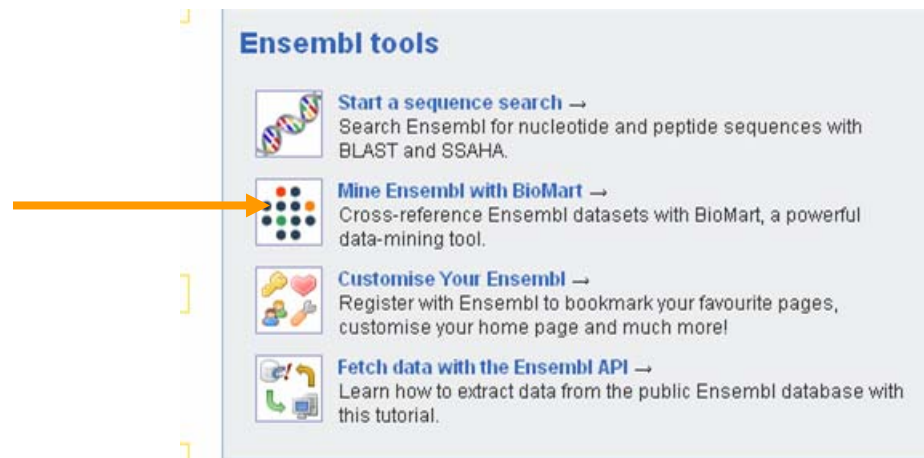






Data Mining in Ensembl with BioMart



Ensembl tools

-  **Start a sequence search** → Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.
-  **Mine Ensembl with BioMart** → Cross-reference Ensembl datasets with BioMart, a 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.

BioMart- Data mining

- BioMart is a search engine that can find multiple terms and put them into a table format.
- Such as: mouse gene (IDs), chromosome and base pair position
- 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... genes on one region of a chromosome associated with a disease

BioMart Data Sets

- Ensembl genes
- Vega genes
- SNPs

- Markers
- “Diseases”
- Gene expression information
- Gene ontology
- Homology predictions
- Protein annotation

Web Interface

New Help Count Results

» **Dataset:**
Homo sapiens genes (NCBI36)

» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID

» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Database:

Dataset:

Choose a **Dataset** above, then use the left panel to navigate through the **Attributes** and **Filters** making your selections in this main panel. To preview the results click the **Results** button in the top panel.

[Mini Tutorial](#)

With BioMart, quickly extract gene-associated information from the Ensembl databases.

Information Flow

- Choose the species of interest (**Dataset**)
- Decide what you would like to know about the genes (**Attributes**)
(sequences, IDs, description...)
- Decide on a smaller geneset using **Filters**.
(enter IDs, choose a region ...)

Web Interface

The screenshot shows a web interface with a top navigation bar containing 'New' and 'Help' buttons. The main content area is divided into two panels. The left panel is highlighted in yellow and contains a tree view with the following structure:

- » Dataset:
 - Homo sapiens genes (NCBI36)
- » Attributes (Features)
 - Ensembl Gene ID
 - Ensembl Transcript ID
- » Filters
 - [None selected]

The right panel contains a 'Database:' dropdown menu set to 'Ensembl 43' and a text input field containing 'Homo sapiens genes (NCBI36)'. Below these, there is a paragraph of text: 'aset above, then use the left panel to navigate through the **Attributes** and **Filters** selections in this main panel. To preview the results click the **Results** button in the top panel.'

Three blue callout boxes with white text provide instructions:

- 'Choose the species of interest' points to the 'Homo sapiens genes (NCBI36)' item in the Dataset list.
- 'Choose what information to view.' points to the 'Attributes (Features)' section.
- 'Choose the geneset using what we know.' points to the text input field.

Three main stages: Dataset, Attributes and Filters.

The First Step: Choose the Dataset

New **Help** **Count** **Results**

» **Dataset:**
Homo sapiens genes (NCBI36)
» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID
» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Database:

Dataset:

Choose a **Dataset** above, then use the left panel to navigate through the **Attributes** and **Filters** making your selections in this main panel. To preview the results click the **Results** button in the top panel.

[Mini Tutorial](#)

Homo sapiens genes are the default.

The Second Step: Attributes

Four output pages.

» **Dataset:**
Homo sapiens genes (NCBI36)

» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID

» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Features Homologs
 Structures Sequences
 SNPs

REGION:

Chromosome Attributes

Chromosome Name Strand
 Gene Start (bp) Band
 Gene End (bp)

GENOMIC FEATURES:

GENE:

EXPRESSION:

PROTEIN:

Attributes are what we want to know about the genes.

The SNP Attribute Page

» **Dataset:**
Homo sapiens genes (NCBI36)

» **Attributes** (SNPs)
Ensembl Gene ID
Ensembl Transcript ID

» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Features Homologs
 Structures Sequences
 SNPs

REGION:

GENE:

GENE ASSOCIATED SNPS:

SNP Attributes

<input type="checkbox"/> Reference ID	<input type="checkbox"/> Validation status
<input type="checkbox"/> Allele	<input type="checkbox"/> Mapweight
<input type="checkbox"/> TSC ID	<input type="checkbox"/> fpctg name
<input type="checkbox"/> HGBASE ID	

SNP Location Attributes

<input type="checkbox"/> Transcript location (bp)	<input type="checkbox"/> Peptide location (aa)
<input type="checkbox"/> SNP Chromosome Strand	<input type="checkbox"/> Chromosome Location (bp)

Gene Location and Effect

<input type="checkbox"/> Location in Gene (coding etc)	<input type="checkbox"/> Synonymous Status
<input type="checkbox"/> Peptide Shift	

Output variation information such as SNP reference ID and alleles.

Filters Allow Gene Selection

» **Dataset:**

Homo sapiens genes (NCBI36)

» **Attributes** (SNPs)

Ensembl Gene ID

Ensembl Transcript ID

» **Filters**

[None selected]

» **Dataset:**

[None Selected]

REGION:

GENOMIC FEATURES:

GENE:

GENE ONTOLOGY:

EXPRESSION:

MULTI SPECIES COMPARISONS:

PROTEIN:

SNP:

**Choose the geneset by region, gene ID(s),
protein/domain type.**

Export Sequence or Tables

» Dataset:
Homo sapiens genes (NCBI36)

» Attributes (Sequences)

- Peptide
- Chromosome
- Ensembl Gene ID
- Biotype

» Filters
[None selected]

» Dataset:
[None Selected]

Display maximum rows as

Export all results to

```
>Y|ENSG00000168757|protein_coding
MRPEGSLTYRVPERLRQSGCGVGRAAQALVCASAKEGTAFRMEAVQEGAAGVESEQAALG
EEAVLLDDIMAEVEVVAEEEGLVERRRQAQAAVPGPGPMTPEALEELLAVQVELE
PVNAQARKAFSRQREKMERRRKPFLDERRGAVIQSVPGFWANVIANHPQMSALITDEDEDM
LSYMSLEVEEEKHRVHLCKIMLFFRSNPYFQNKVITKEYLVNITEYRASHSTPIEWYPD
YEVEAYRRRHNSLNFNWFSDHNFAGSNKIAESPDRSVVRTCGAIPCNTTRG*
>Y|ENSG00000099715|protein_coding
MTVGFNSDISSVVRVNTTCHKCLLSGTYIFAVLLVCVVFHSGAQEKNYTIREEIPENVL
IGNLLKDLNLSLIPNKSLTTTMOFKLVYKTDGVDPLIRIEEDTGEIFTTGARIDREKLCAG
IPRDEHCFYEVEVAILPDEIFRLVKIRFLIEDINDNAPLFPATVINISIPENSAINSKYT
LPAAVDPDVGINGVQNYELIKSQNIFGLDVIETPEGDKMPQLIVQKELDREEKDTYVMKV
KVED
ENAK
MVLVI
TDHE
NDNAI
GMLT
PENLI
TFYVI
AVDNI
```

Ensembl Gene ID	Ensembl Transcript ID	Description
ENSG00000206668	ENST00000383941	Y RNA [Source:RFAM;Acc:RF00019]
ENSG00000206789	ENST00000384062	5S ribosomal RNA [Source:RFAM;Acc:RF00001]
ENSG00000207376	ENST00000384646	7SK RNA [Source:RFAM;Acc:RF00100]
ENSG00000195216	ENST00000352275	Small nucleolar RNA U70 [Source:RFAM;Acc:RF00156]
ENSG00000194751	ENST00000353862	Small subunit ribosomal RNA, 5' domain [Source:RFAM;Acc:RF00001]
ENSG00000206750	ENST00000384023	Y RNA [Source:RFAM;Acc:RF00019]
ENSG00000207372	ENST00000384642	U2 spliceosomal RNA [Source:RFAM;Acc:RF00004]
ENSG00000206986	ENST00000384259	U6 spliceosomal RNA [Source:RFAM;Acc:RF00026]
ENSG00000206894	ENST00000384167	5S ribosomal RNA [Source:RFAM;Acc:RF00001]
ENSG00000206860	ENST00000384133	Eukaryotic type signal recognition particle RNA [Source:RFAM;Acc:RF00001]

Genes and attributes are exported as sequence (Fasta format) or tables.

Query:

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

Query:

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

Query:

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

A Brief Example

The image shows a screenshot of a web interface for dataset selection. On the left is a yellow sidebar with navigation options. The main panel on the right contains dropdown menus for 'Database' and 'Dataset', a text box for instructions, and a link for a 'Mini Tutorial'. A blue callout bubble points to the 'Dataset' dropdown, containing the text 'Change dataset to mouse Mus musculus'.

» Dataset:
Homo sapiens genes (NCBI36)

» Attributes (Features)
Ensembl Gene ID
Ensembl Transcript ID

» Filters
[None selected]

» Dataset:
[None Selected]

Database:

Dataset:

Choose a **Dataset** above, then use the left panel to navigate through the **Attributes** and **Filters** making your selections in this main panel. To preview the results click the **Results** button in the top panel.

[Mini Tutorial](#)

**Change dataset to mouse
*Mus musculus***

A Brief Example

» **Dataset:**
Mus musculus genes (NCBIM36)

» **Attributes (Features)**
Ensembl Gene ID
Ensembl Transcript ID

» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Database:

Dataset:

Choose a **Dataset** above, then use the left panel to navigate through the **Attributes** and **Filters** making your selections in this main panel. To preview the results click the **Results** button in the top panel.

[Mini Tutorial](#)

Dataset has changed.

Attributes (Output Options)

» **Dataset:**
Mus musculus gen...

» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID

» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Features Homologs
 Structures Sequences
 SNPs

REGION:
 GENOMIC:
 GENE:
 PROTEIN:

Attributes allow us to choose what we wish to know.

IDs are found in the 'Features' page.

Attributes (Output Options)

» **Dataset:**
Mus musculus genes (NCBIM36)
» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID
» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Features **Homologs**
 Structures **Sequences**
 SNPs

REGION:
 GENOMIC FEATURES:
 GENE:

Ensembl Attributes

<input checked="" type="checkbox"/> Ensembl Gene ID	<input type="checkbox"/> Ensembl Peptide length
<input checked="" type="checkbox"/> Ensembl Transcript ID	<input type="checkbox"/> Transcript count
<input type="checkbox"/> Ensembl Peptide ID	<input type="checkbox"/> % GC content
<input type="checkbox"/> External Gene ID	<input type="checkbox"/> Description
<input type="checkbox"/> External Gene DB	<input type="checkbox"/> Biotype
<input type="checkbox"/> Ensembl CDS length	<input type="checkbox"/> Source
<input type="checkbox"/> Ensembl cDNA length	<input type="checkbox"/> Status

GO Attributes

<input type="checkbox"/> GO ID	<input type="checkbox"/> GO evidence code
<input type="checkbox"/> GO description	

Ensembl Gene ID is selected

**Default options selected:
Ensembl Gene ID and Transcript ID**

Attributes (Output Options)

» Dataset:
Mus musculus genes (NCBIM36)
» Attributes (Features)
Ensembl Gene ID
Ensembl Transcript ID
Markersymbol ID
» Filters
[None selected]

» Dataset:
[None Selected]

GO Attributes

- GO ID
- GO description
- GO evidence code

External References (max 3)

- CCDS ID
- Codelink ID
- EMBL ID
- EntrezGene ID
- Havana ID
- Illumina ID
- IPI ID
- Markersymbol ID
- Markersymbol Accession
- Mirbase
- PDB ID
- Protein ID
- RefSeq DNA ID
- RefSeq Predicted DNA ID
- UniProt/Swiss-Prot Accession
- Unified UniProt ID
- Unified UniProt Accession
- Uniprot varsplic ID
- Illumina v1
- Imgt gene db
- Imgt ligm db

Microarray Attributes (max 2)

- Agilent Probe
- Affy moe430a
- Affy moe430b

'Markersymbol ID' will give us the MGI ID

↓

**Scroll down to select MGI symbol.
Also select the accession number.**

The Results Table

» **Dataset:**

Mus musculus genes (NCBIM36)

» **Attributes** (Features)

Ensembl Gene ID
Ensembl Transcript ID
Markersymbol ID
Markersymbol Accession

» **Filters**

[None selected]

» **Dataset:**

[None Selected]

Display maximum

10



rows as

HTML



Export all results to

File



Go

Ensembl Gene ID	Ensembl Transcript ID	Markersymbol ID	Markersymbol Accession
ENSMUSG00000071964	ENSMUST00000096694		
ENSMUSG00000053211	ENSMUST00000065545	Zfy2	MGI:99213
ENSMUSG00000053211	ENSMUST00000065545	Zfy1	MGI:99212
ENSMUSG00000068457	ENSMUST00000089879		
ENSMUSG00000068457	ENSMUST00000069309	Uty	MGI:894810
ENSMUSG00000068457	ENSMUST00000044500	Uty	MGI:894810
ENSMUSG00000069053	ENSMUST00000091208		
ENSMUSG00000056673	ENSMUST00000055032	Jarid1d	MGI:99780
ENSMUSG00000069049	ENSMUST00000091197	Eif2s3y	MGI:1349430
ENSMUSG00000069049	ENSMUST00000091194		

‘Results’ give us Gene IDs for all mouse genes in the Ensembl database.

Select a Smaller Geneset

» **Dataset:**
Mus musculus genes (NCBIM36)

» **Attributes** (Features)

- Ensembl Gene
- Ensembl Trans
- Markersymbol
- Markersymbol

» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Select 'Filters'

Expand the REGION panel

- REGION:
- GENOMIC FEATURES:
- GENE:
- GENE ONTOLOGY:
- MULTI SPECIES COMPARISONS:
- PROTEIN:
- SNP:

Instead of all mouse genes, select protein coding genes on chromosome 10.

Select Genes on Chromosome 10

» **Dataset:**
Mus musculus genes (NCBIM36)

» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID
Markersymbol ID
Markersymbol Accession

» **Filters**
Chromosome: 10

» **Dataset:**
[None Selected]

REGION:

Chromosome 10

Base pair
Start 1
End 10000000

Band
Start
End

Marker
Start
End

GENOMIC FEATURES:

GENE:

Select chromosome 10

Instead of all mouse genes, select protein coding genes on chromosome 10.

Select Protein Coding Genes

» **Dataset:**
Mus musculus genes (NCBIM36)

» **Attributes** (Features)
Ensembl Gene ID
Ensembl Transcript ID
Markersymbol ID
Markersymbol Accession

» **Filters**
Chromosome: 10
Gene type : protein_coding

» **Dataset:**
[None Selected]

Entries with a 5' UTR Only Excluded

Entries with a 3' UTR Only Excluded

Gene type
Mt_tRNA
protein_coding
pseudogene
rRNA
snoRNA

Source

Status

GENE ONTOLOGY:

MULTI SPECIES COMPARISONS:

PROTEIN:

SNP:

Gene type: protein coding

Filters are set to chromosome 10 and protein-coding genes. Genes must meet BOTH criteria to be in the result table.

Results (Preview)

For the full result
table: Go

» **Dataset:**

Mus musculus genes (NCBIM36)

» **Attributes** (Features)

Ensembl Gene ID
Ensembl Transcript ID
Markersymbol ID
Markersymbol Accession

» **Filters**

Chromosome: 10

» **Dataset:**

[None Selected]

Display maximum

10

rows as

HTML

Export all results to

File

Go

Ensembl Gene ID	Ensembl Transcript ID	Markersymbol ID	Markersymbol Accession
ENSMUSG000000015202	ENSMUST000000015346	Cnksr3	MGI : 2674130
ENSMUSG000000064065	ENSMUST000000086896	A130090K04Rik	MGI : 2444159
ENSMUSG000000064065	ENSMUST000000058132	A130090K04Rik	MGI : 2444159
ENSMUSG000000064065	ENSMUST000000078070	A130090K04Rik	MGI : 2444159
ENSMUSG00000000766	ENSMUST000000063036	Oprm1	MGI : 97441
ENSMUSG00000000766	ENSMUST000000052751	Oprm1	MGI : 97441
ENSMUSG00000000766	ENSMUST000000092731	Oprm1	MGI : 97441
ENSMUSG00000000766	ENSMUST000000056385	Oprm1	MGI : 97441
ENSMUSG00000000766	ENSMUST000000092729	Oprm1	MGI : 97441
ENSMUSG00000000766	ENSMUST000000000783	Oprm1	MGI : 97441

This is a preview- if you are happy with the
table, click 'Go'.

Full Result Table

Ensembl Gene ID

Transcript ID

MGI symbol

MGI Accession Number

Ensembl Gene ID	Ensembl Transcript ID	Markersymbol ID	Markersymbol Accession
ENSMUSG00000015202	ENSMUST00000015346	Cnksr3	MGI:2674130
ENSMUSG00000064065	ENSMUST00000086896	A130090K04Rik	MGI:2444159
ENSMUSG00000064065	ENSMUST00000058132	A130090K04Rik	MGI:2444159
ENSMUSG00000064065	ENSMUST00000078070	A130090K04Rik	MGI:2444159
ENSMUSG00000000766	ENSMUST00000063036	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000052751	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000092731	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000056385	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000092729	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000000783	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000078634	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST000000100088	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000092728	Oprm1	MGI:97441
ENSMUSG00000000766	ENSMUST00000092734	Oprm1	MGI:97441
ENSMUSG00000043134	ENSMUST00000055501		
ENSMUSG00000058273	ENSMUST00000073045		
ENSMUSG00000019775	ENSMUST00000019909	Rgs17	MGI:1927469
ENSMUSG00000019775	ENSMUST00000064225	Rgs17	MGI:1927469
ENSMUSG00000019774	ENSMUST00000019908	Mtrf11	MGI:1918830
ENSMUSG00000019773	ENSMUST00000019907	Fbxo5	MGI:1914391
ENSMUSG00000019772	ENSMUST00000019906	Vip	MGI:98933
ENSMUSG00000046916	ENSMUST00000051809	Myct1	MGI:1915882
ENSMUSG00000046916	ENSMUST00000091210	Myct1	MGI:1915882
ENSMUSG00000019769	ENSMUST00000041639	Syne1	MGI:1927152
ENSMUSG00000019769	ENSMUST00000056571	Syne1	MGI:1927152

Original Query:

- For all mouse genes on chromosome 10 that are protein coding, I would like to know the **IDs** in both **Ensembl and MGI**.
- In the query:
Filters: what we know
Attributes: columns in the **Result Table**

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

- Either www.biomart.org/martview
- Or click on 'BioMart' from Ensembl

The screenshot shows the Ensembl website interface. At the top right, the navigation menu includes 'HOME', 'BLAST', 'BIOMART', 'ITEMAP', and 'HELP'. The 'BIOMART' link is circled in red. Below the navigation is a search bar titled 'Search Ensembl' with a dropdown menu set to 'All species' and a 'Go' button. Below the search bar is a section titled 'Ensembl tools' with four links: 'Start a sequence search', 'Mine Ensembl with BioMart' (circled in red), 'Customise your Ensembl', and 'Fetch data with the Ensembl API'. To the right of the 'Ensembl tools' section is a 'Popular genomes' section with links to 'Homo sapiens', 'Mus musculus', and 'Danio rerio'. Below that is a 'More genomes' section with links to 'Aedes aegypti', 'Anopheles gambiae', 'Bos taurus', 'Caenorhabditis elegans', 'Canis familiaris', 'Ciona intestinalis', and 'Ciona savignyi'. At the bottom left, there is a 'You are logged in as Giulietta Spudich' message and an 'Ensembl headlines' section for 'Release 42 (December 2006)'.

BioMart team

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

