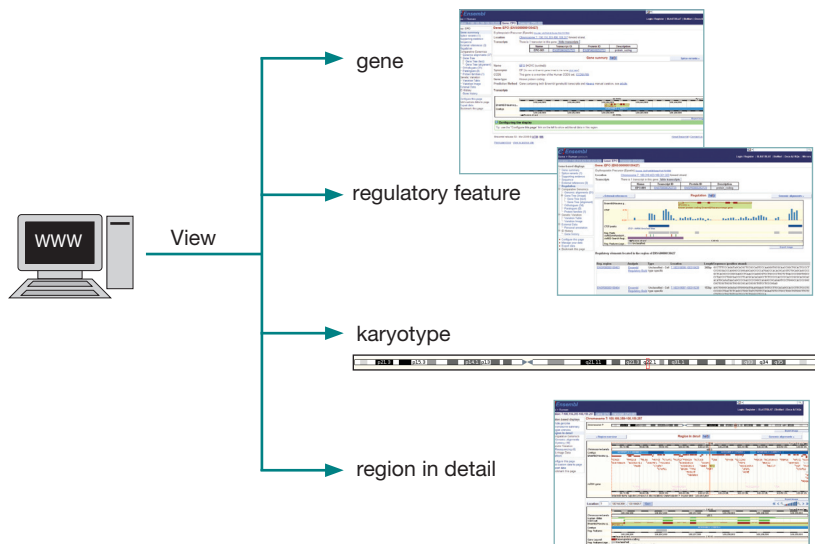


Taxon: Hominae
(Human/Chimp/Gorilla ancestor (9 MYA))

Gene_Count: 2
Branch_Length: 0
Bootstrap: 99
Type: Speciation
Image: expand all sub-trees
Image: collapse this node
Comparison: Jump to Multi-species view
View Sub-tree: Alignment: FASTA
View Sub-tree: Tree: New Hampshire
View Sub-tree: Expand for Jalview

A phylogenetic gene tree (above) for the human insulin precursor gene. The gene tree is used to predict homology, and is the result of multiple sequence alignments using all the species in Ensembl. This gene tree shows speciation events (blue nodes or squares) and duplication events (red squares). Clicking on a node displays an information box, shown left.

- Export sequence, or create a table of gene information using BioMart (see overleaf).
- Upload your own data to the gene, protein, karyotype or location pages.



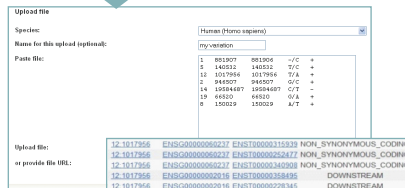
Ensembl can be used to display your own data in a variety of forms. The example above shows the various display options for the human erythropoietin gene, located on chromosome 7.

Accessing Ensembl

The project home page (www.ensembl.org) provides general information and links to home pages for each organism. Data are presented in a series of ‘views’ – from chromosomal region to gene or variation. The species home pages provide quick access to displays of chromosome regions or individual genes and proteins via text searching, interactive chromosome views and BLAST or BLAT similarity searching.

Tool	Description	Online version	API script
Assembly converter	Map your data to the current assembly. Accepted file formats: GFF, GTF, BED, PSL N.B. Export is currently in GFF only	Online version	API script
ID History converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	Online version (max 30 ids)	API script
Variation Effect Predictor	(Formerly SNP Effect Predictor). Upload a set of SNPs in our standard format and export a file containing consequence types. Uploaded tracks can also be viewed on Location pages.	Online version (max 750 SNPs)	API script

Ensembl implements tools in the form of a web interface and API script. Try our *Variation Effect Predictor* to determine whether a variation is already known and its effect (if any) on a transcript.



Variation: rs7578997

Variation class: SNP (rs7578997) source dbSNP_131 - Variants (including SNPs and indels) imported from dbSNP

Synonyms: Illumina_Human660W_quad rs7578997
Illumina_C450_2K rs7578997
Illumina_HumanM_DexV3 rs7578997
dbSNP rs7578997 rs2019284

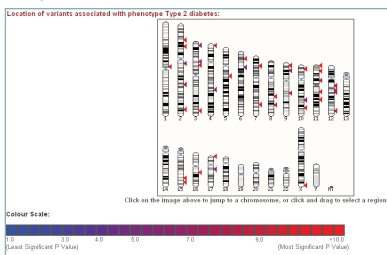
Present in: 1000 genomes (1000 genomes - Low coverage, 1000 genomes - Trio - YRI), Phenotype-associated variations (NHGRI_GWAS_catalog), HapMap (HapMap Phase II, HapMap Phase III), Clinical_SDB variations from dbSNP, ENSEMBL_Venter

Alleles: T/C (ambiguity code Y)
Accepted allele: T

Location: This feature maps to 2-43732023 (forward strand) | [View in location tab](#)

Linked variations: Phenotype Data: Polygenic Context

Hide/Show columns	Source	Study	Associated Gene(s)	Associated variant	Strongest risk allele	P value
Type 2 diabetes (T2D) View on Karyotype	DIHGRI_GWAS_catalog	pubmed18372903	THADA	rs7578997	rs7578997:T	1E-9



Variation pages also show disease and phenotype association from projects like the National Human Genome Research Institute’s GWAS Catalog and the European Genome-phenome Archive (EGA).

Retrieving data from Ensembl with BioMart

BioMart’s web interface can be used to extract information from the Ensembl databases without the need for programming knowledge. It can be used to output sequences or tables of genes, gene positioning information (chromosome and base pair location), and other annotation in HTML, text, or Microsoft Excel format. Among other features, users can translate lists of IDs from one database to another (such as RefSeq IDs to Ensembl genes), list all SNPs for a gene, select genes on any chromosome using an InterPro domain, export gene expression data (for example, GNF data). Data from separate sources can be mined using this integrative tool. Learn how to use BioMart in this video: www.ensembl.org/Help/Movie?id=89

In addition to the web browser, the Perl API (application programming interface) or direct MySQL queries can be used to extract information from the MySQL relational databases. Tables and flat files can be downloaded from the ftp site: www.ensembl.org/info/data/ftp/index.html

The complete software system is also freely available, and is used by groups in academic and industrial settings worldwide. Users are encouraged to become involved in further development of the project; subscribe to ensembl-dev at www.ensembl.org/info/about/contact/mailling.html

Further reading

Flicek, P. *et al.* Ensembl 2011. *Nucleic Acids Res.* 39. Database issue, D800-D806
doi: 10.1093/nar/gkq1064

Ensembl Update 2010. *BMC Bioinformatics*, 11, 240. Series of papers available at: www.biomedcentral.com/series/ENSEMBL2010

www.ensembl.org/info/about/publications.html

Workshops

In the past two years Ensembl gave over a hundred workshops in its genome browser - worldwide! If you would like to learn more about hosting a workshop, visit: www.ensembl.info/workshops

About Ensembl

Ensembl is a joint project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute.

Ensembl is funded primarily by the Wellcome Trust with additional support from the European Molecular Biology Laboratory, the US National Institute of Allergy and Infectious Diseases, the European Union, the UK Biotechnology and Biological Sciences Research Council and the UK Medical Research Council.