

## BIOMART

Use the web-based Biomart in Ensembl to create a dataset using the following parameters:

### Dataset:

Ensembl Genes 100 (or the latest version)

Mouse genes (GRCm38.p6) (or the latest version)

### Filters:

Chromosome 11

Band E2 only

Transcript count >=7

Limit to genes with RefSeq protein (peptide) IDs only

### Attributes:

Default attributes

Add "RefSeq Protein (peptide) ID"

I used the datasets Ensembl Genes 102 (from the archive site) and Mouse Genes (GRCm38.p6).

Screenshots of the results are shown below

The screenshot displays the Ensembl Biomart interface. The sidebar on the left shows the dataset selection process: "Dataset 64 / 56305 Genes", "Mouse genes (GRCm38.p6)", and "Filters" (Chromosome/scaffold: 11, Band Start: E2, Band End: E2, Transcript count >=: 7, With RefSeq peptide ID(s): Only). The "Attributes" section lists "Gene stable ID", "Gene stable ID version", "Transcript stable ID", "Transcript stable ID version", and "RefSeq peptide ID". The main area shows the export options: "Export all results to" (File), "Email notification to" (empty), "View" (10 rows as HTML), and "Unique results only" (checked). The "Go" button is visible. Below the screenshot, the exported data is shown as a text file in Notepad, with the following content:

```
mart_export.txt - Notepad
File Edit Format View Help
Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version, RefSeq peptide ID
ENSMUSG00000034566, ENSMUSG00000034566.10, ENSMUST00000043931, ENSMUST00000043931.8, NP_082138
ENSMUSG000000025137, ENSMUSG000000025137.15, ENSMUST00000026129, ENSMUST00000026129.15, NP_001349930
ENSMUSG000000025137, ENSMUSG000000025137.15, ENSMUST00000026129, ENSMUST00000026129.15, NP_077191
ENSMUSG000000025137, ENSMUSG000000025137.15, ENSMUST00000106188, ENSMUST00000106188.3, NP_001334544
ENSMUSG00000020770, ENSMUSG00000020770.13, ENSMUST00000021116, ENSMUST00000021116.11, NP_766157
ENSMUSG000000020770, ENSMUSG000000020770.13, ENSMUST00000106452, ENSMUST00000106452.1, NP_001272935
ENSMUSG000000025138, ENSMUSG000000025138.14, ENSMUST00000080202, ENSMUST00000080202.11, NP_001350368
ENSMUSG000000025138, ENSMUSG000000025138.14, ENSMUST00000080202, ENSMUST00000080202.11, NP_694696
ENSMUSG000000045775, ENSMUSG000000045775.15, ENSMUST00000106532, ENSMUST00000106532.3, NP_001346537
ENSMUSG000000045775, ENSMUSG000000045775.15, ENSMUST00000092445, ENSMUST00000092445.11, NP_001346535
ENSMUSG000000045775, ENSMUSG000000045775.15, ENSMUST00000092445, ENSMUST00000092445.11, NP_001074403
ENSMUSG000000057948, ENSMUSG000000057948.12, ENSMUST00000075036, ENSMUST00000075036.8, NP_001009573
ENSMUSG000000020773, ENSMUSG000000020773.11, ENSMUST00000106441, ENSMUST00000106441.7, NP_766158
ENSMUSG000000020773, ENSMUSG000000020773.11, ENSMUST00000021120, ENSMUST00000021120.5, NP_001192010
ENSMUSG000000020776, ENSMUSG000000020776.18, ENSMUST00000103031, ENSMUST00000103031.7, NP_001351008
ENSMUSG000000020776, ENSMUSG000000020776.18, ENSMUST00000103031, ENSMUST00000103031.7, NP_001351007
```

## BIOMART - R

Search OMIM.org for "huntington's disease". The first five entries all have this or a similar phrase in the title. Record the five identifiers (six-digit numbers) of those five records. The corresponding biomaRt filter name for these identifiers is "mim\_morbid\_accession". Use biomaRt to retrieve two tables with the following attributes, limiting to the five MIM values you found

### First table

Entrez Gene ID  
HGNC symbol  
Ensembl Gene ID

### Second table

HGNC symbol  
Ensembl Gene ID  
Ensembl Transcript ID

**The top 5 identifiers are: 603218, 604802, 143100, 606438, 607136**

**The table sizes are different because the second one is showing transcripts for each gene, some of which have multiple transcripts**

### R Script

```
library(biomaRt)
ensembl = useMart("ENSEMBL_MART_ENSEMBL", dataset="hsapiens_gene_ensembl",
                  host="www.ensembl.org")
getBM(attributes= c("entrezgene_id", "hgnc_symbol",
                    "ensembl_gene_id"), filters = "mim_morbid_accession",
        values=c("603218", "604802", "143100", "606438", "607136"), mart = ensembl)
getBM(attributes= c("hgnc_symbol", "ensembl_gene_id",
                    "ensembl_transcript_id"), filters = "mim_morbid_accession",
        values=c("603218", "604802", "143100", "606438", "607136"), mart = ensembl)
```

### Tables

```
> getBM(attributes= c("entrezgene_id", "hgnc_symbol",
+                    "ensembl_gene_id"), filters = "mim_morbid_accession",
+        values=c("603218", "604802", "143100", "606438", "607136"), mart = ensembl)
  entrezgene_id hgnc_symbol ensembl_gene_id
1          3064      HTT   ENSG00000197386
2          5621     PRNP   ENSG00000171867
3         57338     JPH3   ENSG00000154118
4          6908     TBP   ENSG00000112592
> getBM(attributes= c("hgnc_symbol", "ensembl_gene_id",
+                    "ensembl_transcript_id"), filters = "mim_morbid_accession",
+        values=c("603218", "604802", "143100", "606438", "607136"), mart = ensembl)
  hgnc_symbol ensembl_gene_id ensembl_transcript_id
1      HTT   ENSG00000197386   ENST00000680239
2      HTT   ENSG00000197386   ENST00000680956
3      HTT   ENSG00000197386   ENST00000680360
4      HTT   ENSG00000197386   ENST00000681528
5      HTT   ENSG00000197386   ENST00000647962
6      HTT   ENSG00000197386   ENST00000649900
7      HTT   ENSG00000197386   ENST00000680291
8      HTT   ENSG00000197386   ENST00000355072
9      HTT   ENSG00000197386   ENST00000648150
10     HTT   ENSG00000197386   ENST00000506137
11     HTT   ENSG00000197386   ENST00000512909
12     HTT   ENSG00000197386   ENST00000510626
```