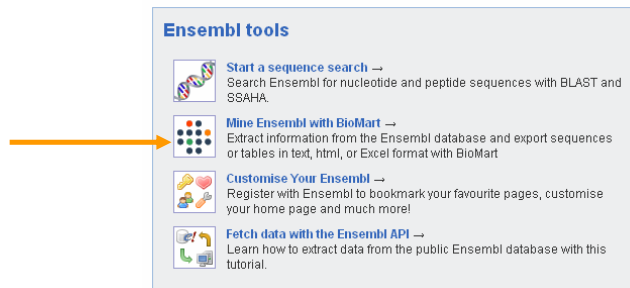


Data Mining in Ensembl with BioMart



Ensembl tools

- Start a sequence search** → Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.
- Mine Ensembl with BioMart** → Extract information from the Ensembl database and export sequences or tables in text, html, or Excel format with BioMart
- Customise Your Ensembl** → Register with Ensembl to bookmark your favourite pages, customise your home page and much more!
- Fetch data with the Ensembl API** → Learn how to extract data from the public Ensembl database with this tutorial.

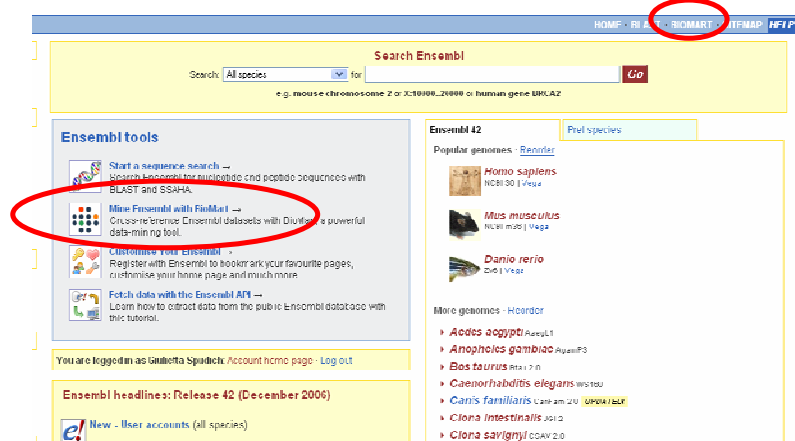
*Dr. Giulietta Spudich
Biosapiens 2007*

BioMart

<http://www.biomart.org/biomart/martview>

<http://www.ensembl.org/biomart/martview>

- Or click on 'BioMart' from Ensembl



HOME | RELEASE | **BIO MART** | RTMAP | HELP

Search Ensembl

Search: All species for Go

e.g. mouse chromosome 2 or X:100000_200000 or human gene BRCA2

Ensembl tools

- Start a sequence search → Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.
- Mine Ensembl with BioMart** → Extract information from the Ensembl database with BioMart's powerful data-mining tool.
- Customise Your Ensembl → Register with Ensembl to bookmark your favourite pages, customise your home page and much more!
- Fetch data with the Ensembl API → Learn how to extract data from the public Ensembl database with this tutorial.

You are logged in as Giulietta Spudich. Account home page - Log out

Ensembl headlines: Release 42 (December 2006)

New - User accounts (all species)

Ensembl 42 [Feed species](#)

Popular genomes - [Reorder](#)

- Homo sapiens** NCBI 30 | View
- Mus musculus** NCBI mouse | View
- Danio rerio** ZFIN | View

More genomes - [Reorder](#)

- Aedes aegypti** RefSeq | View
- Anopheles gambiae** AgParP3
- Bos taurus** RefSeq | View
- Caenorhabditis elegans** WGS | View
- Canis familiaris** UCSC | View
- Ciona intestinalis** JGI | View
- Ciona savignyi** UCSC | View

BioMart- Data mining

- BioMart filters the data in the Ensembl databases, combines multiple terms and puts them into a table format.
- Such as: human genes (HGNC IDs), chromosome and base pair position
- No programming required!

3 of 17

General or Specific Data-Tables

- All the genes for one species
- Or... only genes on one specific region of a chromosome
- Or... only genes on one specific region of a chromosome that have homologues

4 of 17

Information Flow

- **Dataset** (*species, genes or SNPs*)
- Decide on a smaller geneset using **Filters**.
(*enter IDs, choose a region ...*)
- Attach information to your gene set
(**Attributes**)
(*sequences, IDs, description...*)

5 of 17

BioMart

Datasets

- Ensembl genes
- Vega genes
- SNPs

Filters

- IDs, chromosomal region
- Gene Ontology (GO terms)
- Microarray probe IDs
- Protein domains (InterPro)

6 of 17

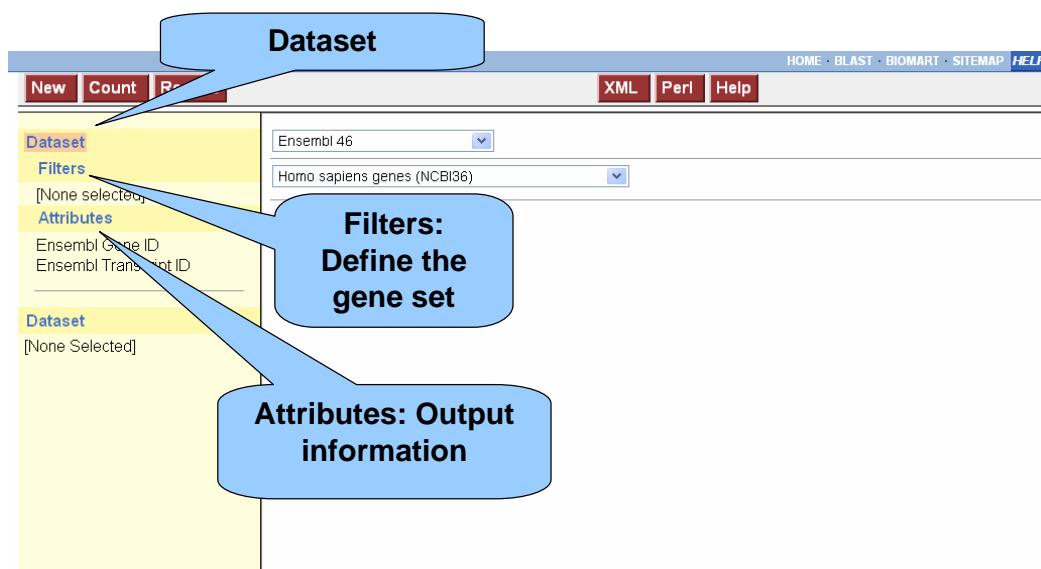
BioMart

Attributes (Output)

- *IDs, chromosomal region*
- *Gene Ontology (GO terms)*
- *Microarray probe IDs*
- *Protein domains (InterPro)*
- *Sequences (gene, peptide, UTR, flanking)*
- *Variations (SNPs) and position in gene*
- *Homologues*

7 of 17

Web Interface



Three main stages: Dataset, Filters and Attributes.

8 of 17

Results

The screenshot shows a web interface for viewing and exporting search results. At the top, there are navigation tabs: 'New', 'Count', 'Results', 'XML', 'Perl', and 'Help'. Below this, there are two main panels. The left panel is a sidebar with 'Dataset' and 'Attributes' sections. The right panel contains a table of results. The table has columns for 'Dataset', 'Filters', and 'Attributes'. The first row shows a dataset with the following attributes: Peptide, Ensembl Gene ID, Chromosome, and Biotype. Below the table, there is a section for 'Export all results to' with a dropdown menu set to 'FASTA' and a 'Go' button. There is also a 'View' section with a dropdown set to '10 rows as' and another 'Go' button. The main content area displays a list of protein sequences in FASTA format, starting with >ENS00000092377|Y|protein_coding and >ENS00000099715|Y|protein_coding.

Tables or sequences

9 of 17

Export tables as...

- Microsoft Excel (xls)
- Text (csv, tsv)
- HTML
- GFF
- XML

Or export sequences in FASTA format

FASTA sequences

- Gene (unspliced)
- Transcript (cDNA)
- Translation (coding)
- UTR (5' or 3')
- Flanking sequence

11 of 17

The Flow

- Choose Dataset (genes or SNPs, species)
- Choose Filters (narrows the gene set)
- Choose Attributes (output options)

12 of 17

Attributes and Filters

- For all human genes on chromosome 10 that are protein coding, I would like to know the IDs in Ensembl.
- In the query:
Filters: what we know
Attributes: what we want to know.

13 of 17

Query:

- For all **human genes** on **chromosome 10** that are **protein coding**, I would like to know the IDs in Ensembl.
- In the query:
Filters: what we know
Attributes: what we want to know.

14 of 17

Query:

- For all human genes on chromosome 10 that are protein coding, I would like to know the **IDs** in **Ensembl**.
- In the query:
Filters: what we know
Attributes: what we want to know.

15 of 17

BioMart – Other Installations

The screenshot displays the BioMart interface for the MCW Proteomics Center. The top navigation bar includes 'GRAMENE Martview MartView' and a search box. Below this, the 'Find in' section shows 'Browse' and 'Genomes' options. The main content area is titled 'PRIDE' and features a 'New' button, 'XML', 'Help', 'Count', and 'Results' buttons. The 'Dataset' section is expanded, showing 'PRIDE' as the selected dataset, with 'Attributes' and 'Filters' both set to '[None selected]'. The 'Database' dropdown is set to 'PRIDE BioMart' and the 'Dataset' dropdown is set to 'PRIDE'. A 'Using MartView' section provides instructions: 1. Choose **Dataset** above, 2. Click **Attributes** and make your selection in the list below, 3. Click **Results** in the top panel. A link to a 'Mini Tutorial' is also present.

Find more at www.biomart.org

16 of 17

BioMart team

- [Arek Kasprzyk](#)
- Benoît Ballester
- Syed Haider
- Richard Holland
- Damian Smedley

