

## **Module 3: Genes and Sequences (Sanger)**

### **v - Data Mining in Ensembl with EnsMart**

- **Introduce EnsMart, a data mining tool for large datasets**

#### **Introduction**

The EnsMart system extends the Ensembl genome browser's capabilities, facilitating rapid retrieval of customised datasets. A wide variety of complex queries are supported, on various types of annotations, for numerous species. These can be applied to many research problems, ranging from SNP selection for candidate gene screening, through cross-species evolutionary comparisons, to microarray annotation. Users can group and refine biological data according to many criteria, including cross-species analyses, disease links, sequence variations, and expression patterns. Both tabulated list data, and biological sequence output can be generated on the fly, in HTML, text, Microsoft Excel and compressed formats. A wide range of sequence types, such as cDNA, peptides, coding regions, UTRs and exons, with additional upstream and downstream regions, can be retrieved. EnsMart can be accessed via a public web site or through a Java application suite.

## MartView

MartView implements the user interfaces to the system. Ensembl pages have different access points to this view. Click on the link to MartView from the zebrafish Ensembl site.

**Zebrafish Genome Browser**

**Ensembl Entry Points**

Search for  with  [Lookup](#)

Show Chr/FPC  From  To  [Lookup](#)

Retrieve a sequence [Export](#) Advanced data retrieval tool [EnsMart](#)

Search your sequence [BLAST/SSAHA](#)

**Zebrafish Genome Project**

The zebrafish genome project is a collaboration between the Sanger Institute and the zebrafish community, announced during the Sanger Institute Zebrafish Workshop 2000 and was started in February 2001.

This Ensembl website features the zebrafish assembly version 4 (Zv4), as released on the 12th July 2004. This assembly was produced by integrating the whole genome shotgun assembly with data from the physical map ([more information](#)).

Datasets used for the analyses that were provided by collaborators are acknowledged [here](#).

The zebrafish sequencing project is funded by the Wellcome Trust.

You may export data from this site. Please see the [Conditions of use](#) for these data.

**Annotation**

Ensembl zebrafish genes (ENSDAR\*) are generated automatically by the Ensembl gene builder. A limited number of Zebrafish clones have also been manually annotated in Vega and these will be imported into Ensembl once the Zebrafish assembly merges the whole genome shotgun and clone sequence data.

**Current Release 31.4d**

Last Update: 02-09-2004  
 Ensembl gene predictions: 23524  
 Genscan gene predictions: 57411  
 Ensembl gene exons: 214844  
 Ensembl gene transcripts: 32062  
 Clones: 4100  
 Scaffolds: 21333  
 Chromosomes: 25  
 Base Pairs: 1571018465  
 Golden Path Length: 1560425332

[What's New](#)

**Browse a Chromosome**

1 2 3 4 5 6 7 8 9 10 11 12 13  
 14 15 16 17 18 19 20 21 22 23 24 25

**Documentation & Help**

About Ensembl [Home](#)

For context-sensitive help on any web page click [Help](#)

Questions or suggestions? Try [Help Desk](#)

Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site) [Documentation](#)

**Ensembl Links and Site Map**

[Download](#) [Export](#) [EnsMart](#) [BLAST/SSAHA](#)

**Other Species**

[Mosquito](#) [Honeybee](#) [C. elegans](#)  
[Dog](#) [C.intestinalis](#) [Fruitfly](#)  
[Fugu](#) [Chicken](#) [Human](#)  
[Mouse](#) [Chimp](#) [Rat](#)  
[S. cerevisiae](#) [Tetraodon](#) [X. tropicalis](#)

Queries in EnsMart are organised into three steps: start, filter and output. The user can navigate between these three stages using the 'Back' and 'Next' buttons provided. Below is a detailed description of each step using MartView as an example.

## Start

The start stage includes the initial selection of the species and focus for the query. Each species is designated with its genome assembly version. There are three possible foci: Ensembl, SNP and Vega. Select Ensembl.

The screenshot shows the bioMart interface at the 'START' stage. The top navigation bar includes buttons for 'new', 'START', 'FILTER', 'OUTPUT', and 'export'. Below this, there are 'new' and 'next' buttons. The main content area is titled 'Select the dataset for this query' and contains two dropdown menus: 'Database: Ensembl 31' and 'Dataset: Danio rerio genes (ZFISH4)'. A yellow box labeled 'Focus' points to the 'Dataset' dropdown, and another yellow box labeled 'species' points to the 'Database' dropdown. Below the dropdowns, there is a section titled 'Using MartView' with instructions on how to use the system. On the right side, there is a 'Summary' sidebar with buttons for 'refresh' and 'Help Desk', and a list of options: 'start', 'filter', and 'output', each with a 'Not yet initialised' status.

## Filter

This stage allows the user to limit the initial search to a subset satisfying particular criteria. A wide range of filter types can be applied, in any combination. The system supports batch querying and a set of external identifiers can be uploaded directly from a file. The region filter allows a search to be carried out on the full genome, on a single chromosome, or on a portion of a chromosome (as determined by markers, bands or base pair coordinates). The availability of other filter options depends on the data content for a particular species and focus. For gene foci, multi-species filters can limit the selection of genes to those associated with homologues in other species, or with an upstream region that is conserved between species. Further filters allow restriction to a particular gene type or to genes that have been mapped to a particular external id set (for example, Affymetrix, EMBL, Gene Ontology or ZFIN identifiers). Searches can also be limited to genes with protein products possessing particular features, such as the presence of a transmembrane domain, signal sequence, or other domain specified using identifiers from domain databases. Access to expression data stored in EnsMart is provided via the eVOC controlled expression vocabulary. Currently two datasets can be accessed in this way: the GNF microarray dataset and EST-derived expression data. Finally, one can restrict searches to genes with SNPs in particular regions (for example, coding or UTR), or to genes that have non-synonymous SNPs.

For example, the following configuration of filters selects genes that satisfy the following criteria:

- placed in chromosome 20
- have at least two transcripts
- have identified orthologous genes in Fugu
- contain a transmembrane domain

The screenshot shows the bioMart interface with the following filter settings:

- REGION:**
  - Chromosome: 20
  - Base pair: Start, End
  - Marker: Start, End
- GENE:**
  - Known genes: Only/Excluded
  - Entries with following IDs (Ensembl Transcript ID(s))
  - Transcript count >= 2
  - Entries with a 5' UTR: Only/Excluded
  - Entries with a 3' UTR: Only/Excluded
  - Biotype data: biotype (protein\_coding), source (ensembl), confidence (KNOWN)
- MULTI SPECIES COMPARISONS:**
  - Homologous Fugu Genes: Only/Excluded
- PROTEIN:**
  - with PROFILE ID(s): Only/Excluded
  - Entries with following ID(s) (Interpro ID(s))
  - Transmembrane domains: Only/Excluded
  - Signal domains: Only/Excluded

Annotations on the left side of the screenshot:

- Select chromosome** (points to Chromosome: 20)
- two or more transcripts** (points to Transcript count >= 2)
- and have transmembrane domains** (points to Transmembrane domains)
- homologous to Fugu genes** (points to Homologous Fugu Genes)

Summary on the right side of the screenshot:

- start:** Dataset: Danio rerio genes(Ensembl), 20024 Entries Total
- filter:** Chromosome name: 20, Transcript count >= 2, Homologous Fugu Genes: Only, Transmembrane domains: Only, 57 Entries pass Filters
- output:** Feature List, 57 Results in Output

For each filter a MartView user can define whether the criteria should be satisfied or not. Click next to advance to the next stage.

### Output

In this stage we can select the format for the output, but first it might be of interest to check how many genes passed our criteria. We can find this information on the right-hand side of the page. In the example above there are 57 entries that have passed the filter.

For the output we require the following data:

- chromosome name (in this case, all should be on chromosome 20) and chromosome start
- Ensembl id for the gene
- ZFIN id if available

The screenshot shows the bioMart interface for Zebrafish. The 'Chromosome name' callout points to the 'Chromosome Name' attribute under 'Chromosome Attributes'. The 'Chromosome start' callout points to the 'Start Position (bp)' attribute. The 'Ensembl id' callout points to the 'Ensembl Gene ID' attribute under 'Gene Attributes'. The 'ZFIN id' callout points to the 'Zfin ID' attribute under 'External References (max 3)'. The right sidebar shows a filter for 'Chromosome name: 20' and 'Transcript count >=: 2'.

- Fugu homologues (which all genes should have since it was specified in the filter)

The screenshot shows the 'Fugu Homolog Attributes' section with the following attributes selected: 'Fugu Ensembl Gene ID', 'Fugu External ID', 'Fugu External DB', 'Fugu Chromosome', 'Fugu Chr Start (bp)', 'Fugu Chr End (bp)', 'Ensembl Peptide ID', '% Coverage', '% Identity', 'Fugu Ensembl Peptide ID', 'Fugu % Coverage', and 'Fugu % Identity'.

- and finally the output format is HTML

The screenshot shows the 'Select the output format' section with the following options: 'HTML' (selected), 'Text, comma separated', 'MS Excel', 'Text, fixed width', 'Text, tab separated', and 'Predefined ADF attributes'.

In order to get the output, click on the Export button. The output for the 57 genes appears in a table with links to the Ensembl database. Click on one of these genes and verify that all the selected criteria have been satisfied.

### **Exercises**

1. Try your own queries. Experiment with different filters and outputs. In particular try with the “sequence” option for output. You can export cDNA, genomic sequences and so on.
2. Dump all predicted coding regions that contain a tubulin domain. How would you approach this query?
3. In the filter stage an extra dataset can be added. In which situation is it useful to query two datasets?