Ensembl Genomes (non-chordates): Quick tour

Paul Kersey [1]
DNA & RNA
Beginner
0.5 hour

This quick tour provides a brief introduction to Ensembl Genomes [2], the non-chordate genome browser.

Learning objectives:

- Basic understanding of Ensembl Genomes and how it can help you to access and analyse genome-scale data
- Know where to find out more about the Ensembl Genomes resource

What is Ensembl Genomes?

Ensembl Bacteria [3], Protists [4], Fungi [5], Plants [6] and Metazoa [7] (collectively, ‘Ensembl Genomes’) are five portals for genome-scale data, developed in close collaboration with scientific communities expert in the biology of individual species. Implemented using the Ensembl software suite for genome analysis and browsing, which was developed for the study of vertebrate genomes (described by Flicek et al., 2012, Ensembl 2012, Nucleic Acids Res. 40 D84-90 [8]), Ensembl Genomes [9] provides a powerful and consistent set of interactive and programmatic interfaces for non-vertebrate genomes, providing access to data including gene predictions, comparative analysis, variant annotation, and transcriptomic alignments. Since its establishment in 2009, the resource has grown rapidly and now contains 70 eukaryotic and 249 bacterial species. Moreover, recent improvements to the site make it easier than ever to upload your own data and to visualise and analyse it in the context of reference annotations (Figure 1).

![Figure 1](image-url) Ensembl Genomes entry page showing the five portals on the right-hand side.

Data sources
Ensembl Genomes (non-chordates): Quick tour

Ensembl Genomes provides access to genome-scale data through a number of interfaces, including a web browser, a search-optimised data warehouse [10], bulk download and various programmatic interfaces. The data come from a variety of sources, including collaborators in the scientific community, publicly available data archives, and computational analysis pipelines run at the EBI and elsewhere. As the scientific scope of the project is exceptionally broad, our goal is to provide an up-to-date view of core annotation as recognised by the relevant scientific communities, integrated with data from other species through the use of shared interfaces and comparative analysis.

Wherever possible, we actively collaborate with community groups in the management of data and the development of services, including:

- **WormBase** [11] (for nematodes)
- **VectorBase** [12] (for invertebrate vectors of human pathogens)
- **Gramene** [13] (for plants)
- **PomBase** [14] (for the fission yeast *Schizosaccharomyces pombe*)
- **PhytoPath** [15] (for plant pathogens)
- **CADRE** [16] (for *Aspergilli*).

We also show canonical annotations from leading model organism databases such as:

- **DictyBase** [17] (for *Dictyostelium discoideum*)
- **FlyBase** [18] (for *Drosophila*)
- **SGD** [19] (for *Saccharomyces cerevisiae*).

**What can you find in Ensembl Genomes?**

The current Ensembl Genomes [9] portals (as of March 2012) are summarised below. They are: Ensembl Metazoa [20], Plants, Protists, Fungi and Bacteria. New genomes are added to the project with each release.

**Ensembl Metazoa** [7]

Contains data from non-chordate [21] metazoan species, including species annotated by WormBase [22], FlyBase [23] and VectorBase [24]: worms, flies and arthropod [25] vectors of human pathogens. Specifically, these include 12 genomes of the genus *Drosophilida*, four genomes of nematode worms and five genomes of insect pathogens (three mosquitoes, the body louse *Pediculus humanus*, and the black-legged tick *Ixodes scapularis*). Additionally, a further seven species are included from outside these groups, ranging from other arthropoda (e.g. the honey bee) to basal metazoans such as *Trichoplax adhaerens*.

**Ensembl Plants** [6]

Developed jointly with Gramene (a resource for plant genomics based at the Cold Spring Harbor Laboratory) and contains data for ten flowering plants (six monocots and four dicots [26]) and one species of moss. Variation databases (recording genome-wide polymorphism [27] across populations) are provided for *Arabidopsis*, rice and grape.

**Ensembl Protists** [4]

Includes the genomes of a number of Apicomplexa species (including the causative agents of
malaria), *Leishmania major*, the slime mould *Dictyostelium discoideum*, two diatom genomes and four plant pathogens.

**Ensembl Fungi** [5]

Includes the genomes of model species such as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Neurospora crassa*, twelve phytopathogens, and data from eight species of *Aspergillus*, prepared in collaboration with the Central *Aspergillus* Data REpository (CADRE [16]).

**Ensembl Bacteria** [3]

Covers a number of clades, each represented by multiple genomes. The bacterial clades are *Escherichia/Shigella* and *Bacillus* (home to the model organisms *E. coli* and *B. subtilis*), and five important clades of pathogenic bacteria: *Borrellia*, *Mycobacterium*, *Neisseria*, *Staphylococcus* and *Streptococcus*, and two genera of arthropod symbionts, *Buchnera* and *Wolbachia*. A further clade covers the archaeal genus *Pyrococcus*. The genomes of between 4 and 77 related strains are captured within each database, and access to inter- and intra-clade comparative genomics is provided.

**What can I do with Ensembl Genomes**

**Using Ensembl Genomes data**

The genome is a natural entry point for many types of bioinformatics data, providing both a reference framework and a scientific context within which the results of transcriptomic or proteomic data can be interpreted.

Using Ensembl Genomes [9], you can:

- Retrieve all or part of a genome sequence.
- Use the sequence alignment search tool BLAST [28] against any genome.
- Examine features (e.g. protein coding genes, non-protein coding genes and SNPs) in a chromosomal region.
- View all alternative transcripts (including variants caused by alternative splicing and promoter usage) for a gene.
- View positions and sequences of mRNAs and proteins that align with a gene.
- Explore homologues and phylogenetic trees across, within and between clades.
- View sequence alignments and conserved regions across species.
- Export sequence, or create a table of gene or SNP information using BioMart (see Getting data from Ensembl Genomes [31]).
- Upload your own data, including next-generation sequencing reads (in common alignment formats such as BAM [32]) to view in the context of public annotation.
- Predict the effect of nucleotide sequence variants on genes and their products.

The taxonomic spread of Ensembl Genomes allows you to view comparative genomic information at many levels, focusing either on a narrow taxonomic range (potentially as small as one bacterial species) or a pan-taxonomic focus (from microbe to man). Additionally, the provision of data for pathogenic species, their vectors and their hosts through a common interface provides a unique
resource for the study of pathogen-mediated diseases of medical and agricultural importance.

**Genome browser**

The genome browser is the main point of access for most users of Ensembl Genomes [9]. The browser displays gene structures, supporting evidence, cross-references, genome-wide assays and a range of other features. It is easily configurable and functions as a client for the DAS [33] protocol, allowing you to rapidly integrate additional data.

The browser enables you to upload your own next generation RNA sequence data in common file formats (such as BAM [32]). You can also visualise your alignments alongside the reference annotation [29] (Figures 2 and 3).

**Figure 2** Visualising a BAM file in Ensembl Genomes. In this example, conventional EST data (green tracks) are visualised against the reference gene annotation [34], and a BAM file (showing RNA-seq [35] data) has also been visualised. Individual read alignments (grey tracks, with red ends on each read), and overall read coverage at each position (top track, grey), are shown.
**Figure 3** Variant annotation in Location View. This example shows a simple visualisation of sequence variations identified at the displayed locus, colour coded by their functional effects on their most proximate gene. This is one of many views that provide access to variation data in Ensembl Genomes.

**Bacterial genomes**

Ensembl Bacteria [3] provides special graphical representations to support the correct depiction of bacterial genomes, including circular chromosomes, alternative translational initiation and polycistronic transcripts (Figure 4).
Figure 4 Ensembl Bacteria web page showing the circular navigation tool. Users can click on the round handles at the end of the selected sector (shown in red) to adjust the region shown in the 'Region in detail' window.

Tools for comparative genomics

The Ensembl [36] comparative genomics analyses (Ensembl Compara [37]) are applied to all domains within Ensembl and Ensembl Genomes [9]. The DNA comparison modules are suitable for use over a narrow taxonomic range, and can be used to support multiple sequence alignments and
to make predictions of ancestral DNA sequence (Figure 5). The protein comparison modules compute the evolutionary history of a gene family, and can be applied over much larger evolutionary distances. A 'gene tree' [38] provides a summary of this information (and the anchor point for outward links to pairwise and multiple-sequence alignments, and orthologue [39]/paralogue [40] lists).

**Figure 5** Visualisation of homology relationships in Ensembl Genomes: viewing only the current gene CRP34 and its immediate orthologues. Using the view options, you can expand selected nodes to see more distant branches of the tree in more detail.

**Visualising genomic polymorphism**

The Ensembl variation schema is used to capture information from population-wide surveys of sequence polymorphism [27], and is currently populated for one mosquito, two fungi, one protist, and four plant species. Data are available for download or can be visualised in the genome browser (Figure 6).

**Figure 6** Visualising genomic variation in a genetic context. The figure shows the location and functional consequence of genomic polymorphisms in the context of a neighbouring gene, and the domain structure of the protein it encodes.
Getting data from Ensembl Genomes

Ensembl Genomes [9] provides access through a variety of interfaces, including a graphical web browser [41] with text and sequence search, an FTP server [42], a Perl [43] interface for programmatic access, and a public MySQL [44] server. Additionally, Ensembl Genomes data is provided in a search-optimised data warehouse [10] (BioMart) that provides facilities for fast, customisable download and analysis. You can link searches between different BioMarts in separate locations, allowing you to combine Ensembl Genomes data with information from external resources, and to construct searches across the genomes of different taxonomic groups.

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on Ensembl Genomes

Support

- There are separate help pages for each Ensembl Genomes [9] portal. For example, for Metazoa [20], the help pages and documentation can be found here [45].
- To use help for a different Ensembl Genomes portal, simply replace the word 'metazoa' in the URL with the names of the other portals (bacteria, fungi, plants, protists).
- Any questions or comments can be sent to our helpdesk [46].

References

Ensembl Genomes


Ensembl BioMarts


Collaborators

Selected databases in Ensembl Genomes have been built by a number of collaborating resources, including the Central Aspergillus RESource [16] (CADRE), Gramene [13], PomBase [14], PhytoPath [15], VectorBase [12], and WormBase [11].
Funding

Ensembl Genomes is supported by the European Commission [51] within its 7th Framework Programme, under the thematic area "Infrastructures", contract number 283496 ("transPLANT"); by the award of BBSRC [52] grants BB/F019793/1, BB/1001077/1 and BH/H531519/1; by the award WT090548MA from the Wellcome Trust [53]; and by the Bill and Melinda Gates Foundation [54].

Contributors

Paul Kersey [1]

EMBL-EBI
Team leader

Paul Kersey did his PhD at the University of Edinburgh on the molecular biology and genetics of the cell cycle of fission yeast, and subsequently worked at the MRC Human Genetics Unit. Since 1999, Paul has worked at EMBL-EBI on various databases, including UniProt, GOA, Integr8, IPI, and Genome Reviews. He currently leads the Ensembl Genomes team, which provides services for genome-scale data from non-vertebrate species, primarily through the use of the Ensembl software platform. The team integrates data from across the taxonomic space, and collaborates closely with particular groups involved in genome annotation and the development of domain-specific resources.

Source URL: http://www.ebi.ac.uk/training/online/course/ensembl-genomes-non-chordates-quick-tour

Links
[1] http://www.ebi.ac.uk/training/online/trainers/pkersey