Data Mining in Ensembl with BioMart
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 an InterPro domain
The First Step: Choose the Dataset

- Dataset Options:
  - Ensembl Genes (release 49)
  - Mus musculus genes (NCBI/37)

- Filters:
  - [None selected]

- Attributes:
  - Ensembl Gene ID
  - Ensembl Transcript ID

- Dataset:
  - [None Selected]
The Second Step: Filters

Filters define which genes we are looking at.
**Attributes attach information**

Determine output columns with Attributes.
### Results

#### Dataset Filters
- [None selected]

#### Attributes
- Peptide
- Ensembl Gene ID
- Chromosome
- Biotype

---

#### Tables or sequences

**Dataset**

- ENS0000000052377

**Attributes**

- Peptide: MTRENEAISISIVKVNTHNCNLCHCSPOTIFYFAVILCAYVFHSAQGRNQRTYREIPERVL
- Ensembl Gene ID: [protein_coding]
- Chromosome: 12
- Biotype: [protein_coding]

---

**Dataset**

- ENS0000000099715

**Attributes**

- Peptide: MVQHNEDISSVYKNTHBCLCNLCSITUHFALCAYVFHSAQGRNQRTYREIPERVL
- Ensembl Gene ID: [protein_coding]
- Chromosome: 12
- Biotype: [protein_coding]
Query:

• For all mouse genes on chromosome 10 that are protein coding, I would like to know the IDs in both Ensembl and MGI. Are there Illumina probes and GO IDs for these genes?

• 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. Are there Illumina probes and GO IDs for these genes?

• 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. Are there Illumina probes and GO IDs for these genes?

• In the query:
  Filters: what we know
  Attributes: what we want to know.
A Brief Example

Change dataset to mouse *Mus musculus*
Select the genes with Filters

- Click Filters.
- Expand the ‘REGION’ panel.

We are looking for mouse genes on chromosome 10 that are protein coding.
Filters (selecting the genes)

Please restrict your query using criteria below

- REGION:
  - Chromosome
  - Base pair
    - Gene Start (bp)
    - Gene End (bp)
- Band
  - Start
  - End

Change this to chromosome 10
Filters (selecting the genes)

Select 'protein coding' in the 'GENE' section.

Click on 'Attributes'
We would like GO terms and IDs in MGI (the Mouse Genome Informatics site).
Attributes (Output)

Click ‘Results’

Scroll down to add ‘Illumina v1’ probes that map to these genes.
The Results Table - Preview

For the full result table: click ‘Go’ or View ‘ALL’ rows.

‘Results’ shows Gene IDs, GO terms, and Illumina probes for all protein coding mouse genes on chromosome 10.
<table>
<thead>
<tr>
<th>Ensembl Gene ID</th>
<th>Ensembl Transcript ID</th>
<th>GO ID</th>
<th>GO description</th>
<th>MGI symbol</th>
<th>Illumina probes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015346</td>
<td>GO:0005515</td>
<td>protein binding</td>
<td>C6kscr3</td>
<td>a0d83236.13.428_30-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015346</td>
<td>GO:0005737</td>
<td>cytoplasm</td>
<td>C6kscr3</td>
<td>a0d83236.13.428_30-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015346</td>
<td>GO:0005966</td>
<td>regulation of signal transduction</td>
<td>C6kscr3</td>
<td>a0d83236.13.428_30-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015621</td>
<td>GO:0005515</td>
<td>protein binding</td>
<td>C6kscr3</td>
<td>a0d83236.13.428_30-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015621</td>
<td>GO:0005737</td>
<td>cytoplasm</td>
<td>C6kscr3</td>
<td>a0d83236.13.428_30-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015621</td>
<td>GO:0005966</td>
<td>regulation of signal transduction</td>
<td>C6kscr3</td>
<td>a0d83236.13.428_30-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015621</td>
<td>GO:0005617</td>
<td>membrane</td>
<td>A130090K04Rik</td>
<td>a10d83236.13.428 2.215-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015617</td>
<td>GO:0005617</td>
<td>membrane</td>
<td>A130090K04Rik</td>
<td>a10d83236.13.428 2.215-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015617</td>
<td>GO:0005617</td>
<td>membrane</td>
<td>A130090K04Rik</td>
<td>a10d83236.13.428 2.215-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015617</td>
<td>GO:0005617</td>
<td>membrane</td>
<td>A130090K04Rik</td>
<td>a10d83236.13.428 2.215-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015617</td>
<td>GO:0005617</td>
<td>membrane</td>
<td>A130090K04Rik</td>
<td>a10d83236.13.428 2.215-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015618</td>
<td>GO:0005617</td>
<td>membrane</td>
<td>A130090K04Rik</td>
<td>a10d83236.13.428 2.215-S</td>
</tr>
<tr>
<td>ENSMUSG00000015202</td>
<td>ENSMUST00000015618</td>
<td>GO:0005617</td>
<td>membrane</td>
<td>A130090K04Rik</td>
<td>a10d83236.13.428 2.215-S</td>
</tr>
<tr>
<td>ENSMUSG00000064065</td>
<td>ENSMUST000000086586</td>
<td>GO:0005518</td>
<td>rhodopsin-like receptor activity</td>
<td>Oprm1</td>
<td>a10d839174.18.263-S</td>
</tr>
<tr>
<td>ENSMUSG00000064065</td>
<td>ENSMUST000000058132</td>
<td>GO:0005518</td>
<td>rhodopsin-like receptor activity</td>
<td>Oprm1</td>
<td>a10d839174.18.263-S</td>
</tr>
<tr>
<td>ENSMUSG00000064065</td>
<td>ENSMUST000000058132</td>
<td>GO:0005518</td>
<td>rhodopsin-like receptor activity</td>
<td>Oprm1</td>
<td>a10d839174.18.263-S</td>
</tr>
<tr>
<td>ENSMUSG00000064065</td>
<td>ENSMUST0000000105615</td>
<td>GO:0005518</td>
<td>rhodopsin-like receptor activity</td>
<td>Oprm1</td>
<td>a10d839174.18.263-S</td>
</tr>
<tr>
<td>ENSMUSG00000064065</td>
<td>ENSMUST0000000105615</td>
<td>GO:0005518</td>
<td>rhodopsin-like receptor activity</td>
<td>Oprm1</td>
<td>a10d839174.18.263-S</td>
</tr>
<tr>
<td>ENSMUSG00000064065</td>
<td>ENSMUST0000000105615</td>
<td>GO:0005518</td>
<td>rhodopsin-like receptor activity</td>
<td>Oprm1</td>
<td>a10d839174.18.263-S</td>
</tr>
<tr>
<td>ENSMUSG00000064065</td>
<td>ENSMUST0000000105615</td>
<td>GO:0005518</td>
<td>rhodopsin-like receptor activity</td>
<td>Oprm1</td>
<td>a10d839174.18.263-S</td>
</tr>
</tbody>
</table>
For all mouse genes on chromosome 10 that are protein coding, I would like to know the IDs in both Ensembl and MGI. Are there Illumina probes and GO IDs for these genes?

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
BioMart Data Sets

- Ensembl genes
- Vega genes
- Variations
BioMart around the world...

BioMart started at Ensembl...
To where has it travelled?
Central Portal

Powered by BioMart software:

- BioMart Central Portal
- Ensembl
- HapMap
- HTGT
- Dicybase
- Wormbase
- Gramene
- Europhenome
- Rat Genome Database
- ProSpeGe
- ArrayExpress DW
- Eurexpress
- GermOnLine
- PRIDE
- PepSeeker
- VectorBase
- Pancreatic Expression Database
- Reactome
- EU Rat Mart
- Paramecium DB

Third party software with BioMart Plugin:

- Bioclipse
- biomart-BioConductor
- Cytoscape
- Galaxy
- Taverna
- WebLab

www.biomart.org
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.

For a bookmarkable version of this page, click [here]
Population frequencies

Inter-population comparisons

Gene annotation
DictyBase
GRAMENE

www.gramene.org
How to Get There

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

• Or click on ‘BioMart’ from Ensembl