

Need help?

www.ensemblgenomes.org/info/helpdesk

Any questions or comments can be sent to our helpdesk at:

helpdesk@ensemblgenomes.org

Dr Paul Kersey
EMBL-EBI
Wellcome Trust Genome Campus
Cambridge
CB10 1SD, UK

Ensembl Genomes

www.ensemblgenomes.org

Ensembl Bacteria, Protists, Fungi, Plants and Metazoa (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 (Flicek *et al.* 2011), Ensembl Genomes 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.

What is Ensembl Genomes?

Ensembl Genomes provides access to genome-scale data through a number of interfaces, including a web browser, a query optimised data warehouse, 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. A broad palette of expertise would be required to annotate every genome within the scope of the project, so our goal is to provide an up-to-date view of the core annotation as recognised by the scientific community, 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 (for nematodes), VectorBase (for invertebrate vectors of human pathogens), Gramene (for plants), PomBase (for the fission yeast) and CADRE (for *Aspergilli*). We also show canonical annotations from leading model organism databases such as DictyBase, FlyBase and SGD.

The current set of Ensembl Genomes (as of May 2011) consists of the following:

Ensembl Metazoa contains data from non-chordate metazoan species, including worms, flies and arthropod vectors of human pathogens, representing species annotated by WormBase, FlyBase and VectorBase. Specifically, these include 12 genomes of the genus *Drosophila*, 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*.

EMBL-EBI



Services | Research | Training | Industry

The screenshot shows the Ensembl Genomes website interface. At the top, there are navigation tabs for Bacteria, Protists, Fungi, Plants, Metazoa, and Vertebrates. The main content area is titled 'Ensembl Genomes: Extending Ensembl across the taxonomic space.' Below this, there is a 'Releases' section with a sub-heading 'What's New in Release 9 (19th April 2011)'. The text describes updates to the Ensembl software across all datasets and the addition of new genomes in various taxonomic groups. It lists updates for Ensembl Bacteria, Ensembl Protists, Ensembl Fungi, Ensembl Plants, and Ensembl Metazoa. A 'Future Releases' section mentions the scheduled release for 12th July 2011. There is also a 'Working with communities' section at the bottom right, which lists partners like the Original AgriBio One Network, GRAMENE, and VectorBase.

Ensembl Genomes entry page.

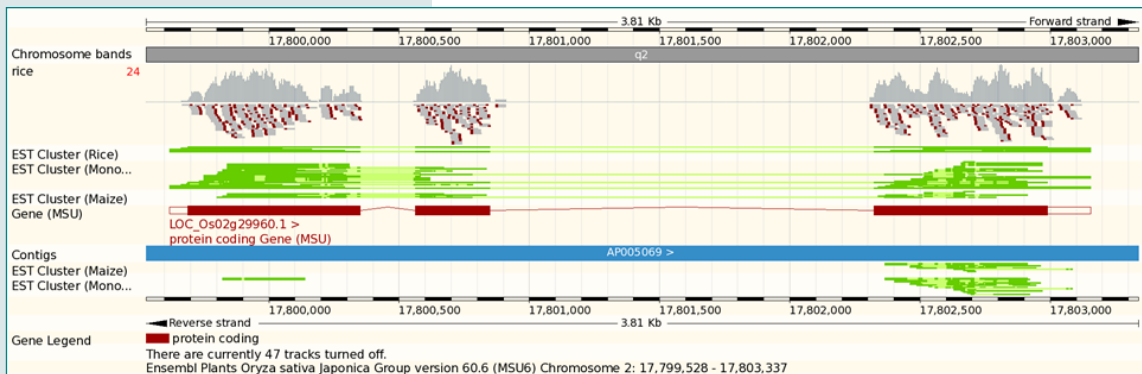
Ensembl Plants is developed jointly with Gramene (a resource for plant genomics based at the Cold Spring Harbor Laboratory) and contains data for ten flowering plants (six monocots and four dicots) and one species of moss. Variation databases (recording genome-wide polymorphism across populations) are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

Ensembl Protists includes the genomes of a number of Apicomplexa species of human pathogens (including the causative agents of malaria), *Leishmania major*, the slime mould *Dictyostelium discoides*, two diatom genomes, and four plant pathogens.

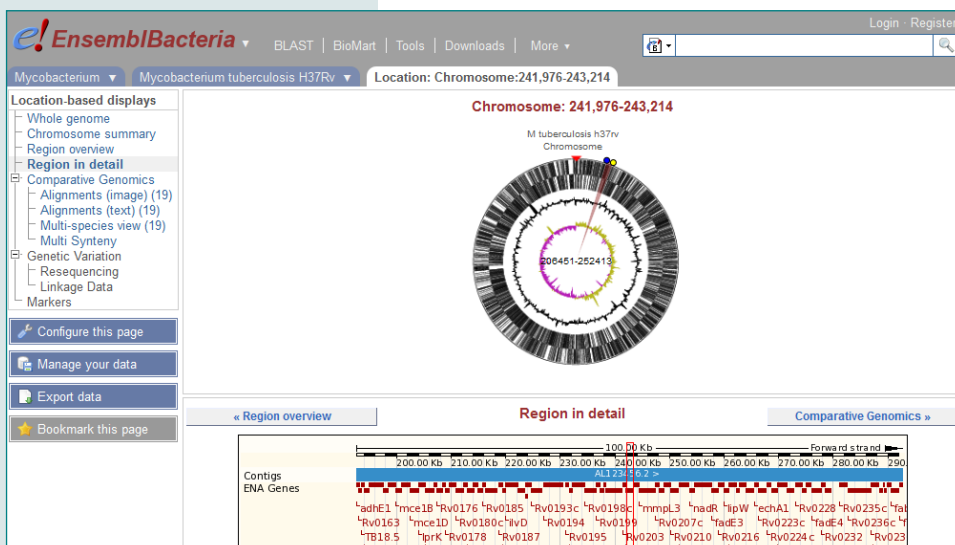
Ensembl Fungi includes the genomes of model species such as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Neurospora crassa*, four fungal phytopathogens, and data from eight species of *Aspergillus*, prepared in collaboration with the Central Aspergillus Data Repository (CADRE).

Ensembl Bacteria 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: *Boreilia*, *Mycobacterium*, *Neisseria*, *Staphylococcus* and *Streptococcus*, and two genera of arthropod symbionts, *Buchnera* and *Wollbachia*. 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.

New genomes are added to the project with each release.



Upload your own next generation RNA sequence data in common file formats (such as BAM). You can also visualise your alignments alongside the reference annotation.



Special graphical representations have been developed to support the correct depiction of bacterial genomes, including circular chromosomes, alternative translational initiation and polycistronic transcripts.

What can I do with Ensembl Genomes?

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, you can:

- Retrieve all or part of a genome sequence.
- Use the sequence alignment search tool BLAST against any genome.
- Link to genome annotation from microarray results.
- 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 overleaf).
- Upload your own data, including next-generation sequencing reads (in common alignment formats such as BAM) to view in the context of public annotation.
- Predict the effect of nucleotide sequence variants on genes and their products.

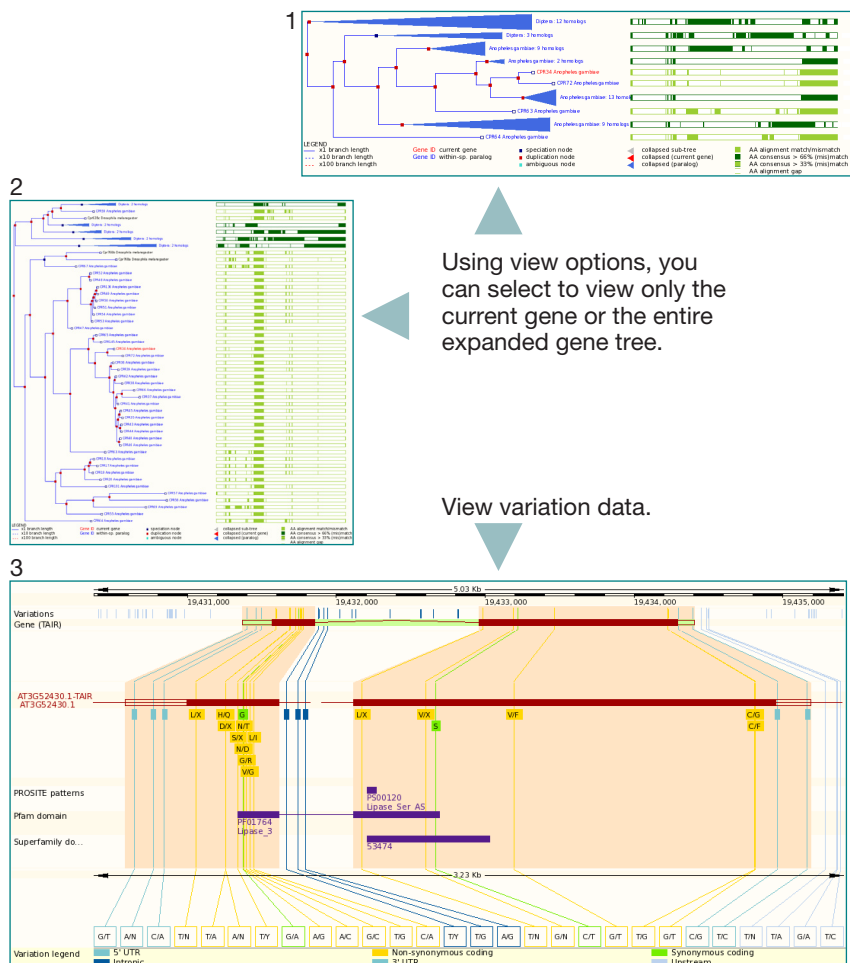
The taxonomic spread of Ensembl Genomes allows users to view comparative genomic information at many levels, focusing either on a narrow taxonomic range (potentially as small as one bacterial species) or a pan-taxonomic focus (from microbe to man). Additionally, the provision of data for pathogenic species, 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.



The genome browser is the main point of access for most users of Ensembl Genomes. 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 protocol, allowing the rapid integration of additional data.

Tools for comparative and population genomics

The Ensembl comparative genomics pipeline, Ensembl Compara, is applied to all domains within Ensembl and Ensembl Genomes. The DNA comparison module is 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. The protein comparison module computes gene trees for the identification of evolutionary relationships, and can be applied over much larger evolutionary distances. The Ensembl variation schema is used to capture information from population-wide surveys of sequence polymorphism, and is currently populated for one mosquito, one yeast, one protist, and four plant species. Data are available for download or can be visualised in the genome browser.



Visualisation of homology relationships (images 1 and 2) and genomic polymorphism (image 3) in Ensembl Genomes.

Retrieving data from Ensembl Genomes

Ensembl Genomes provides access through a variety of interfaces, including a graphical web browser (www.ensemblgenomes.org) with text and sequence search, an FTP server (<ftp://ftp.ensemblgenomes.org/pub>), a Perl interface for programmatic access, and a public MySQL server. Additionally, Ensembl Genomes data is provided in a query-optimised, denormalised data warehouse (BioMart) that provides facilities for fast, customisable download and analysis. Queries can be chained between different BioMarts in separate locations, allowing the combination of Ensembl Genomes data with information from external resources, and the construction of queries between the genomes of different taxonomic groups. ●

About Ensembl Genomes

Ensembl Genomes is supported by the European Commission within Research Infrastructures of the FP7 Capacities Specific Programme as part of the SLING project, grant agreement number 226073 (Integrating Activity); by the award of BBSRC grants BB/F019793/1, BB/1001077/1 and BH/H531519/1; by the award WT090548MA from the Wellcome Trust; and by the Bill and Melinda Gates Foundation.

Selected databases in Ensembl Genomes have been built by a number of collaborating resources, including CADRE, Gramene, PomBase, PhytoPath, VectorBase, and WormBase.