Ensembl EMBL-EBI wellcome trust sanger institute **Tutorial**

Getting started with Ensembl

www.ensembl.org

Ensembl provides genes and other **annotation** such as regulatory regions, conserved base pairs across species, and mRNA protein mappings to the genome. These data are accessible via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interface (**Perl API**). Ensembl comparative analyses, variation mappings and gene determinations are freely available to the scientific community within the context of genomic assemblies. The Ensembl gene set reflects a comprehensive transcript set based on protein and mRNA evidence in **UniProt** and **NCBI RefSeq** databases. 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! Learn how to use the Ensembl genome browser and BioMart by following this introductory e-learning course, consisting of 6 modules outlined below. It is based on a recent update of the Ensembl genome browser. To use exactly the Ensembl used in these tutorials, follow along at the archive site for Ensembl v52 http://dec2008.archive.ensembl.org/index.html. Or, give the live site (with the most recent updates) a try at www.ensembl.org!

Modules

1 Introduction to Ensembl
2 Search for a gene in any species
3 View information in the **Gene** and **Transcript** pages
4 View a region of a chromosome in the **Location** page
5 Using **BioMart** to export sequence and gene annotation

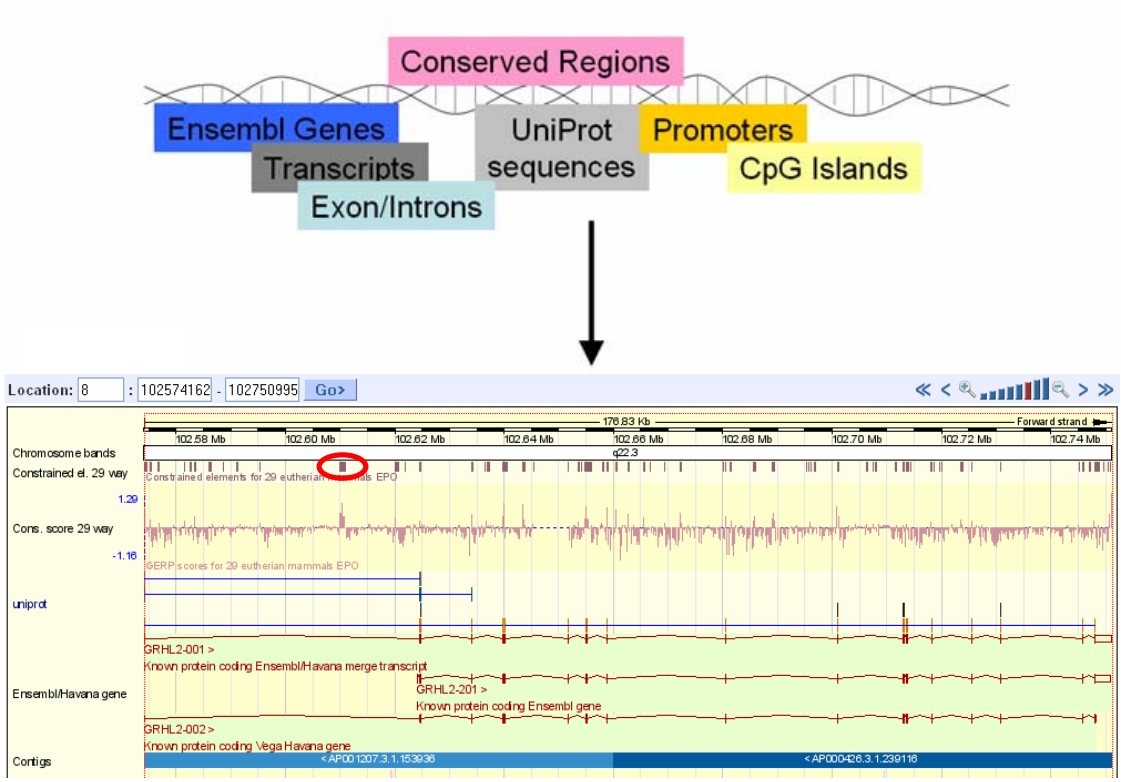
Module 1 – Introduction

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

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 fly 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 52), however the gene sets are determined on the order of once a year. A new browser at www.ensemblgenomes.org is now being set up to access non-chordates such as bacteria, plants, fungi and more.



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 (covered in **module 4**) for the GRHL2 gene in human. The example above shows modules of conserved sequence reflecting conservation scores of sequence identity on a base pair level across 29 species. Conserved regions are displayed as dark modules that represent local regions of alignment. One of the modules is circled in red. You would only have to click on this module to see which regions of the 29 species are highly conserved.

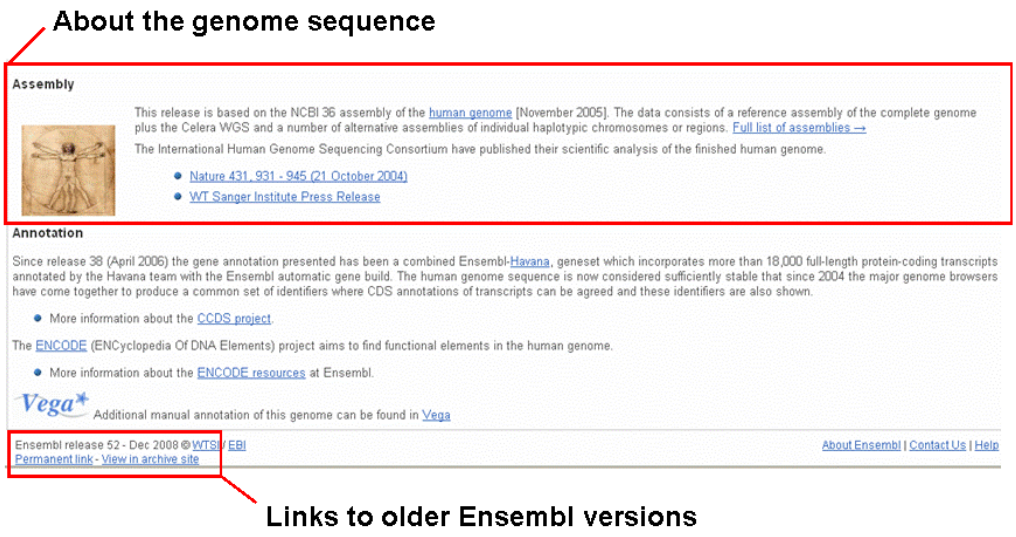
Just under the modules depicting the alignments are blue 'UniProt' tracks. These show 3 proteins from the UniProt knowledge-base (**UniProtKB**) that align to this region of the human genome. The aligned protein sequence to the genome is shown as filled boxes, and connecting lines are gaps in the alignment. You may see some of the aligned sequence in these UniProt proteins match to exons of the three Ensembl transcripts below. 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**). For more about Vega and Havana, see **module 3**. The 'Region in Detail' view, which will be explained in more detail in **module 4**, is useful for comparing Ensembl gene models with current proteins and mRNAs in other databases like **NCBI RefSeq**, **EMBL-Bank**, and, in this example, UniProt. Everything in this view is aligned to the genome (the blue bar).

Ensembl includes a variety of pages in its genome browser that can be accessed by anyone via the world-wide web. In addition to human, mouse, and rat (Ensembl's main foci) annotations are provided for zebrafish, chicken, cow, dog, chimpanzee, platypus, yeast, soil worm and more.

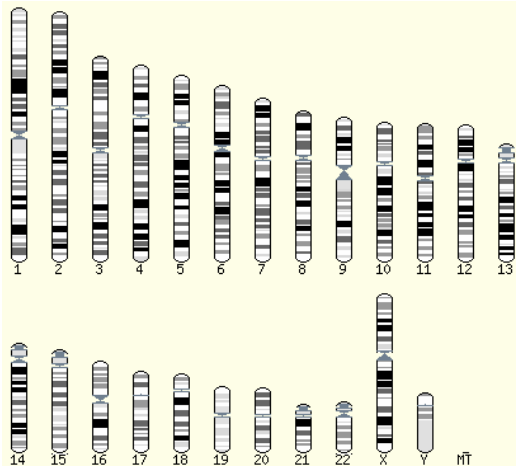
Although the focus is on chordates, Ensembl does have imported gene sets for model organisms such as yeast, worm and fly.

How are Ensembl genes and transcripts determined? All Ensembl transcripts are based on proteins and mRNA from the **UniProtKB** database (UniProt/Swiss-Prot and UniProt/TrEMBL) and **RefSeq.** These proteins and mRNAs are aligned against a genomic sequence assembly imported from a relevant sequencing centre or consortium. (Ensembl doesn't sequence the genomes, or assemble them.) Transcripts are clustered into the same gene if they have overlapping coding sequence. Each transcript is given a list of mRNAs and proteins it is based upon. This 'supportive evidence' underlying every Ensembl transcript can be accessed through the browser.

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.



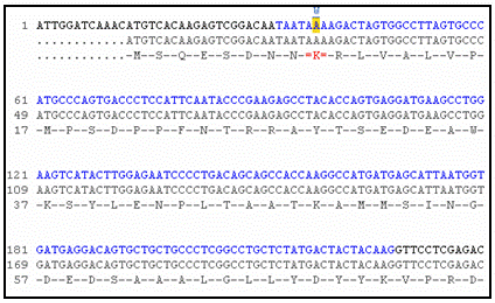
Links to the human karyotype, a summary of gene and genome information, and the most common **InterPro** domains in the genome are found at the left of this index page.



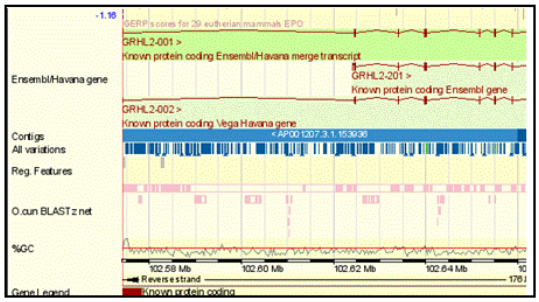
Summary

Assembly:	NCBI 36, Oct 2005
Database version:	52.36n
Base Pairs:	3,253,037,807
Golden Path Length:	3,093,120,360
Genebuild by:	Ensembl
Genebuild started:	Dec 2006
Genebuild released:	Oct 2007
Genebuild last updated/patched:	Oct 2008

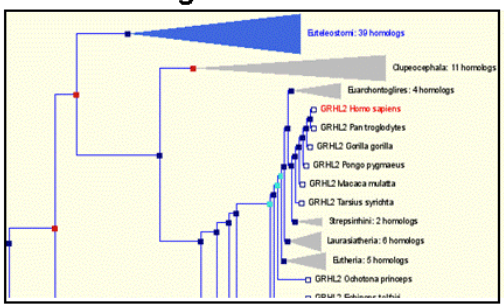
Transcript Sequence w/Variations



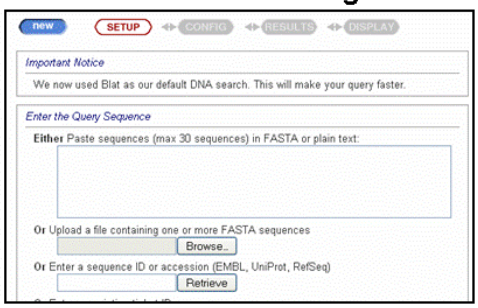
Genes, SNPs, and Conserved Regions



Homologues in Gene Trees



BLAST and BLAT aligners



Ensembl devotes separate pages and views in the browser to display a variety of information types. View variations within the cDNA sequence alongside the protein translation in the **transcript** pages **(module3)**. Compare conserved regions with the position of genes and population variation in the **Region in Detail** view **(module 4)**. See homology relationships in the **gene trees**, or perform a **BLAST** or **BLAT** search against any species in Ensembl.

Sequence Alignments and Homology

Whole genome alignments are available in Ensembl. Most alignments are pairwise, between two species, though we do have some multi-species alignments available. Global aligners have been developed within the comparative genomics project (see reference at the end of this tutorial). In addition, protein homology is predicted for every species in Ensembl through the construction of phylogenetic gene trees using the longest translation of each gene in Ensembl. These **gene trees** can be accessed from the **Gene** page (see **module 3**). For more information see:

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

Retrieving Data from Ensembl

BioMart (www.biomart.org) 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 (see **module 5**).

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 40 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

Contact us/ Workshops

Any questions or comments can be sent to our helpdesk at:
helpdesk@ensembl.org

In the last two years Ensembl gave over a hundred workshops in how to use the genome browser, worldwide! If you would like to host a workshop, please contact us:

http://www.ensembl.org/info/about/outreach/index.html

Glossary

Annotation - The addition of information related to genes or genomic sequence

API (Application Programming Interface) – A series of modules written to interact with a database for easy extraction of information. In Ensembl, it is written in Perl and updated every release. See more here: www.ensembl.org/info/docs/api

BioMart - A popular interface that extracts genes and associated annotation from the database and presents it as formatted tables in Microsoft Excel, HTML or txt in response to the user's specification. Flexible and fast, BioMart can also be used to export sequences, or to connect information across different databases. See module 5.

BLAST- Basic Local Alignment Search Tool - A sequence-alignment program that searches a sequence database to find the optimal alignment to a query.

BLAT – BLAST-like Alignment Tool – A sequence-alignment program on the nucleotide level similar to BLAST, but quicker and demanding of exact-matches. (See reference at the end).

Chordate – An animal belonging to the phylum Chordata, which is a group of animals that have, atleast at one time in their life cycle, a notochord (hollow, dorsal nerve chord).

Contig – In Ensembl, a contiguous region of sequence information imported from a sequence consortium as a part of the sequence assembly.

EMBL-Bank – A database of nucleotide sequences maintained at EMBL-EBI. www.ebi.ac.uk/embl/

Gene tree – (In Ensembl) Phylogenetic trees used to determine homology information between the longest protein for every gene in the 40+ species in Ensembl. Read more about the prediction method here: www.ensembl.org/info/docs/compara/homology_method.html

GO (Gene Ontology) - An organized hierarchy of terms associated with genes and proteins produced by the Gene Ontology Consortium www.geneontology.org. These are used to classify genes and proteins according to biological processes, cellular components, and molecular functions.

InterPro – A database of protein families, motifs, domains and structures http://www.ebi.ac.uk/interpro/

SNP- Single Nucleotide Polymorphism A common sequence variation involving one or a few base pair changes that occurs in DNA at a stable level within a population. Most SNPs in Ensembl are imported from dbSNP (the SNP repository maintained by NCBI).

Syntenic region – a long (100kb or more) continous section of nucleic acids with conserved identity.

UniProtKB - A database of protein information http://www.uniprot.org/

UTR (Untranslated Region) - The untranslated regions of the spliced mRNA. 5' UTR is the portion of an mRNA from the 5' end to the position of the first codon used in translation. The 3'

UTR is the portion of an mRNA from the position of the last codon that is used in translation to the 3' end.

RefSeq - A database of gene and protein information maintained at NCBI
http://www.ncbi.nlm.nih.gov/RefSeq/

VEGA or Havana – The Vertebrate Genome Annotation (VEGA) http://vega.sanger.ac.uk consortium manually annotates specific human, mouse, and zebrafish clones. The gene set from Havana (the part of the VEGA consortium based at the Wellcome Trust Sanger Institute) is compared with the Ensembl transcripts. Identical transcripts are merged, and the remaining are also included in the Ensembl genome browser and corresponding databases.

Did You Know?

Ensembl has a glossary of terms here:

http://www.ensembl.org/common/Help/Glossary

Further Reading

Hubbard, T.J.P. *et. al*, **Ensembl 2009**. *Nucleic Acids Res.* **37,** D690-D697 (2009)

Kent, WJ, **BLAT—the BLAST-like alignment tool** *Genome Res.* Apr 12(4):656-64 (2002)

See more at our publications page:
www.ensembl.org/info/about/publications.html

What to do next

Search for a gene in the next module (**module 2**) to enter the pages of the browser. Or, use our BioMart tool to extract data from the databases in **module 5**.

Watch the related video!

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

'Browsing Ensembl'