Module 3 - Genes and Sequences
v. Data Mining using BioMart

Aims

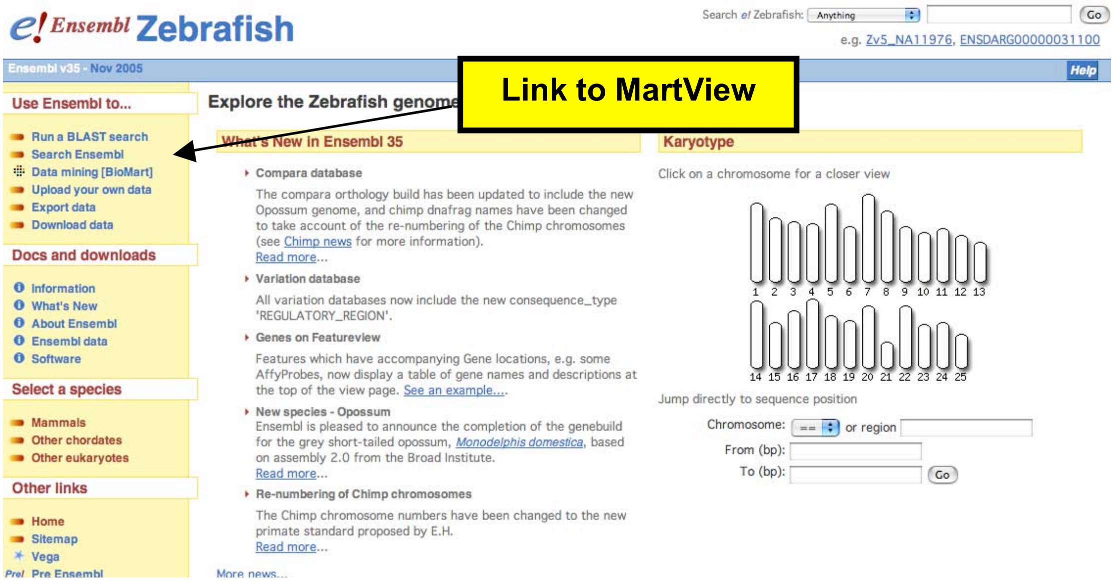
- Introduce BioMart, a data mining system for large datasets

Introduction

The BioMart 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. Follow the link to BioMart from the left-hand menu toolbar:



Compara database

The compara orthology build has been updated to include the new Opossum genome, and chimp dnafrag names have been changed to take account of the re-numbering of the Chimp chromosomes (see Chimp news for more information).
Read more...

Variation database

All variation databases now include the new consequence_type 'REGULATORY_REGION'.

Genes on Featureview

Features which have accompanying Gene locations, e.g. some AffyProbes, now display a table of gene names and descriptions at the top of the view page. See an example....

New species - Opossum

Ensembl is pleased to announce the completion of the genebuild for the grey short-tailed opossum, *Monodelphis domestica*, based on assembly 2.0 from the Broad Institute.
Read more...

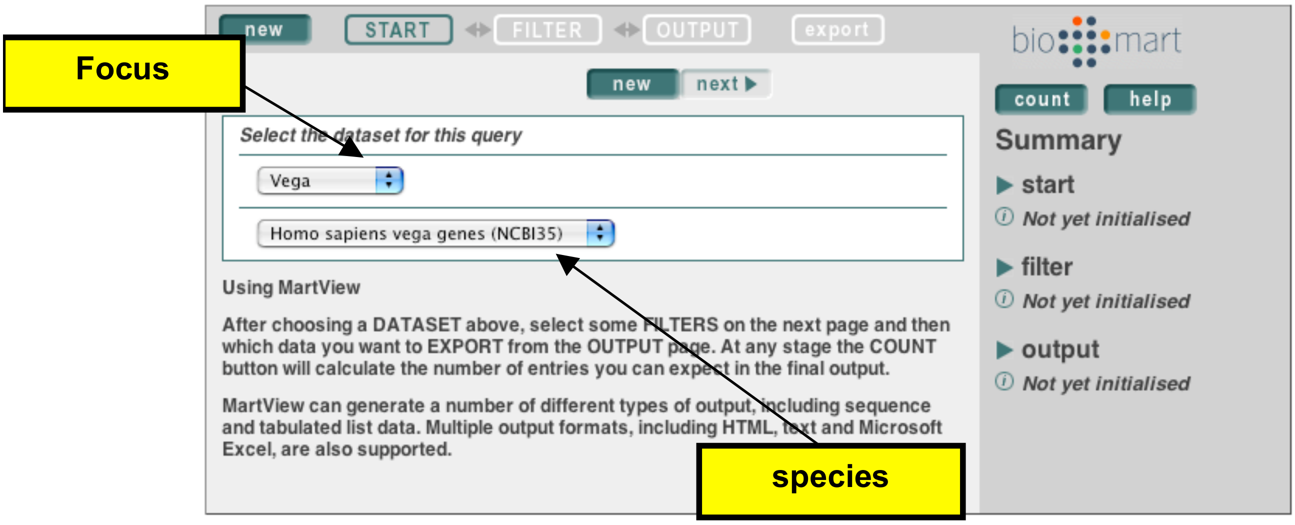
Re-numbering of Chimp chromosomes

The Chimp chromosome numbers have been changed to the new primate standard proposed by E.H.
Read more...

Queries in BioMart 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 the dataset for Ensembl and *Danio rerio* in the species box and click ‘next’.



Select the dataset for this query

Vega

Homo sapiens vega genes (NCBI35)

Using MartView

After choosing a DATASET above, select some FILTERS on the next page and then which data you want to EXPORT from the OUTPUT page. At any stage the COUNT button will calculate the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.

Summary

▶ start
Not yet initialised

▶ filter
Not yet initialised

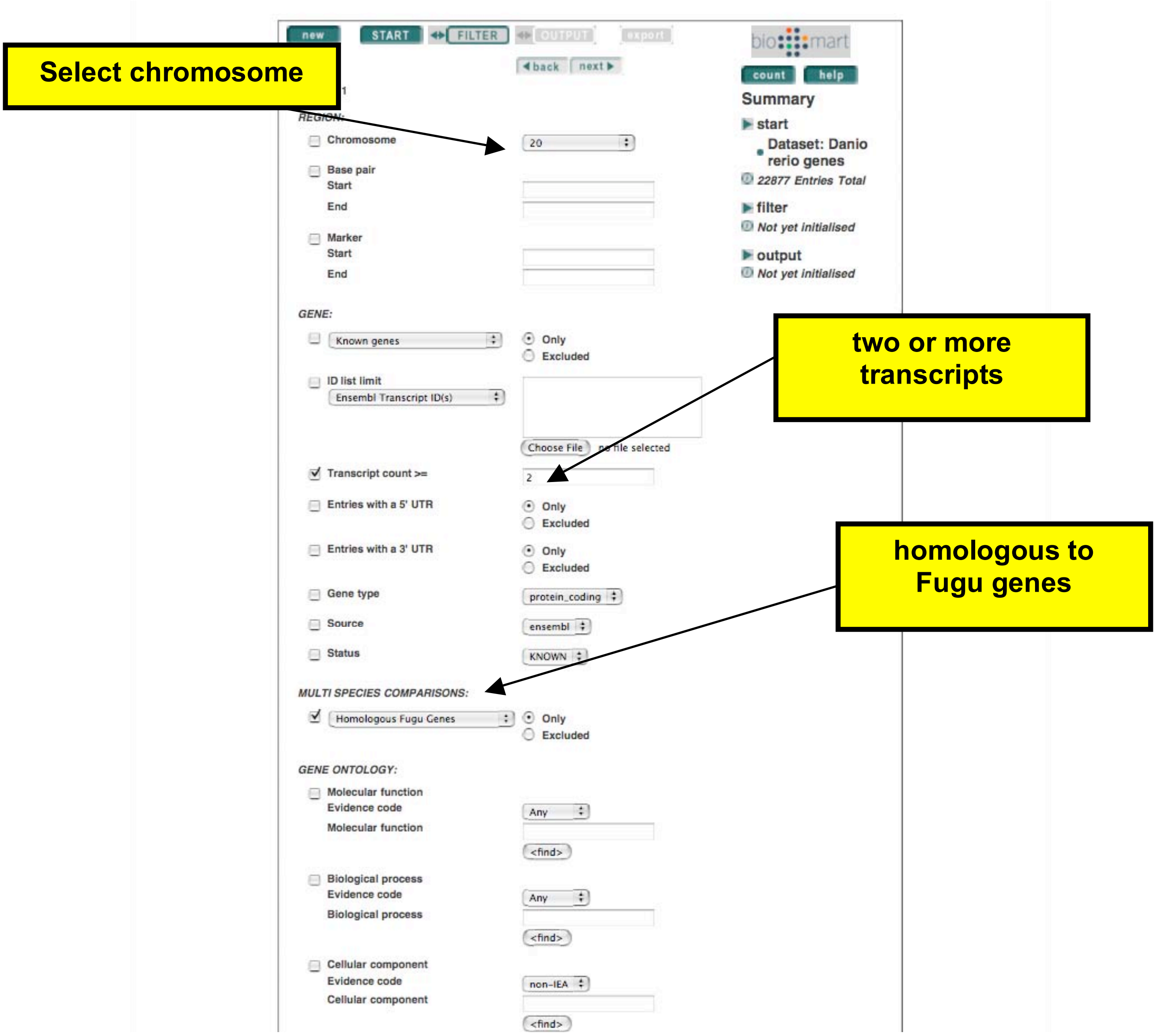
▶ output
Not yet initialised

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 BioMart 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



Select chromosome

two or more transcripts

homologous to Fugu genes

Summary

start
Dataset: Danio rerio genes
22877 Entries Total

filter
Not yet initialised

output
Not yet initialised

REGION:

□ Chromosome 20

□ Base pair
Start
End

□ Marker
Start
End

GENE:

□ Known genes ● Only ○ Excluded

□ ID list limit Ensembl Transcript ID(s)
no file selected

☑ Transcript count >= 2

□ Entries with a 5' UTR ● Only ○ Excluded

□ Entries with a 3' UTR ● Only ○ Excluded

□ Gene type protein_coding

□ Source ensembl

□ Status KNOWN

MULTI SPECIES COMPARISONS:

☑ Homologous Fugu Genes ● Only ○ Excluded

GENE ONTOLOGY:

□ Molecular function
Evidence code Any
Molecular function <find>

□ Biological process
Evidence code Any
Biological process <find>

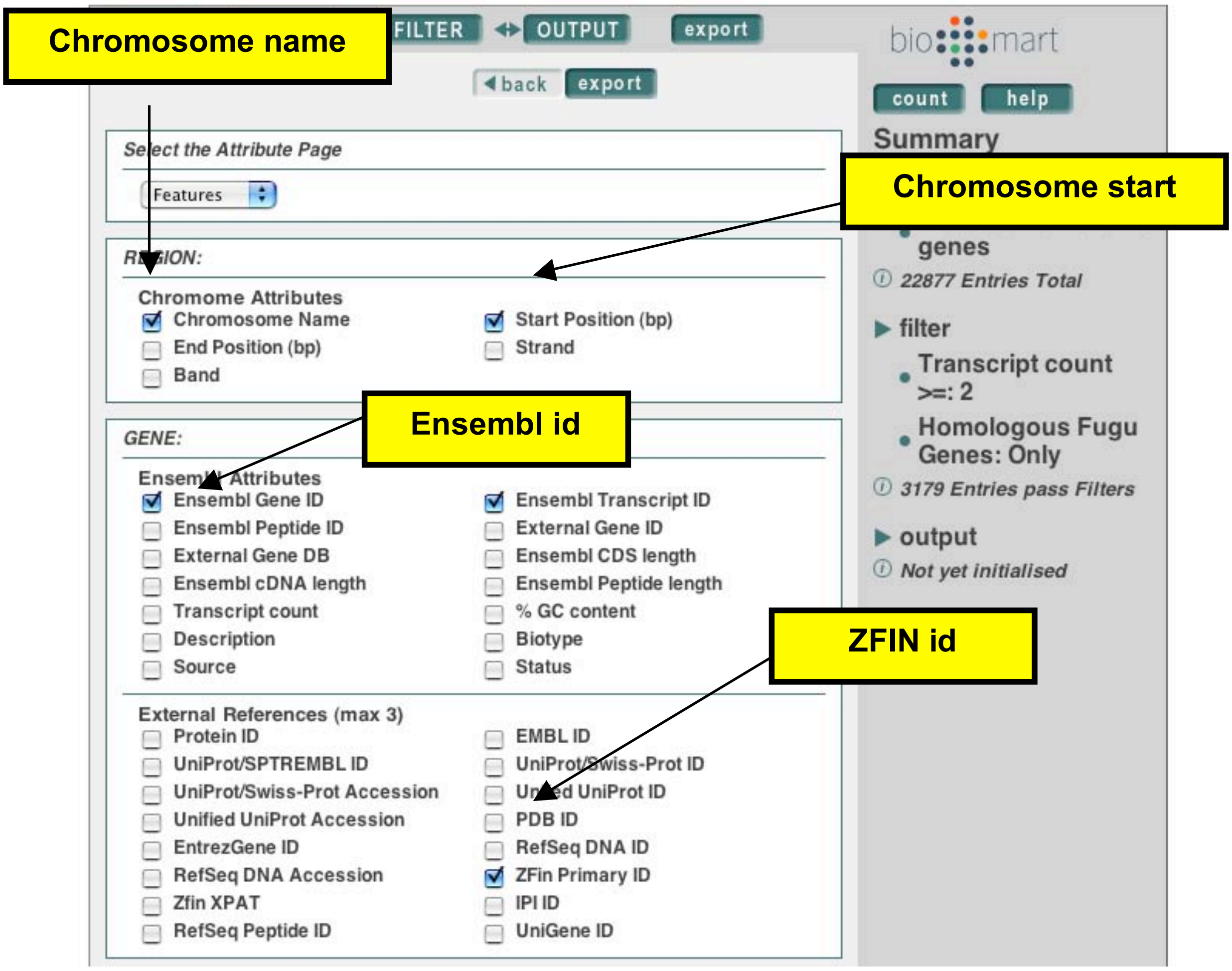
□ Cellular component
Evidence code non-IEA
Cellular component <find>

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. For the output we can for example 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



Select the Attribute Page

Features

REGION:

Chromome Attributes

☑ Chromosome Name
☐ End Position (bp)
☐ Band
☑ Start Position (bp)
☐ Strand

GENE:

Ensembl Attributes

☑ Ensembl Gene ID
☐ Ensembl Peptide ID
☐ External Gene DB
☐ Ensembl cDNA length
☐ Transcript count
☐ Description
☐ Source
☑ Ensembl Transcript ID
☐ External Gene ID
☐ Ensembl CDS length
☐ Ensembl Peptide length
☐ % GC content
☐ Biotype
☐ Status

External References (max 3)

☐ Protein ID
☐ UniProt/SPTREMBL ID
☐ UniProt/Swiss-Prot Accession
☐ Unified UniProt Accession
☐ EntrezGene ID
☐ RefSeq DNA Accession
☐ Zfin XPAT
☐ RefSeq Peptide ID
☐ EMBL ID
☐ UniProt/Swiss-Prot ID
☐ Unified UniProt ID
☐ PDB ID
☐ RefSeq DNA ID
☑ ZFin Primary ID
☐ IPI ID
☐ UniGene ID

Summary

- genes

22877 Entries Total

▶ filter

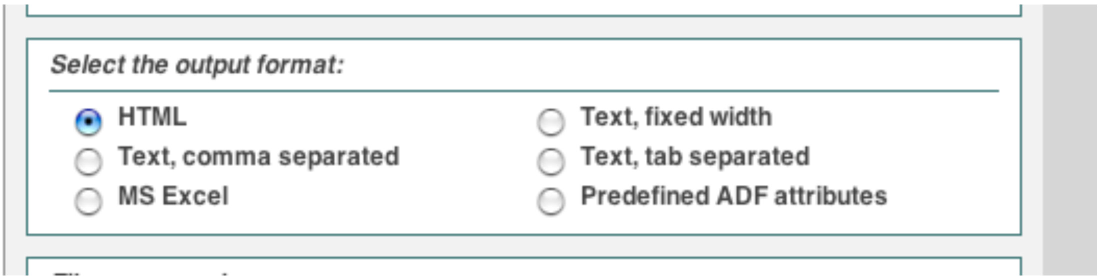
- Transcript count >=: 2
- Homologous Fugu Genes: Only

3179 Entries pass Filters

▶ output

Not yet initialised

- and finally the output format is HTML



Select the output format:

● HTML
○ Text, comma separated
○ MS Excel
○ Text, fixed width
○ Text, tab separated
○ Predefined ADF attributes

In order to get the output, click on the Export button. The output for the 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?

4. In the Filter step the user can specify terms from the ZFIN control vocabularies for anatomical and developmental stages data. For example get a list of all the genes known to be expressed in the blastula developmental stage. Click on ‘find’ to get the full list of terms.