

Data Mining with BioMart



The screenshot shows the Ensembl BioMart web interface. At the top, the Ensembl logo is on the left, and navigation links for 'Home', 'Login / Register', 'BLAST/BLAT', and 'BioMart' are on the right. The 'BioMart' link is circled in black. Below the navigation bar is a toolbar with buttons for 'New', 'Count', 'Results', 'URL', 'XML', 'Perl', and 'Help'. The main content area is divided into two sections. On the left, under the heading 'Dataset', it says '[None selected]'. On the right, there is a dropdown menu labeled '- CHOOSE DATABASE -' with a downward arrow.

www.ensembl.org/biomart/martview
www.biomart.org/biomart/martview

What is BioMart?

- search engine that can find multiple terms and put them into a table format.
- A web interface to mine Ensembl data
- No programming required

General or Specific Data-Tables

- All the genes for one species
- Or... only genes on one specific region of a chromosome
- Or... make BioMart select genes
(I.e. all transcripts that match a microarray probe set, GO term, or InterPro domain).

Results

Dataset 1 / 49506 Genes

Homo sapiens genes (GRCh37)

Filters

HGNC symbol: [ID-list specified]

Attributes

Ensembl Gene ID

Ensembl Transcript ID

Description

Associated Gene Name

MIM Morbidity

Dataset

[None Selected]

Export all results to

File

TSV

☐ Unique results only

Go

Email notification to

View

10

rows as

HTML

☐ Unique results only

New

Count

Results

URL

XML

Perl

Help

Dataset 1 / 49506 Genes

Homo sapiens genes (GRCh37)

Filters

HGNC symbol: [ID-list specified]

Attributes

Protein

Ensembl Gene ID

Ensembl Transcript ID

Associated Gene Name

Dataset

[None Selected]

Export all results to

File

FASTA

☐ Unique results only

Go

Email notification to

View

10

rows as

FASTA

☐ Unique results only

```

>ENSG00000001626|ENST00000446805|CFTR
MFYGIFFLYLGEVTKAVQPLLGR I IASYDPDNKEERS
MQRSPLEKASVSKLFFSWIRPILRKGYRQRLSDIYQIPSVDSADNLSEKLEREWDR
LASKKNPKLINALRRCFFWRFMFYGIFFLYLGEVTKAVQPLLGR I IASYDPDNKEERSIA
IYLGIGLCLLFIVRTLLHPIAIFGLHHIGMQMRIA MFSLIYKKGLALAHFVWIAPLQVAL
LMGLIWELLQASAFQGLGFLIVLALFQAGLG RMMMKYRDQRAKISERLVITSEMIENIQ
SVKAYCWEEAMEKMIENLRQTELKLRKAAYVRYFNSSAFFFSGFFVFLSVLPYALIKG
IILRKIFTTISFCIVLRMAVTRQFPWAVQTWYDSLGAINKIQDFLQKQEKYKLTLEYNLTT
EVMENVTAFWEEGFELFEKAKQNNNNRKT SNGDDSLFFSNFSLGTPVLKDI NFKIER
GQLLAVAGSTGAGKTSLLMVIMGELEPSE GKIKHSGRISFCSQFSWIMPGTIKENI IFGV
SYDEYRYSVIAKQLEEDISKFAEKDNIVLGEGGITLSGGQRARISLARAVYKDADLYL
LDSPPFGYLDVLTEKEIFESCVC KLMANKTRILVTSKMEHLKKADKILILHEGSSYFYGT
SELQNLQPDFSSKLMGCD SFDQFSAERNRSILTETLHRSLEGDAPVSWTETKKQSFQQT
GEFGEKRKNSILNPINSIRKFSIVQKTP LQMNGIEEDSDEPLERRLSLPDSEQGEAILP
RISVISTGPTLQARRRQSVNLMT HSVNQGN IHRKTTASTRKVSLAPQANLTELD IYSR
RLSQETGLEISEEINEEDLKECFDDMESIPAVTTWNTYLR YITVHKSLIFVL IWCLVIF
LAEVAASLVVLWLLGNIPLODKGNSTHSRNN SYAVIITSTSSYYVFYIYVGVA DTL LAMG

```

Tables or sequences

The First Step: Choose the Dataset

The screenshot shows a web application interface for dataset selection. The top navigation bar includes buttons for 'New', 'Count', 'Results', 'URL', 'XML', 'Perl', and 'Help'. The left sidebar contains three sections: 'Dataset' (highlighted), 'Filters' (showing '[None selected]'), and 'Attributes' (listing 'Ensembl Gene ID' and 'Ensembl Transcript ID'). The main content area features two dropdown menus. The first dropdown menu is set to 'Ensembl Genes 59', and the second dropdown menu is set to 'Homo sapiens genes (GRCh37)'.

Dataset: Current Ensembl, Human genes

The Second Step: Filters

New Count Results	★ URL XML Perl Help
Dataset Homo sapiens genes (GRCh37)	Please restrict your query using criteria below
Filters [None selected]	
Attributes Ensembl Gene ID Ensembl Transcript ID	
Dataset [None Selected]	
	☐ REGION:
	☐ GENE:
	☐ TRANSCRIPT EVENT:
	☐ GENE ONTOLOGY:
	☐ EXPRESSION:
	☐ MULTI SPECIES COMPARISONS:
	☐ PROTEIN DOMAINS:
	☐ VARIATIONS:

Filters: Define a gene set

Attributes attach information

New Count Results	★ URL 📄 XML
Dataset Homo sapiens genes (GRCh37)	Please select columns to be included
Filters [None selected]	☒ Features ☐ Transcript Event
Attributes (circled)	☐ Structures ☐ Homologs
Ensembl Gene ID	☐ Variations ☐ Sequences
Ensembl Transcript ID	☐ GENE:
	☐ EXTERNAL:
	☐ EXPRESSION:
	☐ PROTEIN DOMAINS:
Dataset [None Selected]	

Attributes: Determine output columns

Query

For the human CFTR gene, export the Entrez Gene ID(s) and matching Affy HG U133-PLUS-2 probeset(s)

Query:

For the **human CFTR gene**, export the Entrez Gene ID(s) and matching Affy HG U133-PLUS-2 probeset(s)

- In the query:

Filters: what we know

Attributes: what we want to know.

Query:

For the human CFTR gene, export the **Entrez Gene ID(s)** and matching **Affy HG U133-PLUS-2** probeset(s)

- In the query:

Filters: what we know

Attributes: what we want to know

A Brief Example

The screenshot shows a web interface with a top navigation bar containing buttons for 'New', 'Count', 'Results', 'URL', 'XML', 'Perl', and 'Help'. On the left, a sidebar has a 'Dataset' section with '[None selected]'. The main area features a dropdown menu currently displaying 'Ensembl Genes 59'. A blue callout bubble with a pointer to the dropdown arrow contains the text: 'Use the current Ensembl (archives are also available)'.

This screenshot shows the same web interface, but the dropdown menu now displays 'Homo sapiens genes (GRCh37)'. The left sidebar is expanded, showing sections for 'Dataset', 'Filters', and 'Attributes'. Under 'Attributes', 'Ensembl Gene ID' and 'Ensembl Transcript ID' are listed. A blue callout bubble with a pointer to the selected option in the dropdown contains the text: 'Select *Homo sapiens genes*'.

Select the Genes with Filters

Click Filters

New Count Results URL XML Perl Help

Please restrict your query using criteria below

Dataset
Homo sapiens genes (GRCh37)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

☒ REGION:
☒ **GENE:**
☒ GENE ONTOLOGY:
☒ EXPRESSION:
☒ MULTI SPECIES COMPARISONS:
☒ PROTEIN DOMAINS:

Expand the 'GENE' panel.

Expand the GENE panel to enter in the gene ID(s).

Filters (and Count)

New Count Results

URL XML Perl Help

Dataset
Homo sapiens genes (GRCh37)

Filters
HGNC curated gene name [e.g. SRY]: [ID-list specified]

Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset
[None Selected]

Please restrict your query using criteria below

REGION:

GENE:

☐ Limit to genes ...

☒ Only
☐ Excluded

with Illumina HumanWG 6 v1 ID(s)

☒ ID list limit

HGNC curated gene name [e.g. SRY]
CFTR

Change this to HGNC
curated name. Enter
“CFTR” in the box.

Click “Count” to see if genes passed through
your filters.

Attributes (Output Options)

The screenshot shows a web interface with a top navigation bar containing buttons for 'New', 'Count', 'Results', 'URL', 'XML', 'Perl', and 'Help'. On the left, a sidebar lists 'Dataset 1 / 51737 Genes', 'Filters', and 'Attributes' (which is selected). Below 'Attributes', there are options for 'Ensembl Gene ID' and 'Ensembl Transcript ID'. The main content area has a heading 'Please select columns to be included in the output and hit 'Results' when ready'. It contains several radio button options: 'Features' (selected), 'Structures', 'Transcript Event', 'Homologs', 'Variations', and 'Sequences'. Below these are expandable sections: 'GENE:', 'EXTERNAL:' (circled), 'EXPRESSION:', and 'PROTEIN DOMAINS:'.

Click on 'Attributes'

'Attributes' allows you to output information.

Attributes (Output Options)

Select 'EntrezGene ID'

External References (max 3)

- | | |
|---|--|
| ☐ UCSC ID | ☐ MIM Morbid Accession |
| ☐ PDB ID | ☐ MIM Morbid Description |
| ☐ Clone based Ensembl gene name | ☐ MIM Gene Accession |
| ☐ Clone based Ensembl transcript name | ☐ MIM Gene Description |
| ☐ Clone based VEGA gene name | ☐ miRBase Accession(s) |
| ☐ Clone based VEGA transcript name | ☐ miRBase ID(s) |
| ☐ CCDS ID | ☐ Protein (Genbank) ID |
| ☐ EMBL (Genbank) ID | ☐ RefSeq DNA ID |
| ☐ Ensembl LRG transcript | ☐ RefSeq Predicted DNA ID |
| ☒ EntrezGene ID | ☐ RefSeq Protein ID |
| ☐ VEGA transcript ID(s) (OTTT) | ☐ RefSeq Predicted Protein ID |
| ☐ VEGA gene ID(s) (OTTG) | ☐ RefSeq Genomic ID(s) |
| ☐ Ensembl transcript (where OTTT shares CDS with ENST) | ☐ Rfam ID |
| | ☐ Unigene ID |

Attributes (Output Options)

Select the Affy Platform
'HG U133-PLUS-2' in the
'Microarray' section

Microarray (max 2)

- ☐ Affy HC G110
- ☐ Affy HG FOCUS
- ☒ Affy HG U133-PLUS-2
- ☐ Affy HG U133A_2
- ☐ Affy HG U133A
- ☐ Affy HG U133B
- ☐ Affy HG U95AV2
- ☐ Affy HG U95B
- ☐ Affy HG U95C
- ☐ Affy HG U95D
- ☐ Affy HG U95E
- ☐ Affy HG U95A

- ☐ Affy HuGene FL
- ☐ Affy HuEx 1_0 st v2
- ☐ Affy HuGene 1_0 st v1
- ☐ Affy U133 X3P
- ☐ Agilent CGH 44b
- ☐ Agilent WholeGenome
- ☐ Codelink
- ☐ Illumina HumanWG 6 v1
- ☐ Illumina HumanWG 6 v2
- ☐ Illumina HumanWG 6 v3
- ☐ Illumina Human HT 12
- ☐ Phalanx OneArray

The Results Table - Preview

For the full result table: click “Go” or View “ALL” rows.

Dataset 1 / 51737 Genes
Homo sapiens genes (GRCh37)

Filters
HGNC curated gene name [e.g. SRY]: [ID-list specified]

Attributes
Ensembl Gene ID
Ensembl Transcript ID
EntrezGene ID
Affy HG U133-PLUS-2

Dataset
[None Selected]

Export: all results to File
[Go](#)

Email notification to

View 10 rows as HTML ☐ Unique results only

Ensembl Gene ID	Ensembl Transcript ID	EntrezGene ID	Affy HG U133-PLUS-2
ENSG00000001626	ENST00000446805		
ENSG00000001626	ENST00000003084	1080	215702_s_at
ENSG00000001626	ENST00000003084	1080	205043_at
ENSG00000001626	ENST00000426809		215702_s_at
ENSG00000001626	ENST00000468795		215702_s_at
ENSG00000001626	ENST00000468795		215703_at
ENSG00000001626	ENST00000454343		215702_s_at
ENSG00000001626	ENST00000454343		205043_at

Other Export Options (Attributes)

- ❖ Sequences: UTRs, flanking sequences, cDNA and peptides, etc
- ❖ Gene IDs from Ensembl and external sources (MGI, Entrez, etc)
- ❖ Microarray data
- ❖ Protein Functions/descriptions (Interpro, GO)
- ❖ Orthologous gene sets
- ❖ SNP/ Variation Data

How to Get There

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

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

- Or click on 'BioMart' from Ensembl

The screenshot displays the Ensembl BioMart interface. At the top, the Ensembl logo is on the left, and navigation links (Login / Register, BLAST/BLAT, BioMart, Docs & FAQs) are on the right. The 'BioMart' link is circled in red. Below the header, the location is set to 'Human' and 'Location: 6:133,017,695-133,161,157'. A sidebar on the left lists 'Location-based displays' including 'Whole genome', 'Chromosome summary', 'Region overview', 'Region in detail' (selected), 'Comparative Genomics', 'Genomic alignments', 'Synteny (10)', 'Genetic Variation', 'Resequencing (6)', 'Linkage Data', and 'Markers'. Below these are links to 'Configure this page', 'Add/manage custom data', 'Export data', and 'Bookmark this page'. The main content area shows 'Chromosome 6: 133,017,695-133,161,157'. It includes a chromosome ideogram with a red box highlighting the region of interest, and a detailed view of the region. The detailed view shows a genomic track with various features: 'Chromosome bands', 'Contigs' (AL357034.18.1.127703, AL513524.8.1.141029, AL032821.2.1.143563), 'Ensembl/Havana gene' (RP1-248E1.2, MOXD1, STX7, RP11-295F4.4, RP11-295F4.1, TAAR8, TAAR6, TAAR5, TAAR2, TAAR1, VNN1, VNN3, C6orf192, RP5-1181K21.2, AL139102.13, RP11-203B4.1, RP1-55C23.4, RP1-55C23.5, VNN2, RPS12), and 'ncRNA gene' (SNORD100, SNORD100, SNORA33). A scale bar at the top of the detailed view indicates a 1.00 Mb range from 132.80 Mb to 133.80 Mb. The 'Forward strand' is indicated on the right.

Central Portal



Powered by BioMart software:

- [BioMart Central Portal](#)
- [Ensembl](#)
- [Ensembl Bacteria](#)
- [Ensembl Metazoa](#)
- [Ensembl Protists](#)
- [Dictybase](#)
- [Wormbase](#)
- [Gramene](#)
- [Euromphenome](#)
- [UniProt](#)
- [InterPro](#)
- [HGNC](#)
- [Rat Genome Database](#)
- [DroSpeGe](#)
- [ArrayExpress DW](#)
- [Eurexpress](#)
- [HapMap](#)
- [GermOnLine](#)
- [PRIDE](#)
- [PepSeeker](#)
- [VectorBase](#)
- [HTGT](#)
- [Pancreatic Expression Database](#)
- [Reactome](#)
- [EU Rat Mart](#)
- [Paramecium DB](#)
- [International Potato Center \(CIP\)](#)

Third party software with BioMart Plugin:

[Bioclipse](#) [biomaRt-BioConductor](#) [Cytoscape](#) [Galaxy](#) [Taverna](#) [WebLab](#) [Ruby API](#)

www.biomart.org

Worked Example

This module should do the following:

- Show you how to export multiple data types from Ensembl for gene IDs or chromosomal regions.

Ensembl Core Databases

Relational Database

- Normalised
- Each data point stored only once

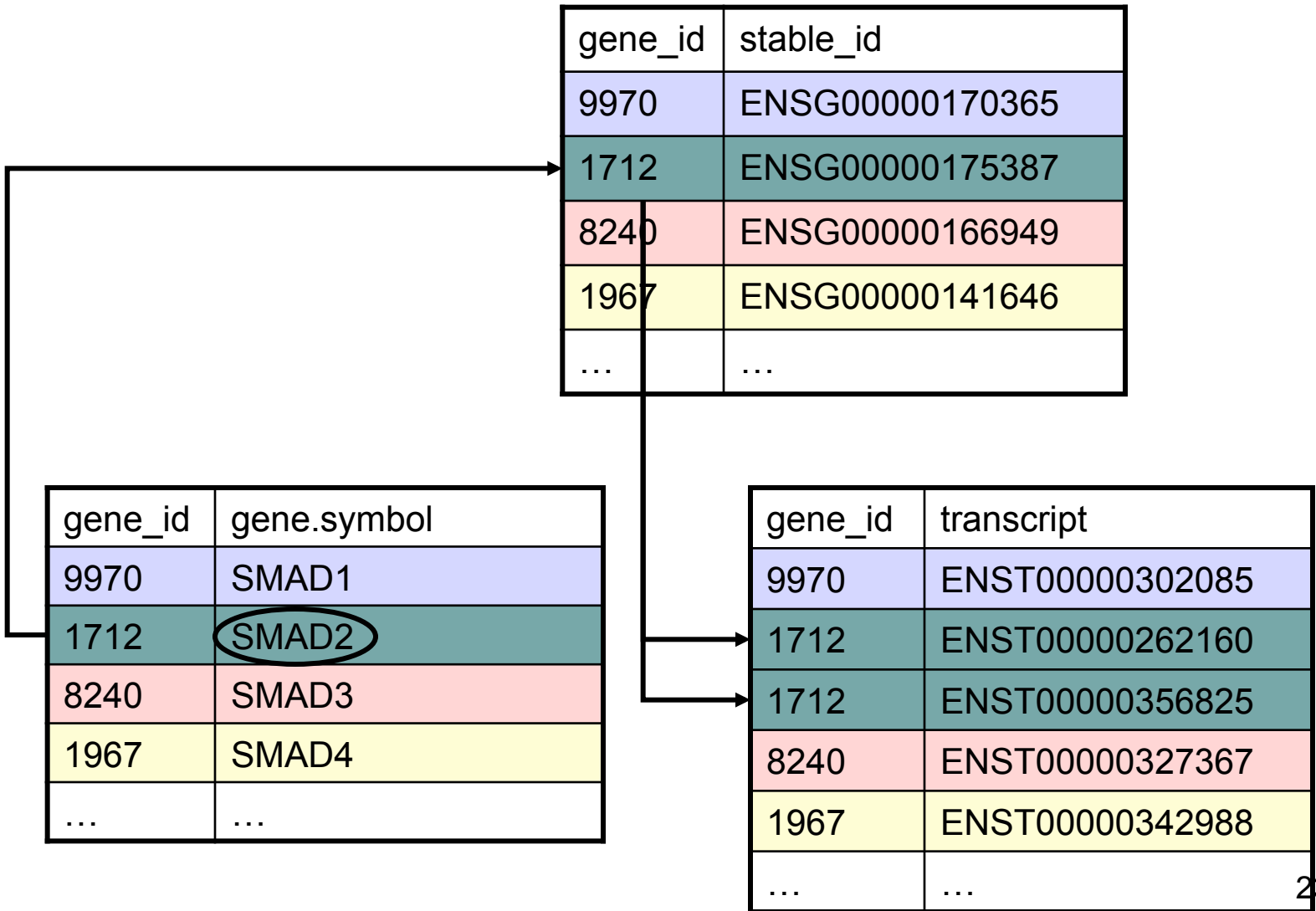
Therefore:

- Quick updates
- Minimal storage requirements

But:

- Many tables
- Many joins for complicated queries
- Slow for data mining applications

Normalised Schema



BioMart Database

Data warehouse

- De-normalised
- Query-optimised

Therefore:

- Fast and flexible
- Ideal for data mining

But:

- Tables with apparent “redundancy”
- Needs rebuilding from scratch for every release from normalised core databases

De-Normalised Schema

gene_id	transcript_id	gene.symbol
ENSG00000170365	ENST00000302085	SMAD1
ENSG00000175387	ENST00000262160	SMAD2
ENSG00000175387	ENST00000356825	SMAD2
ENSG00000166949	ENST00000327367	SMAD3
ENSG00000141646	ENST00000342988	SMAD4
...	...	...

Information Flow

DATASET

SPECIES
FOCUS

FILTER

REGION
GENE
EXPRESSION
HOMOLOGY
PROTEIN
SNP
SWISSPROT
EMBL
REFSEQ
GO
INTERPRO
AFFYMETRIX

ATTRIBUTES

REGION
GENE
EXPRESSION
HOMOLOGY
PROTEIN
SNP
FASTA
GTF
HTML
TEXT
EXCEL
FILE