

## Ensembl, a database for comparative genomics.

<http://www.ensembl.org>

William Rossie

## Highlights

- Introduction
- Data handling
- New features
- Enhancements
- Future directions

## Introduction

- Used by comparative genomic scientists.
- Framework to genomic sequence analyses.
- Intergrates manual annotated genes from external sources.
- Leading source of genome annotation.
- Open source project of genome annotation, for
  - Sequence analysis
  - Data storage
  - Visualisation

## Data handling:

### 1. Searching and viewing genomes

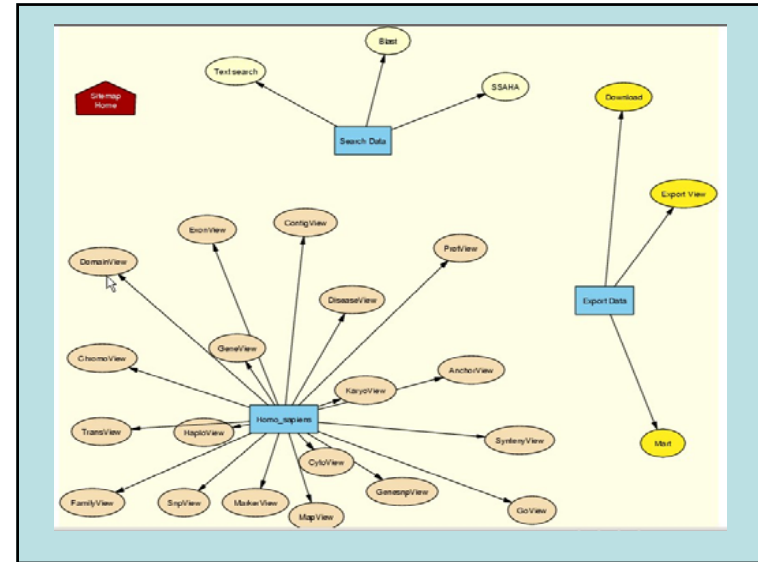
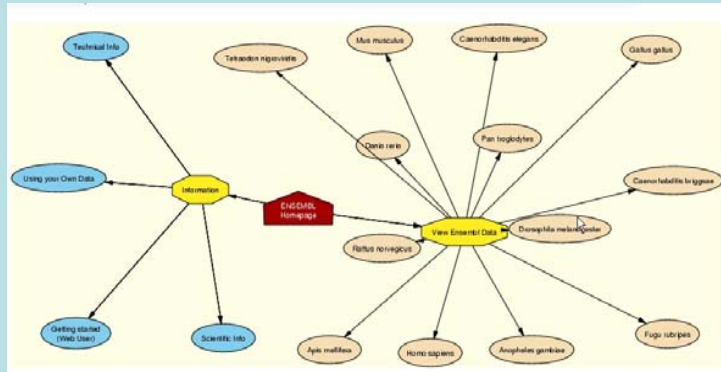
- Compares data from:

- Vertebrates
- Invertebrates
- Worms

Species - Ensembl v26			
Human	pre/	NCBI 34	Jul 04
Mouse		NCBI m33	Jul 04
Zebrafish		WTSI Zv4	Sep 04
Rat		RGSC 3.1	Jul 04
Chicken		WASHUC1	Jul 04
Mosquito		MCZ 2	Apr 04
Fugu		Fugu v2.0	May 04
Fruitfly		BDGP 3.1	Jul 03
Chimp		CHIMP1	May 04
Honeybee		Amel1.1	Sep 04
Tetraodon		TETRAODONT	Sep 04
Cow	pre/	Btau 1.0	
Dog	pre/	BROADD1	
X.tropicalis	pre/	UG3	
C. elegans		WS 118	Apr 04
C. briggsae		cb25.ana8	Jul 03

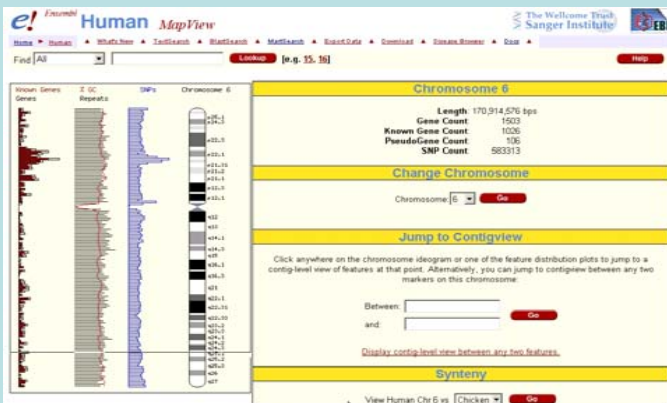
## Data handling:

### 1. Searching and viewing genomes



## Data handling:

### 1. Viewing; Human mapview

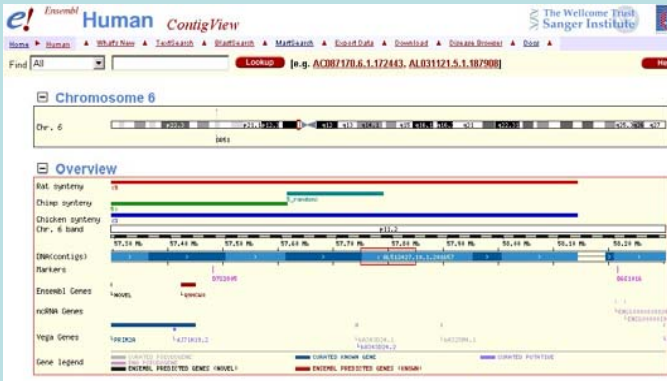


## Data handling:

### 1. Viewing; BLAST view

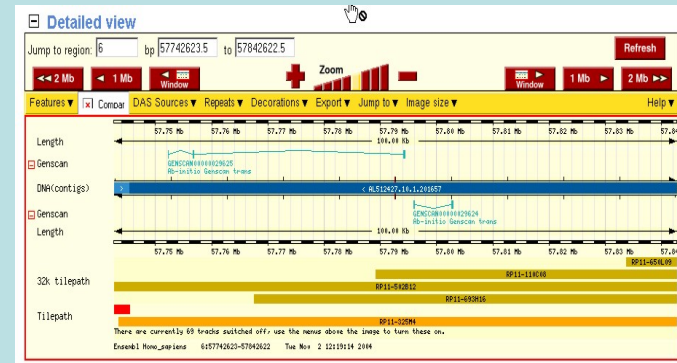
## Data handling:

### 1.Viewing; Contig view (overview)



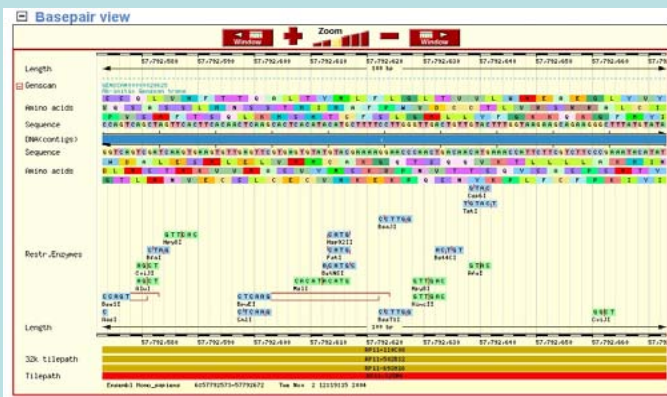
## Data handling:

### 1.Viewing; Contig view (detailed)



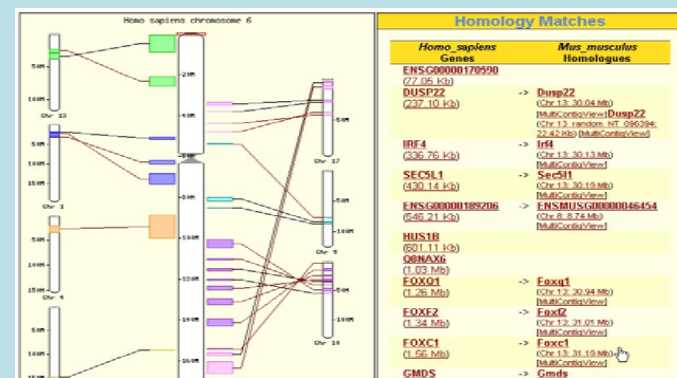
## Data handling:

### 1.Viewing; Contig view (basepair)



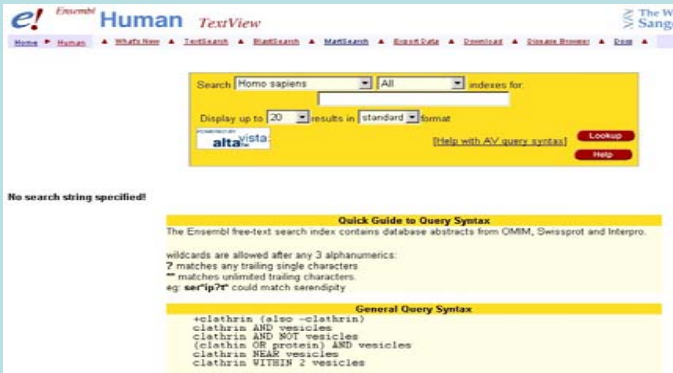
## Data handling:

### 1.Viewing; Synteny view



## Data handling:

### 1. Viewing; Text view



The screenshot shows the Ensembl Human TextView search page. At the top, there is a navigation menu with links like Home, Human, What's New, TextView, BLAST, and more. The main search area has a search box with 'Homo sapiens' selected, a dropdown for 'All' indexes, and a search button. Below the search box, there is a 'Display up to 20 results in standard format' option. A search string 'alta' is entered in the search box. Below the search box, there is a 'No search string specified!' message. To the right, there is a 'Quick Guide to Query Syntax' section with a list of query examples: '+clathrin (also -clathrin)', 'clathrin AND vesicles', 'clathrin AND NOT vesicles', '(clathrin OR protein) AND vesicles', 'clathrin NEAR vesicles', and 'clathrin WITHIN 2 vesicles'.

## Data handling:

### 2. Data uploading

#### Uploading Your Data

You can upload your own data into Ensembl and visualise it on contigview displays along with all existing Ensembl data.

Please **READ THE UPLOAD INSTRUCTIONS CAREFULLY** before uploading any data. Your data must be **formatted correctly**, but instructions page has detailed information about the data formats.

Please read and understand the [Ensembl policy on uploaded data privacy](#).

**Note:** due to server space restrictions there may be a limit imposed on the amount of data you can upload in a single operation and attach them all to a single display.

**Data Expiry:** To prevent large amounts of stale data accumulating on our server we will operate a [data "sweeping" policy](#). If the data is from the server.

If you upload data with features specified in chromosomal (i.e. assembly) coordinates these will be invalidated when a new Ensembl genome assembly is released. It is the responsibility of the user to make sure that their uploaded data coordinates and the Ensembl assembly coordinates are [instructions about uploading data](#) for more information.

Email:

Password:

Upload File:

- or paste data into the box below:

## Data handling:

### 3. Data downloading



The screenshot shows the 'Ensembl Links and Site Map' page. It features a yellow header with the title. Below the header, there are four buttons: 'Download', 'Export', 'EnsMart', and 'BLAST/SSAHA'. To the right of the buttons is a site map diagram showing a central node connected to several other nodes. A mouse cursor is pointing at the 'Download' button.

## Features

- 1. Comparative genome analysis.
- 2. Apollo annotator viewer/editor.
- 3. EnsemblMart (datamining).
- 4. Otter (extended schema for gene curation).

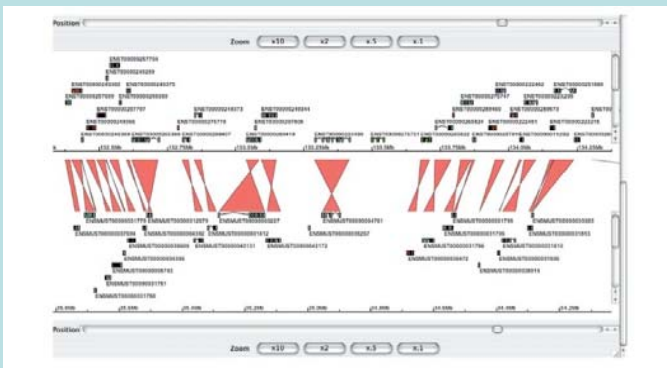
**Features:** 1. Comparative genome analysis.

- Fugu project and Ensembl.
- Linking and defining synteny regions.
- Constructing a catalogue of orthologous genes.
- Types of comparative information:
  - a. Fine grained DNA-DNA analysis.
  - b. Orthologous protein information.
  - c. Large scale synteny data.

**Features:** 2. Apollo.

- Viewer/editor
- Java based browser.
- Open source.
- Support DAS (distributed annotator system)
- Annotator client for Otter.
- Viewing of DNA-DNA alignment, contigview and protein comparison.

**Features:** 2. Apollo; Synteny view



**Features:** 3. EnsemblMart

- Datamining for genomes.
- MySQL database, query optimised.
- View via martview.

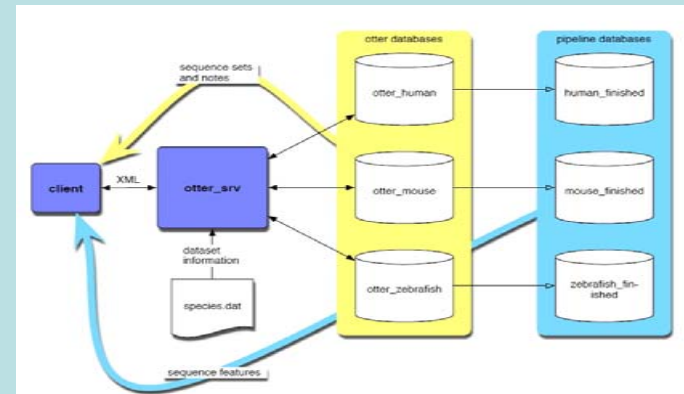
## Features: 4. Otter

- Ensembl database with extended schema for gene curation.
- Associated client/server system.
- Communicate with clients via XML format.

```
<otter>
  <sequence_set>
    <sequence_fragment>
      <accession>
      <locus>
      <transcript>
        <exon>
      <feature_set>

    <stable_id>
    <author>
    <start> <end> <strand>
```

## Features: 4. Otter; overview



## Enhancements

- Ensembl genome annotation.
- Website.
- Software.
- Data analysis pipeline.

## Enhancements

### 1. Ensembl genome annotation.

- Ensembl EST genebuilder: gene prediction.
  - Combination of exonerate, BLAST and EST2Genome.
  - Transcript processed by Genomewise (finding the largest ORF across each one).
- Predicting pseudo genes automatically.
- DNA synteny generated between each group automatically.

## Enhancements

### 2. Website.

- **New tracks:** Eponine track (shows transcription start site prediction).
- **New interface:**
  - Martview (data mining interface).
  - Goview (ontology of gene function process and location terms).
  - Haploview (interface to haplotype data contigview).
- **DAS integration server:**
  - DAS tracks on contigview include;
    - NCBI transcription models.
    - NCBI genomescan prediction.
    - Assembly transcript models.
    - Ensembl mapped RefSeqs.

## Enhancements

### 2. Website.

- **Basepair panel to contigview:**
  - Showing nucleotide
  - Six frame amino acid translation
  - Restriction enzyme site features.
- **Pre-processing of SNP data.**
- **Contigview:** include labelled syntenic blocks.
- **Dotterview:** web interface to DOTTER program (shows dotplot of DNA similarity).

## Enhancements

### 3. Software.

- Reused to build Contigview like webviews of a virtual database composed entirely of different DAS sources.
- Ensembl pipeline used to support gene curation by Wormbase or Havana (stores also gene annotation from Otter).
- **Power the Vega website** (curated annotation of vertebrate genome collection from a number of annotated groups into a single database).

## Enhancements

### 3. Software.

- Hold consistency between biological objects and aware adaptor objects consistent style of function names.
- Perl code base, adaption of parallel java code base, with a common design between the two language bindings.
  - Perl for the biological part.
  - Java for the stable ID transfer and as a backend data adaptor for Apollo.

## Enhancements

### 4. Data analysis pipeline.

- Improvement of processing of ESTs and cDNA as part of the EST gene analysis.
- Adjustment of maximum intron size between vertebrates and invertebrates.
- Improved handling from complex data condition
- Use of different scheduler systems (like PBS GridEngine and LSF).
- Compact storage of gapped alignments.

## Future directions

- Providing genome information.
- Providing baseline annotation for a number of genomes.
- Data expanding of new genomes.
- Technology improvement (complete comparative information between multiple vertebrates).
- Professional uses.
- Technical documentation manuals.
- Intergration with rich ontology based systems (like GenomeKnowledgeBase).

## Conclusion

- Great database for comparative biologists.
- Good links with other databases.
- Up to date.