Data Mining in Ensembl with BioMart

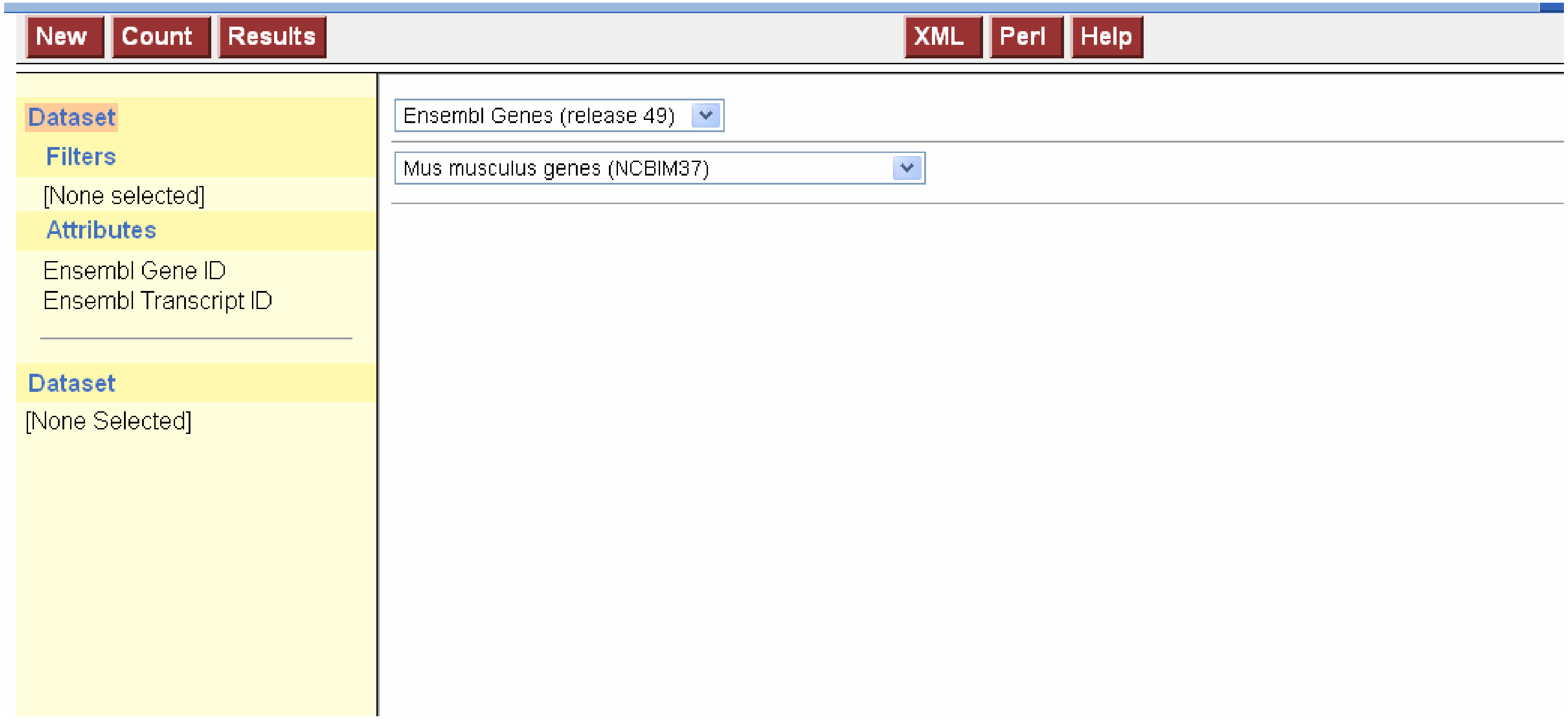
BioMart- Data mining

- 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



[None selected]

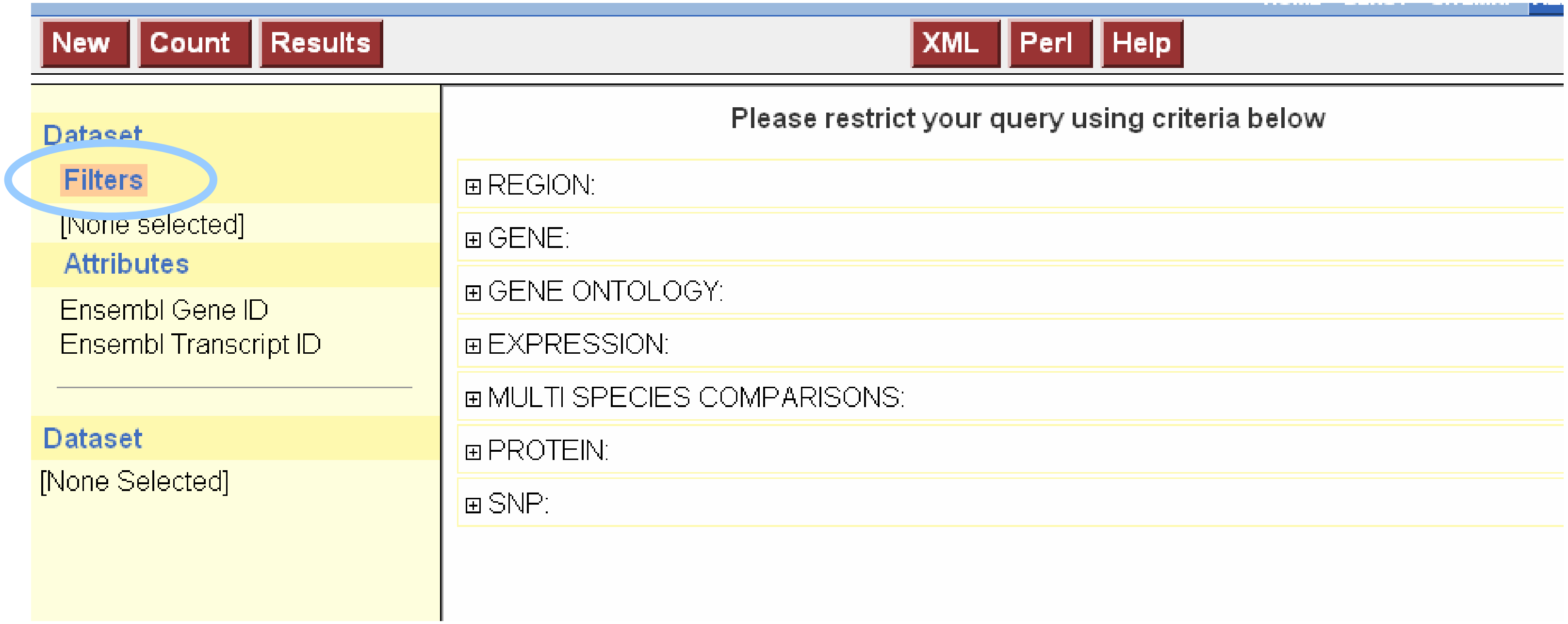
Ensembl Gene ID
Ensembl Transcript ID

[None Selected]

Ensembl Genes (release 49)

Mus musculus genes (NCBIM37)

The Second Step: Filters



Please restrict your query using criteria below

⊞ REGION:

⊞ GENE:

⊞ GENE ONTOLOGY:

⊞ EXPRESSION:

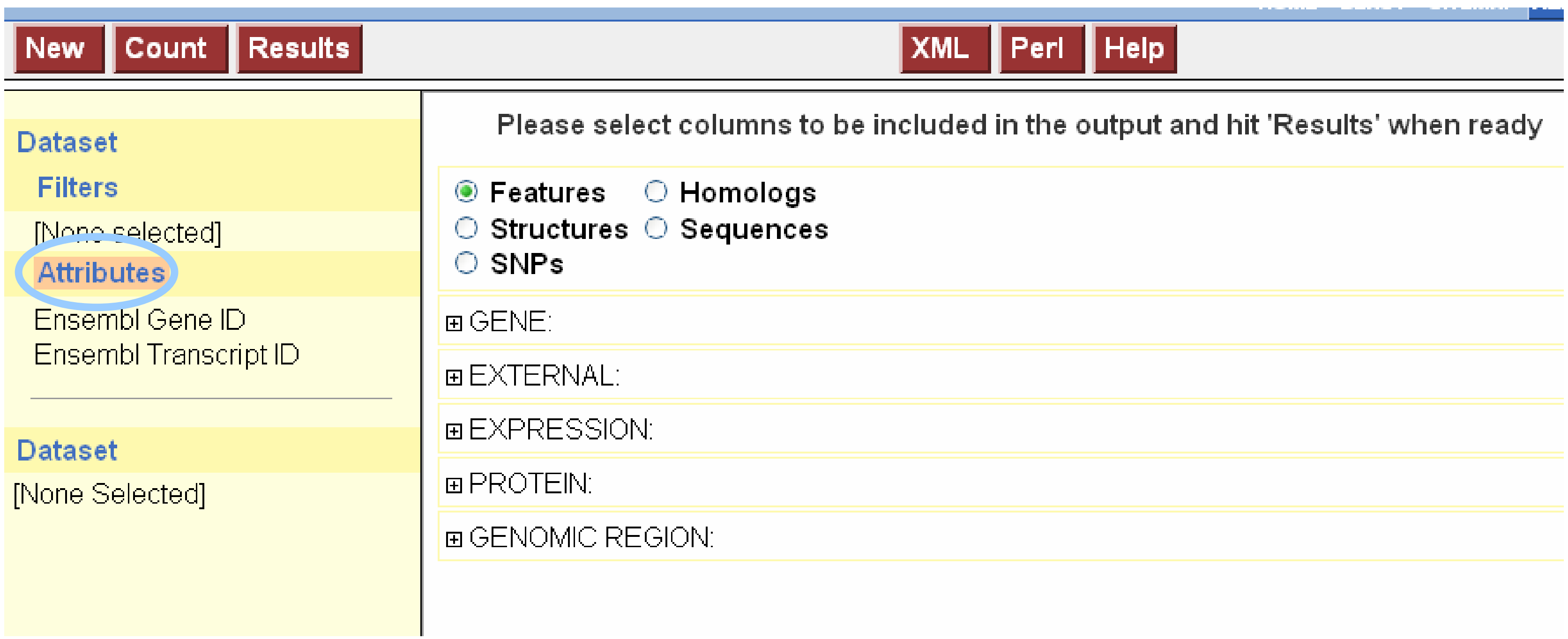
⊞ MULTI SPECIES COMPARISONS:

⊞ PROTEIN:

⊞ SNP:

Filters define which genes we are looking at.

Attributes attach information



Please select columns to be included in the output and hit 'Results' when ready

● Features ○ Homologs
○ Structures ○ Sequences
○ SNPs

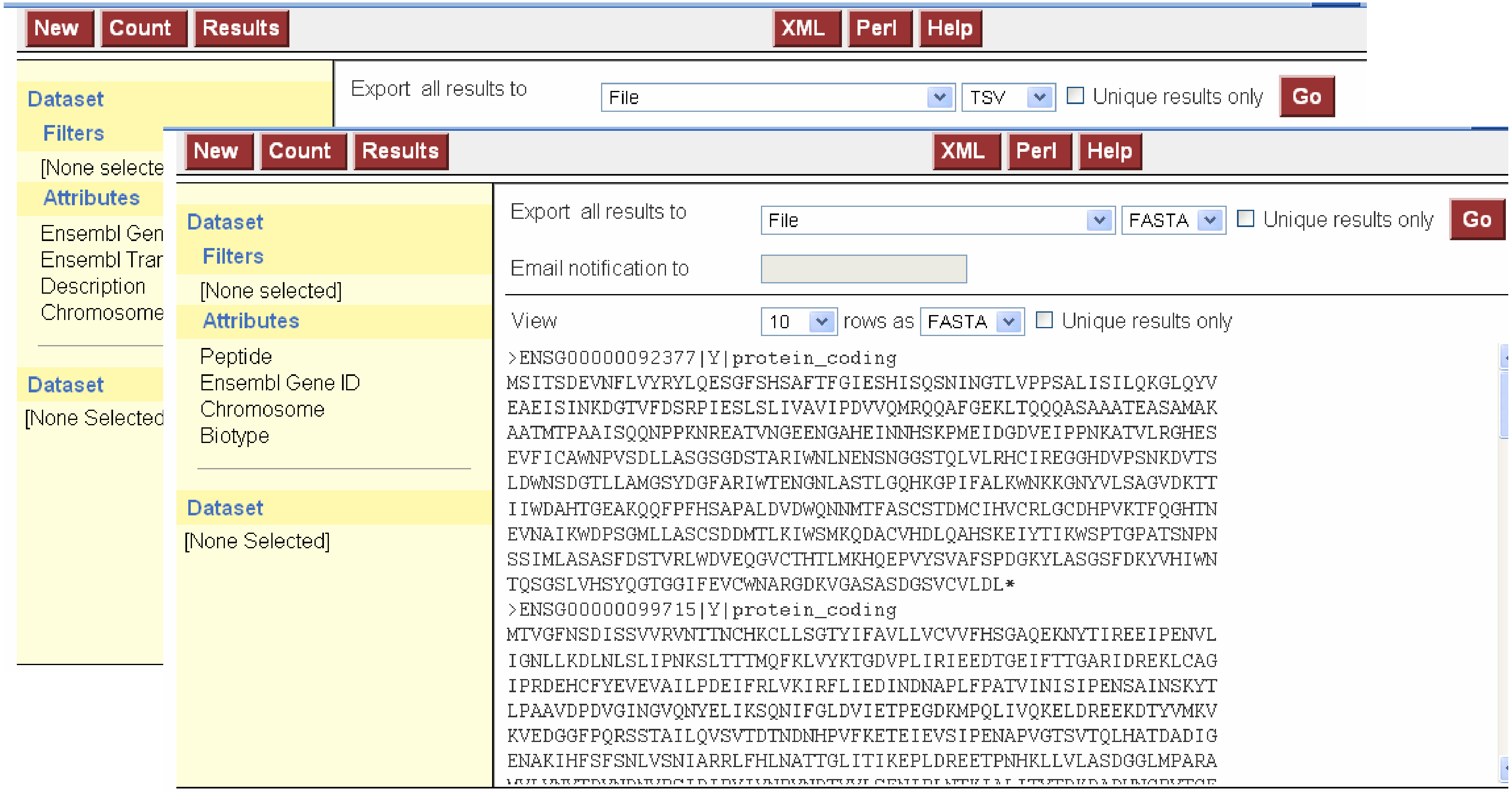
Filters: [None selected]

Attributes: Ensembl Gene ID, Ensembl Transcript ID

Dataset: [None Selected]

Determine output columns with Attributes.

Results



Dataset
Filters
[None selecte
Attributes
Ensembl Gen
Ensembl Trar
Description
Chromosome
Dataset
[None Selected

Export all results to File TSV □ Unique results only

Dataset
Filters
[None selected]
Attributes
Peptide
Ensembl Gene ID
Chromosome
Biotype
Dataset
[None Selected]

Export all results to File FASTA □ Unique results only

Email notification to

View 10 rows as FASTA □ Unique results only

```
>ENSG00000092377|Y|protein_coding
MSITSDEVNFLVYRYLQESGFSHSAFTFGIESHISQSNINGTLVPPSALISILQKGLQYV
EAEISINKDGTVFDSRPIESLSLIVAVIPDVVQMRQQAFGEKLTQQQASAAATEASAMAK
AATMTPAAISQQNPPKNREATVNGEENGAHEINNHSKPMEIDGDVEIPPNKATVLRGHES
EVFICAWNPVSDLLASGSGDSTARIWNLNENSNGGSTQLVLRHCIREGGHDVPSNKDVTS
LDWNSDGTLLAMGSYDGFARIWTENGNLASTLGQHKGPIFALKWNKKGNYVLSAGVDKTT
IIWDAHTGEAKQQFPFHSAPALDVDWQNNMTFASCSTDMCIHVCRLGCDHPVKTFQGHTN
EVNAIKWDPSGMLLASCSDDMTLKIWSMKQDACVHDLQAHSKEIYTIKWSPTGPATSNPN
SSIMLASASFDSTVRLWDVEQGVCTHTLMKHQEPVYSVAFSPDGKYLASGSFDKYVHIWN
TQSGSLVHSYQGTGGIFEVCWNARGDKVGASASDGSVCVLDL*
>ENSG00000099715|Y|protein_coding
MTVGFNSDISSVVRVNTTNCHKCLLSGTYIFAVLLVCVVFHSGAQEKNYTIREEIPENVL
IGNLLKDLNLSLIPNKSLTTTMQFKLVYKTGDVPLIRIEEDTGEIFTTGARIDREKLCAG
IPRDEHCFYEVEVAILPDEIFRLVKIRFLIEDINDNAPLFPATVINISIPENSAINSKYT
LPAAVDPDVGINGVQNYELIKSQNIFGLDVIETPEGDKMPQLIVQKELDREEKDTYVMKV
KVEDGGFPQRSSTAILQVSVTDTNDNHPVFKETEIEVSIPENAPVGTSVTQLHATDADIG
ENAKIHFSFSNLVSNIARRLFHLNATTGLITIKEPLDREETPNHKLLVLASDGGLMPARA
```

Tables or sequences

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**.

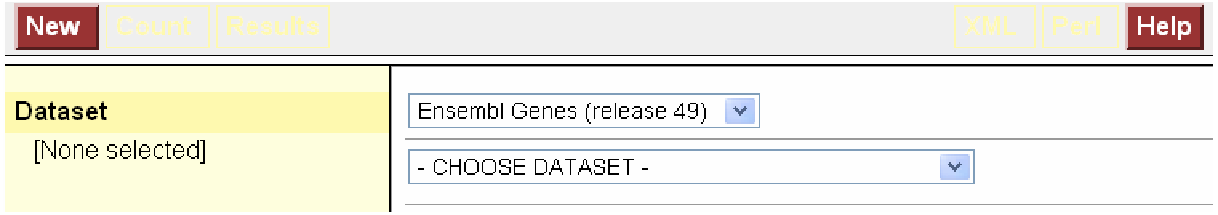
  Are there **Illumina probes** and **GO IDs** for these genes?

- In the query:

  Filters: what we know

  Attributes: what we want to know.

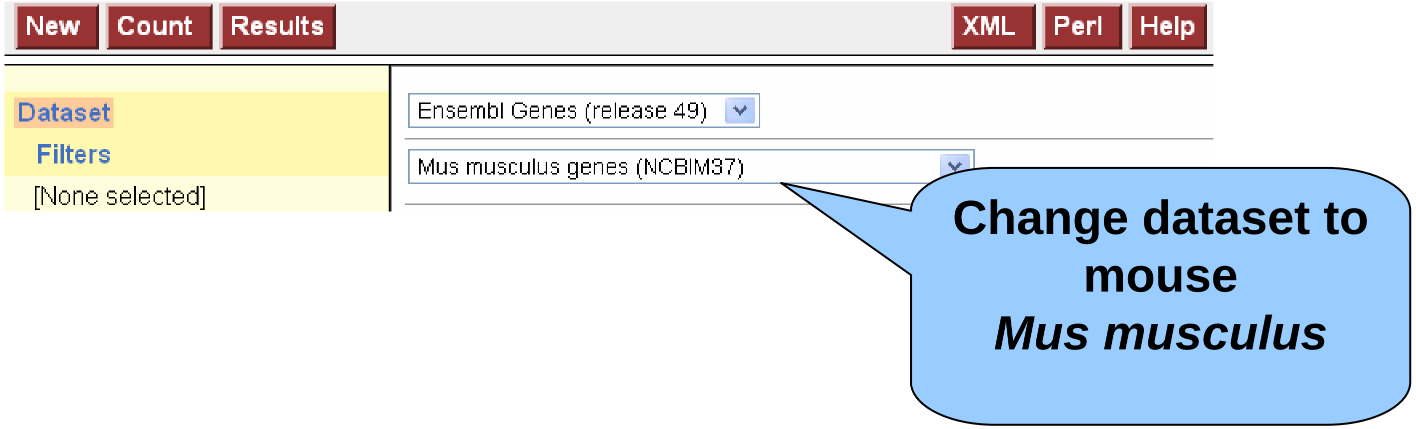
A Brief Example



Dataset
[None selected]

Ensembl Genes (release 49)

- CHOOSE DATASET -



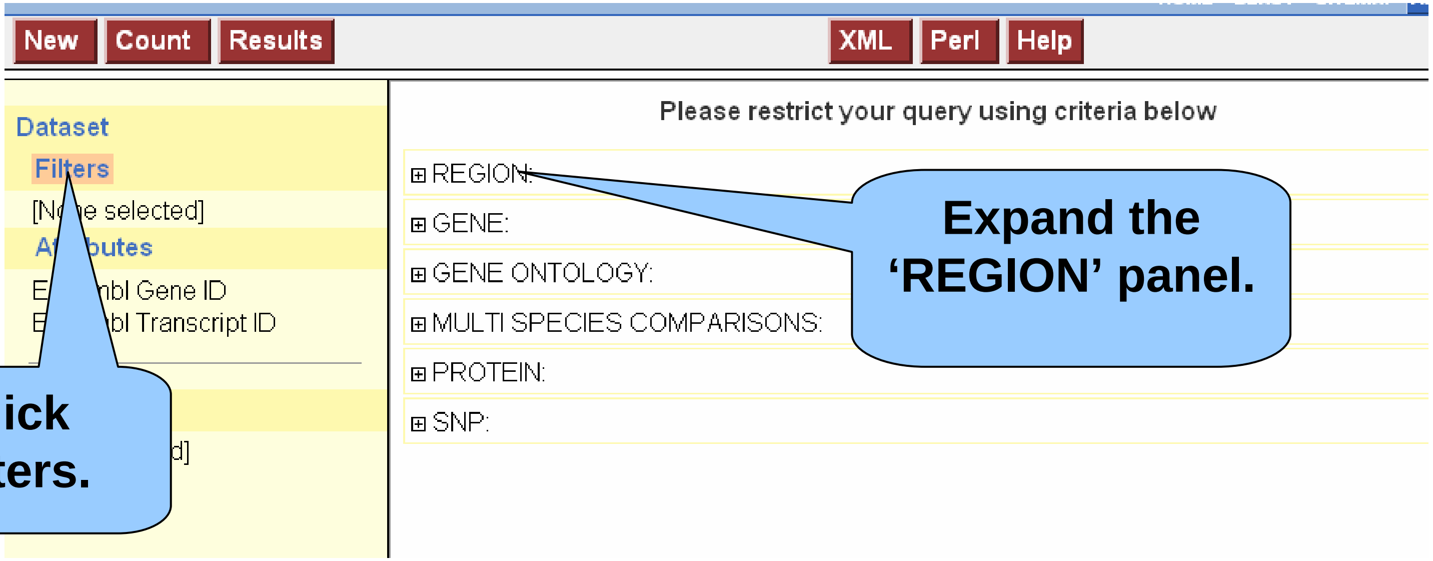
Dataset
Filters
[None selected]

Ensembl Genes (release 49)

Mus musculus genes (NCBIM37)

Change dataset to mouse
Mus musculus

Select the genes with Filters



Please restrict your query using criteria below

⊞ REGION:

⊞ GENE:

⊞ GENE ONTOLOGY:

⊞ MULTI SPECIES COMPARISONS:

⊞ PROTEIN:

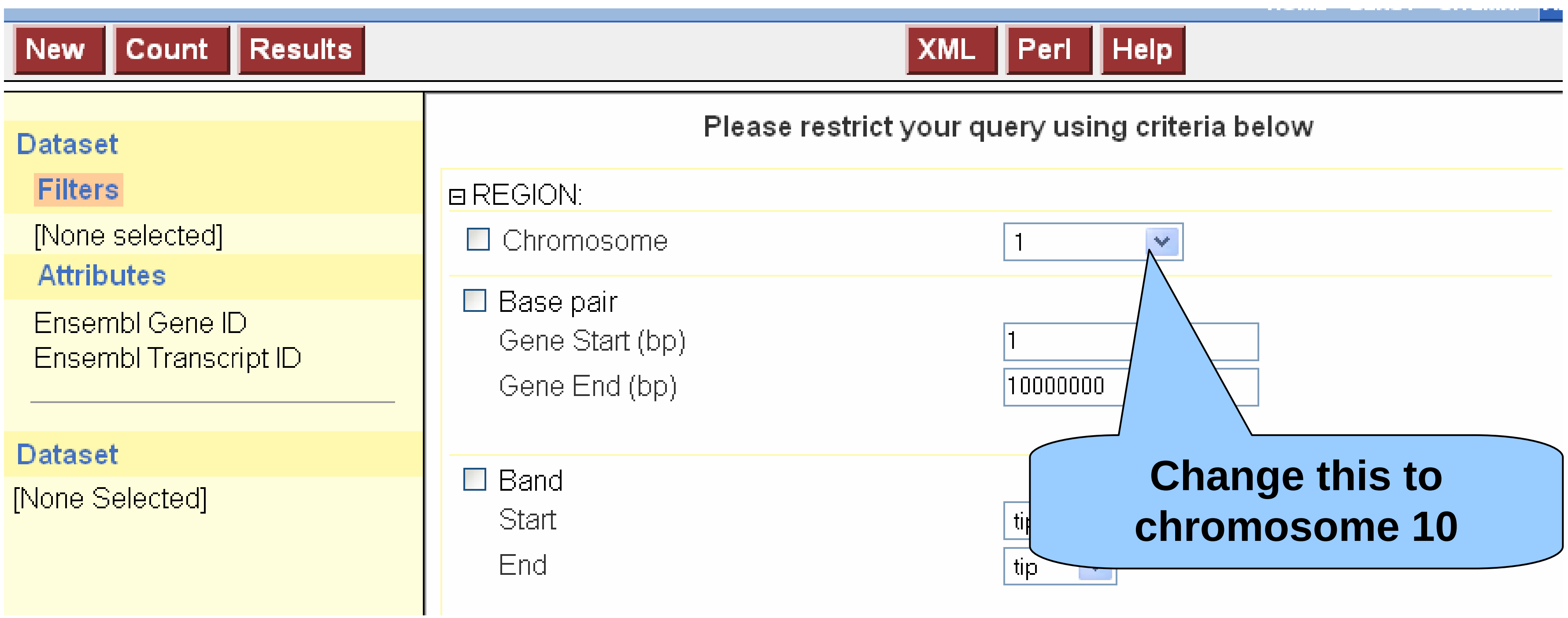
⊞ SNP:

Click Filters.

Expand the ‘REGION’ panel.

We are looking for mouse genes on chromosome 10 that are protein coding.

Filters (selecting the genes)



[None selected]

Ensembl Gene ID
Ensembl Transcript ID

[None Selected]

Please restrict your query using criteria below

⊟ REGION:

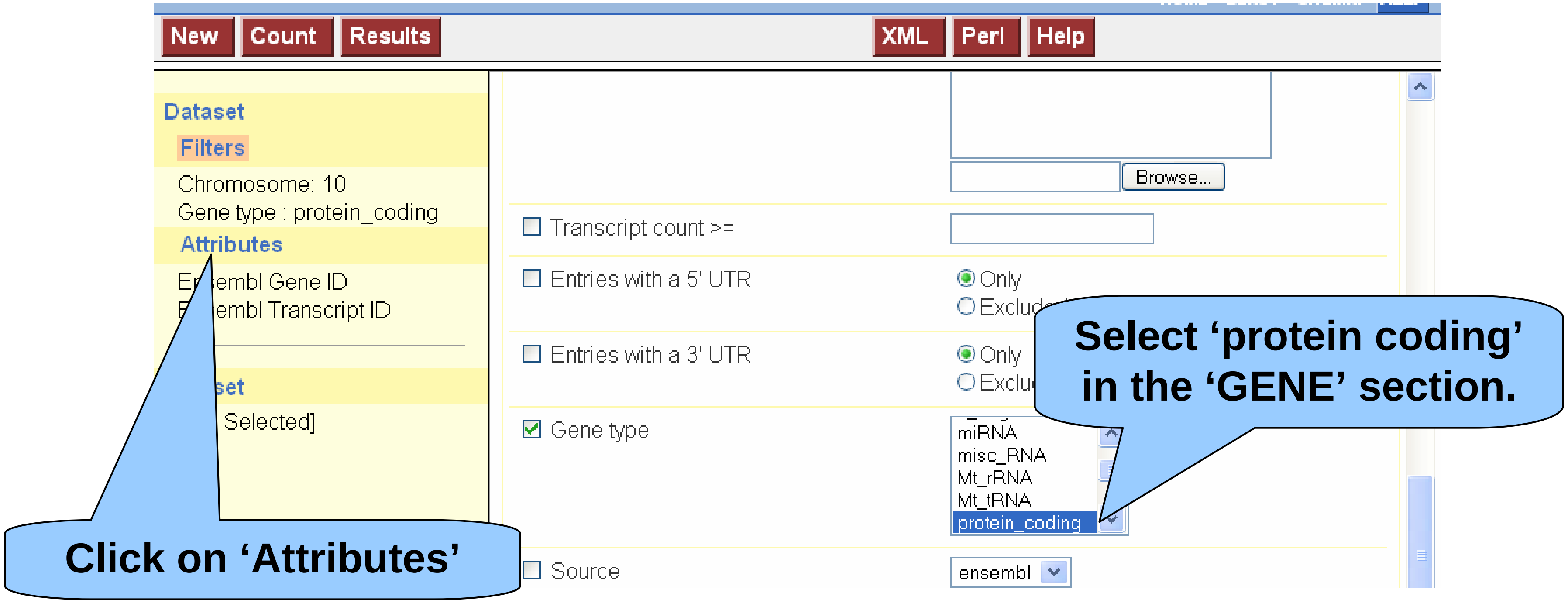
□ Chromosome 1

□ Base pair
Gene Start (bp) 1
Gene End (bp) 10000000

□ Band
Start tip
End tip

Change this to chromosome 10

Filters (selecting the genes)



Dataset

Filters

Chromosome: 10

Gene type : protein_coding

Attributes

E[ ]embl Gene ID

E[ ]embl Transcript ID

set

Selected]

☐ Transcript count >=

☐ Entries with a 5' UTR ● Only ○ Exclud

☐ Entries with a 3' UTR ● Only ○ Exclu

☑ Gene type

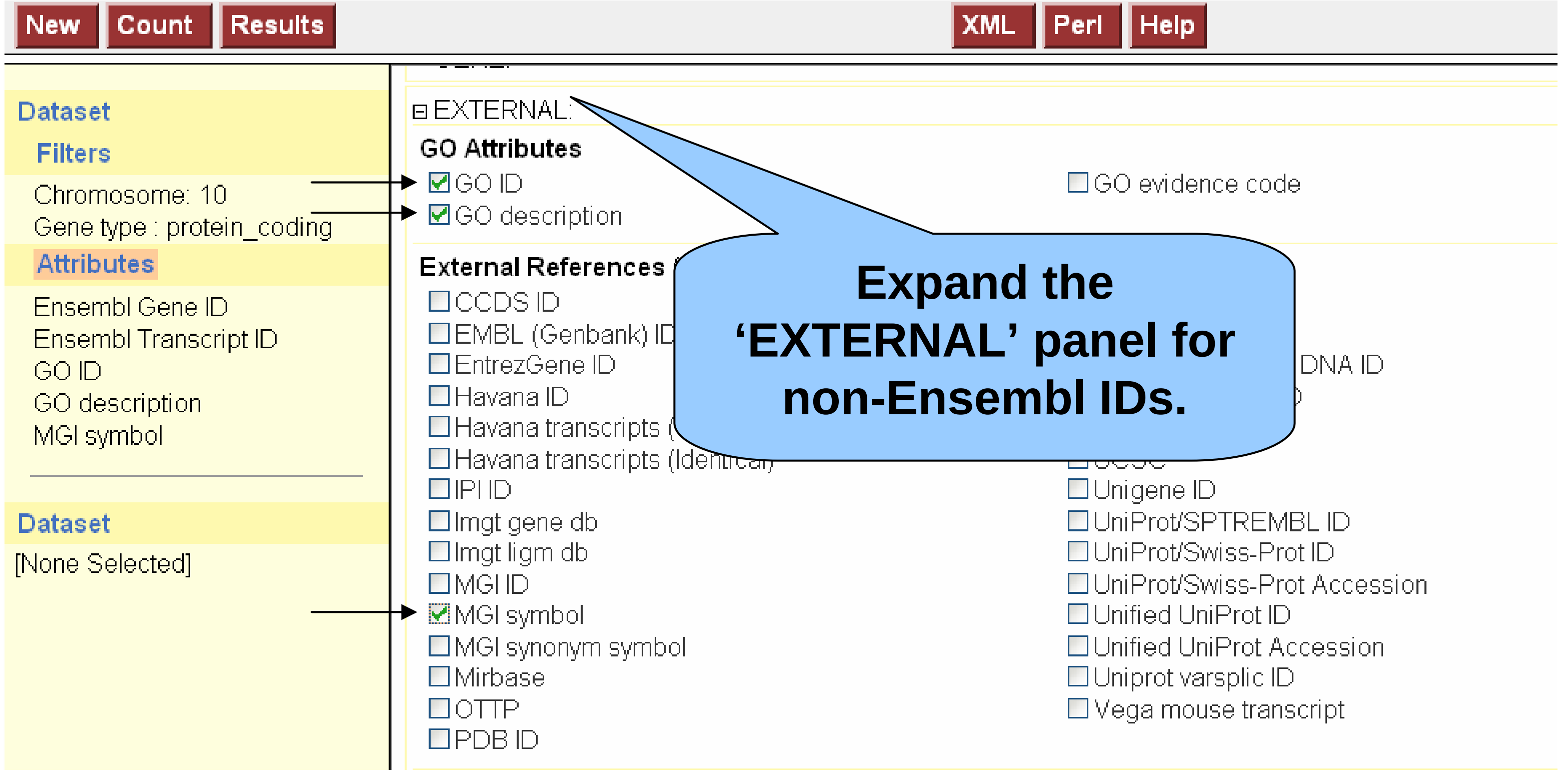
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
protein_coding

☐ Source ensembl

Select ‘protein coding’ in the ‘GENE’ section.

Click on ‘Attributes’

Attributes (Output Options)



Dataset

Filters

Chromosome: 10
Gene type : protein_coding

Attributes

Ensembl Gene ID
Ensembl Transcript ID
GO ID
GO description
MGI symbol

Dataset

[None Selected]

⊟ EXTERNAL:

GO Attributes

- ☑ GO ID
- ☑ GO description
- □ GO evidence code

External References (

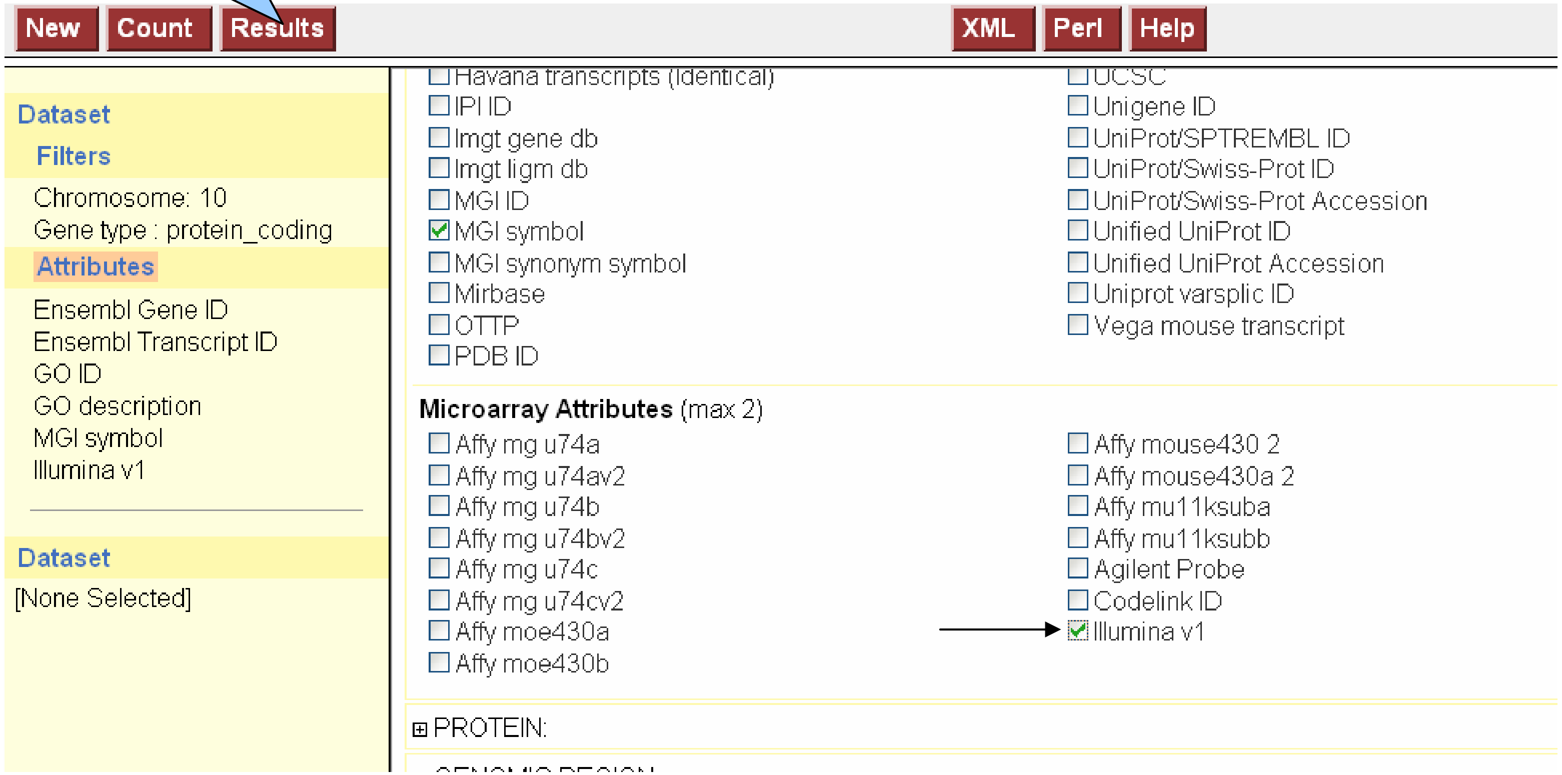
- □ CCDS ID
- □ EMBL (Genbank) ID
- □ EntrezGene ID
- □ Havana ID
- □ Havana transcripts (
- □ Havana transcripts (Identical)
- □ IPI ID
- □ Imgt gene db
- □ Imgt ligm db
- □ MGI ID
- ☑ MGI symbol
- □ MGI synonym symbol
- □ Mirbase
- □ OTTP
- □ PDB ID
- DNA ID
- □ Unigene ID
- □ UniProt/SPTREMBL ID
- □ UniProt/Swiss-Prot ID
- □ UniProt/Swiss-Prot Accession
- □ Unified UniProt ID
- □ Unified UniProt Accession
- □ Uniprot varsplic ID
- □ Vega mouse transcript

Expand the ‘EXTERNAL’ panel for non-Ensembl IDs.

We would like GO terms and IDs in MGI (the Mouse Genome Informatics site).

Attributes (Output)

Click ‘Results’



Dataset

Filters

Chromosome: 10
Gene type : protein_coding

Attributes

Ensembl Gene ID
Ensembl Transcript ID
GO ID
GO description
MGI symbol
Illumina v1

Dataset

[None Selected]

□ Havana transcripts (identical)
□ IPI ID
□ Imgt gene db
□ Imgt ligm db
□ MGI ID
☑ MGI symbol
□ MGI synonym symbol
□ Mirbase
□ OTTP
□ PDB ID

□ UCSC
□ Unigene ID
□ UniProt/SPTREMBL ID
□ UniProt/Swiss-Prot ID
□ UniProt/Swiss-Prot Accession
□ Unified UniProt ID
□ Unified UniProt Accession
□ Uniprot varsplic ID
□ Vega mouse transcript

Microarray Attributes (max 2)

□ Affy mg u74a
□ Affy mg u74av2
□ Affy mg u74b
□ Affy mg u74bv2
□ Affy mg u74c
□ Affy mg u74cv2
□ Affy moe430a
□ Affy moe430b

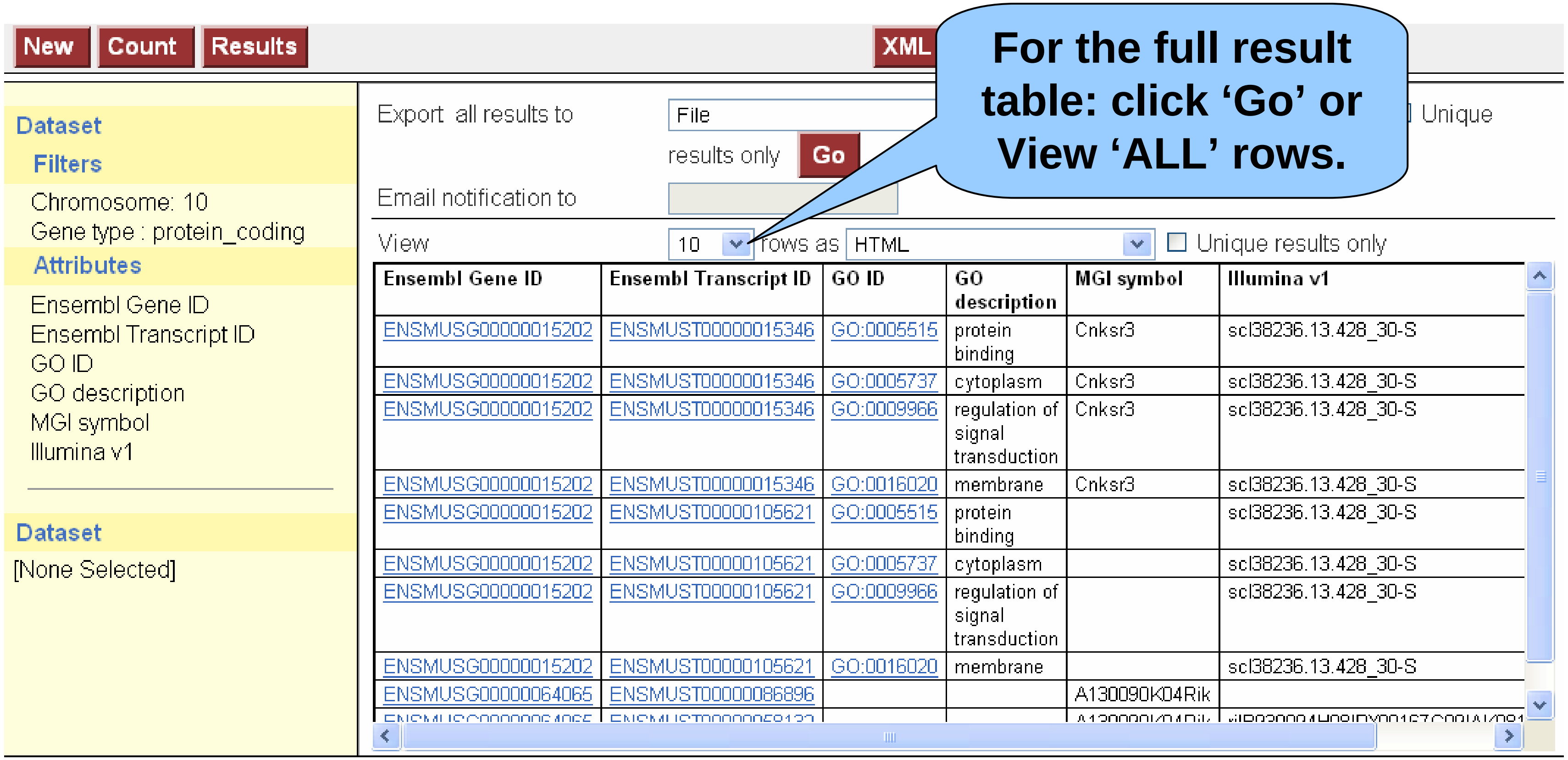
□ Affy mouse430 2
□ Affy mouse430a 2
□ Affy mu11ksuba
□ Affy mu11ksubb
□ Agilent Probe
□ Codelink ID
☑ Illumina v1

⊞ PROTEIN:

GENOMIC REGION

Scroll down to add ‘Illumina v1’ probes that map to these genes.

The Results Table - Preview



Dataset

Filters

Chromosome: 10
Gene type : protein_coding

Attributes

Ensembl Gene ID
Ensembl Transcript ID
GO ID
GO description
MGI symbol
Illumina v1

Dataset

[None Selected]

Export all results to File □ Unique
results only Go

Email notification to

View 10 rows as HTML □ Unique results only

Ensembl Gene ID	Ensembl Transcript ID	GO ID	GO description	MGI symbol	Illumina v1
ENSMUSG00000015202	ENSMUST00000015346	GO:0005515	protein binding	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000015346	GO:0005737	cytoplasm	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000015346	GO:0009966	regulation of signal transduction	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000015346	GO:0016020	membrane	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0005515	protein binding		scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0005737	cytoplasm		scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0009966	regulation of signal transduction		scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0016020	membrane		scl38236.13.428_30-S
ENSMUSG00000064065	ENSMUST00000086896			A130090K04Rik	
ENSMUSG00000064065	ENSMUST00000058132			A130090K04Rik	[illegible]

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.

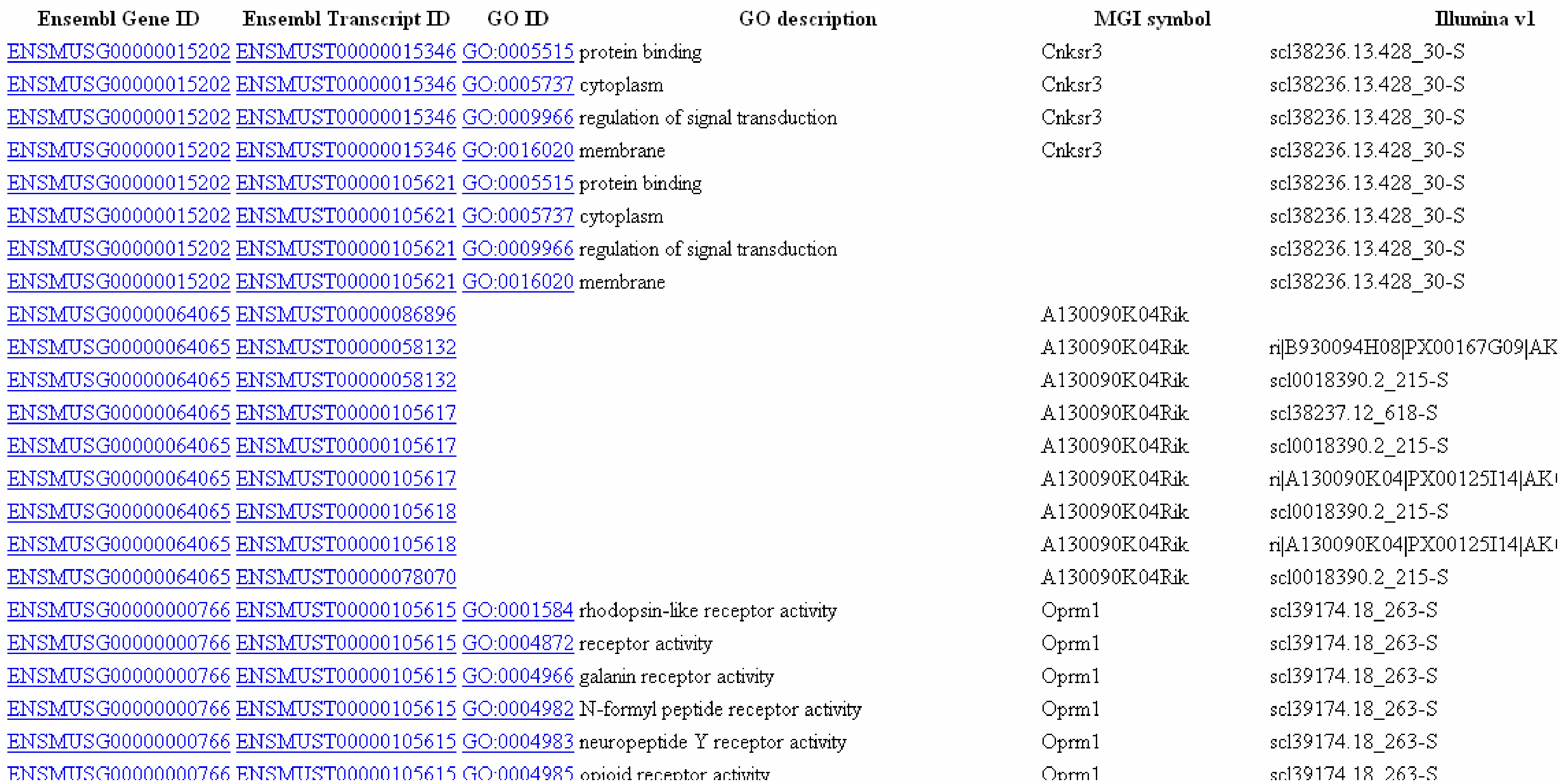
Full Result Table

Ensembl Gene and Transcript IDs

GO terms

MGI symbol

Illumina probes



Ensembl Gene ID	Ensembl Transcript ID	GO ID	GO description	MGI symbol	Illumina v1
ENSMUSG00000015202	ENSMUST00000015346	GO:0005515	protein binding	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000015346	GO:0005737	cytoplasm	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000015346	GO:0009966	regulation of signal transduction	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000015346	GO:0016020	membrane	Cnksr3	scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0005515	protein binding		scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0005737	cytoplasm		scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0009966	regulation of signal transduction		scl38236.13.428_30-S
ENSMUSG00000015202	ENSMUST00000105621	GO:0016020	membrane		scl38236.13.428_30-S
ENSMUSG00000064065	ENSMUST00000086896			A130090K04Rik	
ENSMUSG00000064065	ENSMUST00000058132			A130090K04Rik	ri|B930094H08|PX00167G09|AK
ENSMUSG00000064065	ENSMUST00000058132			A130090K04Rik	scl0018390.2_215-S
ENSMUSG00000064065	ENSMUST00000105617			A130090K04Rik	scl38237.12_618-S
ENSMUSG00000064065	ENSMUST00000105617			A130090K04Rik	scl0018390.2_215-S
ENSMUSG00000064065	ENSMUST00000105617			A130090K04Rik	ri|A130090K04|PX00125I14|AK
ENSMUSG00000064065	ENSMUST00000105618			A130090K04Rik	scl0018390.2_215-S
ENSMUSG00000064065	ENSMUST00000105618			A130090K04Rik	ri|A130090K04|PX00125I14|AK
ENSMUSG00000064065	ENSMUST00000078070			A130090K04Rik	scl0018390.2_215-S
ENSMUSG00000000766	ENSMUST00000105615	GO:0001584	rhodopsin-like receptor activity	Oprm1	scl39174.18_263-S
ENSMUSG00000000766	ENSMUST00000105615	GO:0004872	receptor activity	Oprm1	scl39174.18_263-S
ENSMUSG00000000766	ENSMUST00000105615	GO:0004966	galanin receptor activity	Oprm1	scl39174.18_263-S
ENSMUSG00000000766	ENSMUST00000105615	GO:0004982	N-formyl peptide receptor activity	Oprm1	scl39174.18_263-S
ENSMUSG00000000766	ENSMUST00000105615	GO:0004983	neuropeptide Y receptor activity	Oprm1	scl39174.18_263-S
ENSMUSG00000000766	ENSMUST00000105615	GO:0004985	opioid receptor activity	Oprm1	scl39174.18_263-S

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**.

  Are there **Illumina probes** and **GO IDs** for these genes?

- In the query:

  Filters: what we know

  Attributes: columns in the **Result Table**

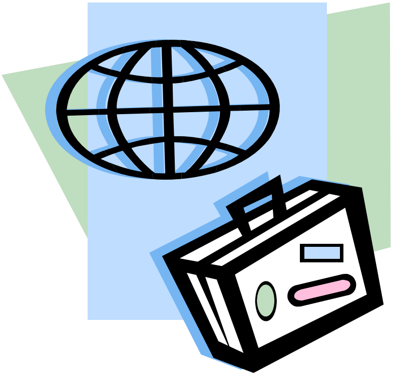
Other Export Options (Attributes)

- ❖ 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

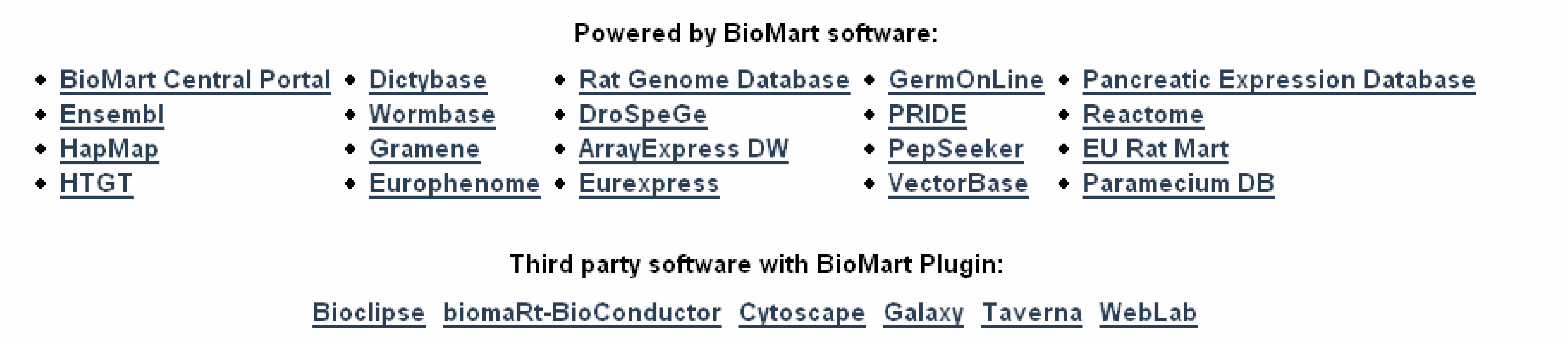
BioMart around the world…



BioMart started at Ensembl…

To where has it travelled?

Central Portal



Powered by BioMart software:

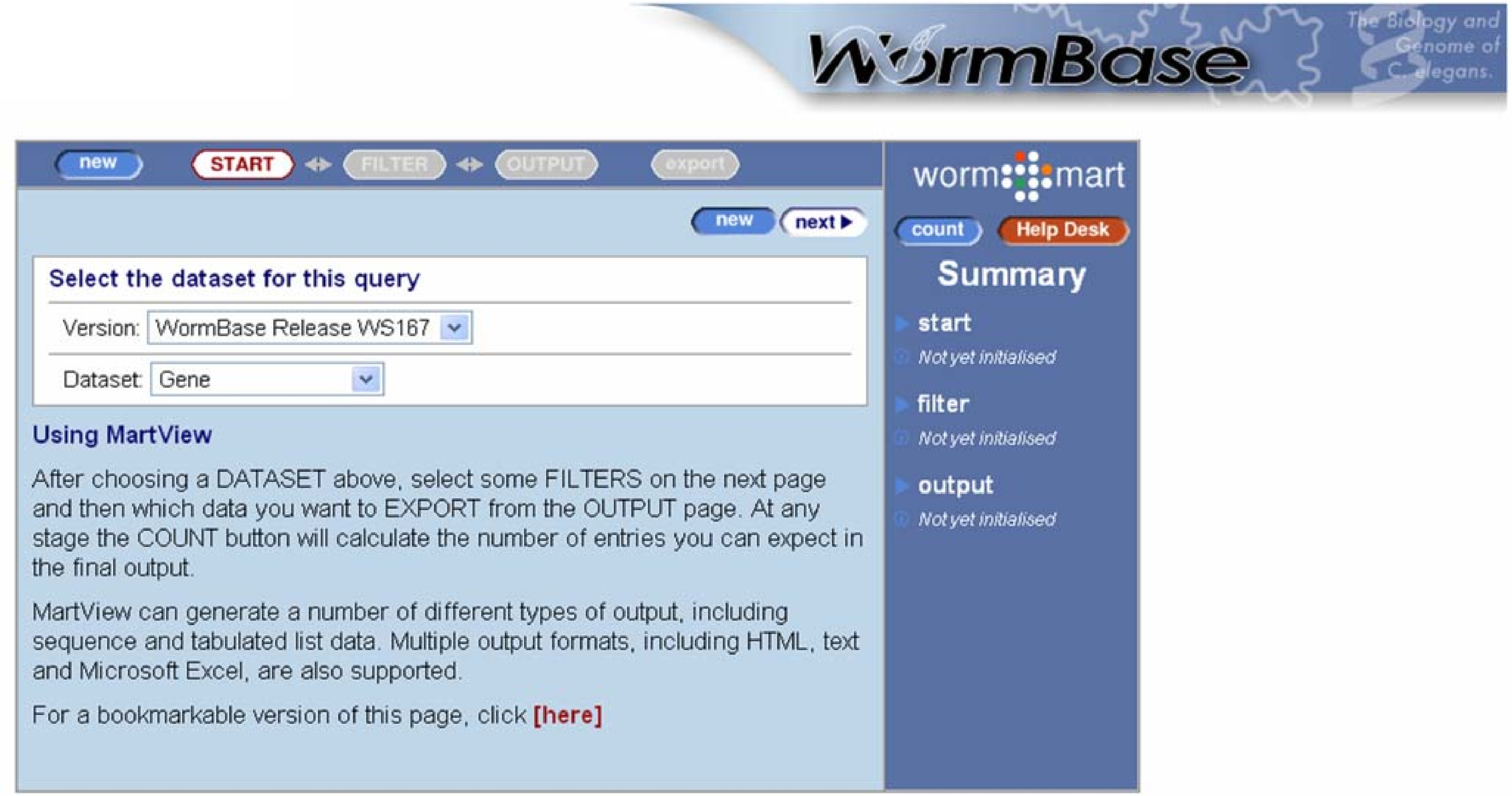
- BioMart Central Portal
- Ensembl
- HapMap
- HTGT
- Dictybase
- Wormbase
- Gramene
- Europhenome
- Rat Genome Database
- DroSpeGe
- ArrayExpress DW
- Eurexpress
- GermOnLine
- PRIDE
- PepSeeker
- VectorBase
- Pancreatic Expression Database
- Reactome
- EU Rat Mart
- Paramecium DB

Third party software with BioMart Plugin:

Bioclipse biomaRt-BioConductor Cytoscape Galaxy Taverna WebLab

www.biomart.org

WormBase



Select the dataset for this query

Version: WormBase Release WS167

Dataset: Gene

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]

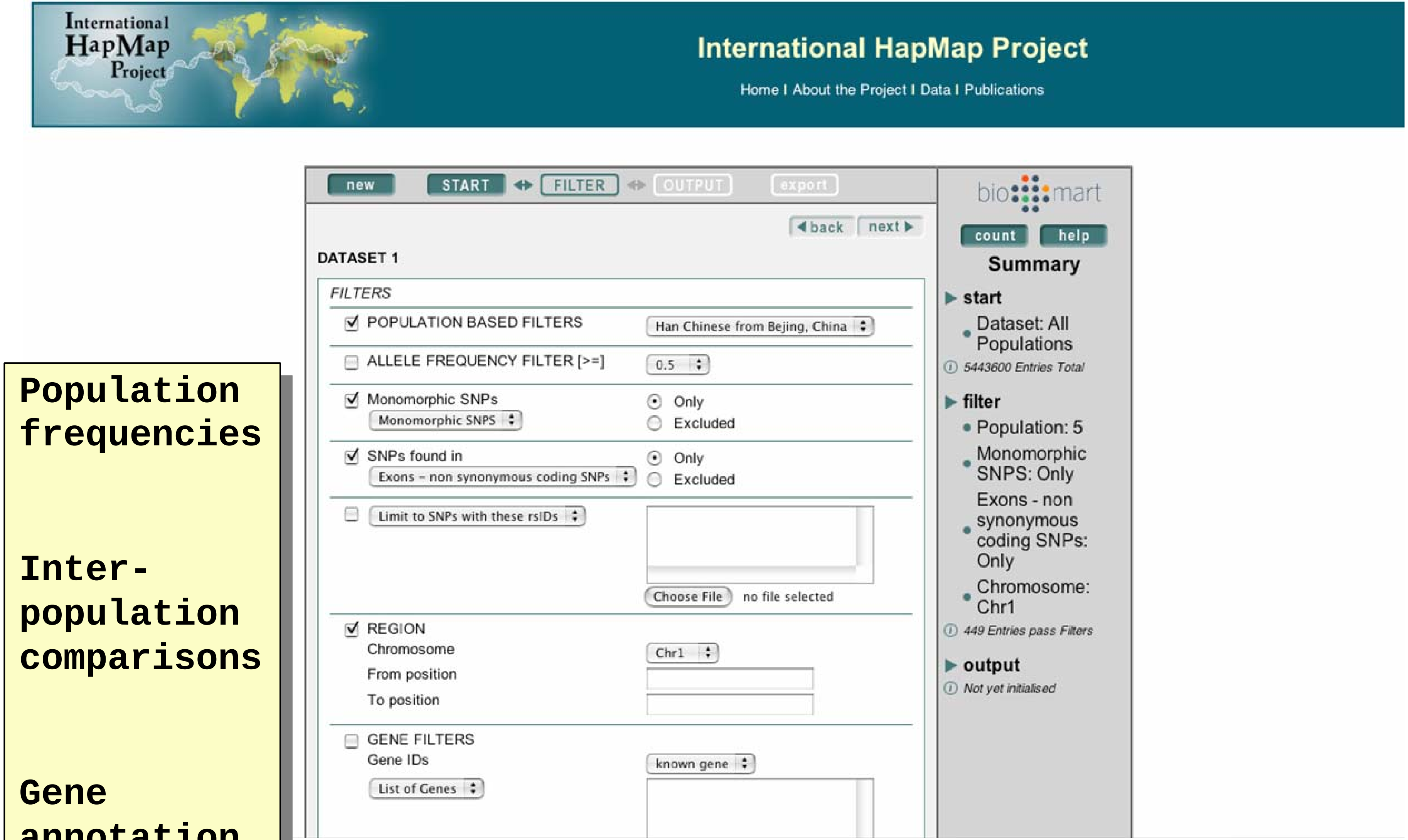
Summary

start
Not yet initialised

filter
Not yet initialised

output
Not yet initialised

HapMap



International HapMap Project

Home | About the Project | Data | Publications

DATASET 1

FILTERS

☑ POPULATION BASED FILTERS — Han Chinese from Bejing, China

□ ALLELE FREQUENCY FILTER [>=] — 0.5

☑ Monomorphic SNPs — Monomorphic SNPS — ● Only ○ Excluded

☑ SNPs found in — Exons - non synonymous coding SNPs — ● Only ○ Excluded

□ Limit to SNPs with these rsIDs — Choose File — no file selected

☑ REGION
Chromosome — Chr1
From position
To position

□ GENE FILTERS
Gene IDs — known gene
List of Genes

Summary

start

- Dataset: All Populations

5443600 Entries Total

filter

- Population: 5
- Monomorphic SNPS: Only
- Exons - non synonymous coding SNPs: Only
- Chromosome: Chr1

449 Entries pass Filters

output

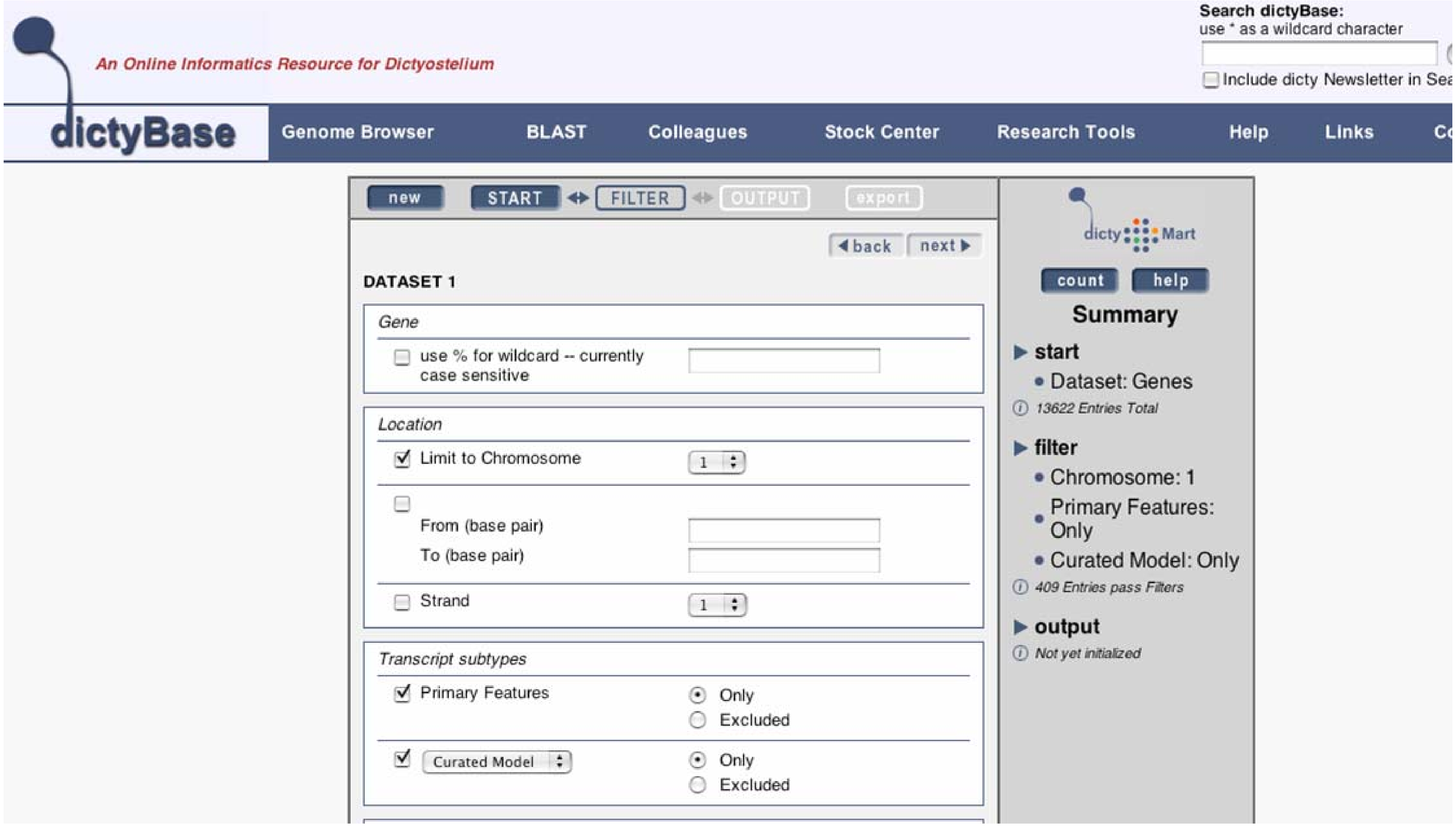
Not yet initialised

Population frequencies

Inter-population comparisons

Gene annotation

DictyBase



Search dictyBase:
use * as a wildcard character

□ Include dicty Newsletter in Sea

DATASET 1

Gene

□ use % for wildcard -- currently case sensitive

Location

☑ Limit to Chromosome 1

□ From (base pair)
To (base pair)

□ Strand 1

Transcript subtypes

☑ Primary Features ● Only ○ Excluded

☑ Curated Model ● Only ○ Excluded

Summary

▶ **start**

- Dataset: Genes

13622 Entries Total

▶ **filter**

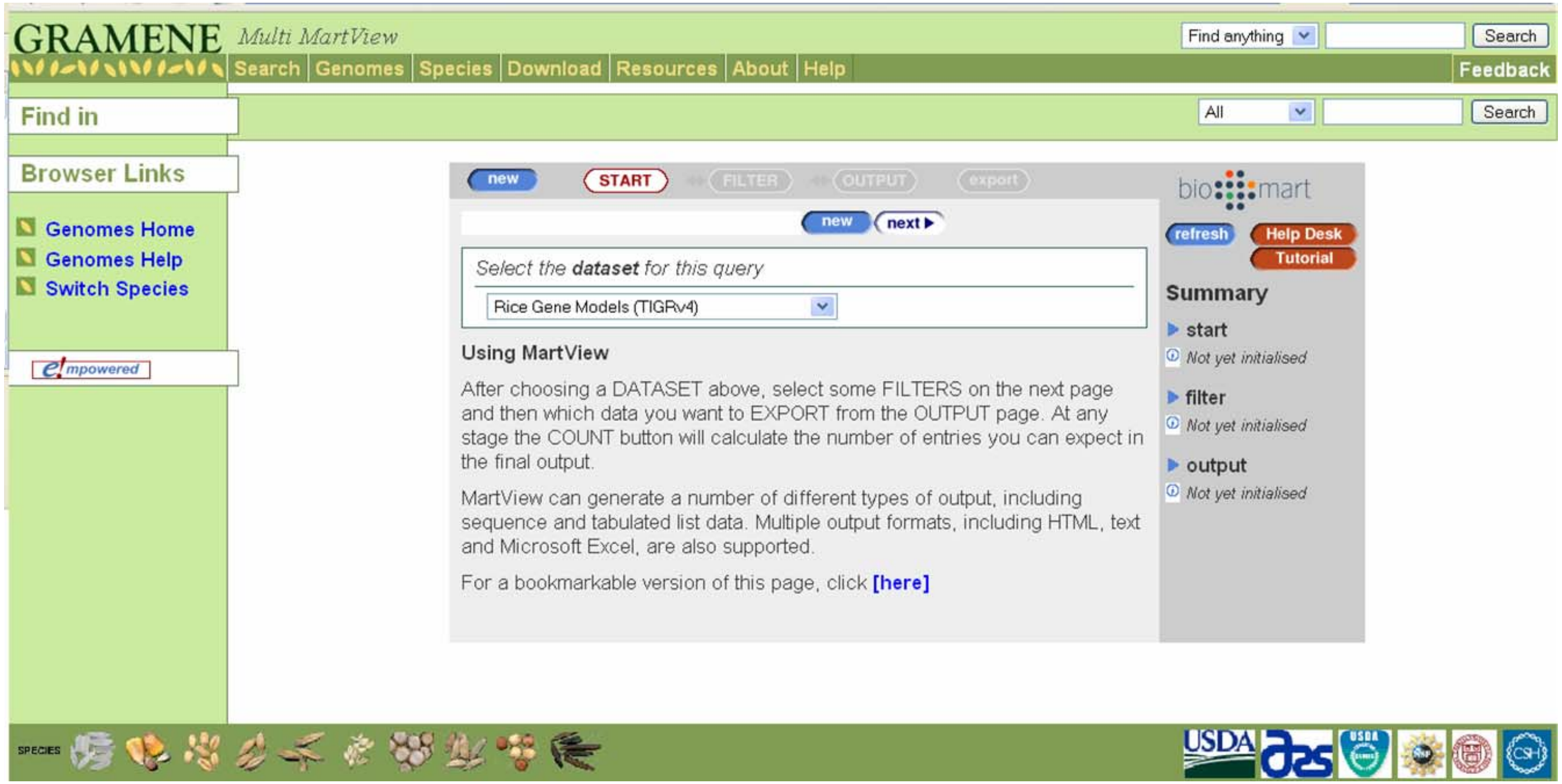
- Chromosome: 1
- Primary Features: Only
- Curated Model: Only

409 Entries pass Filters

▶ **output**

Not yet initialized

GRAMENE



Select the **dataset** *for this query*

Rice Gene Models (TIGRv4)

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]**

Summary

▶ **start**

Not yet initialised

▶ **filter**

Not yet initialised

▶ **output**

Not yet initialised

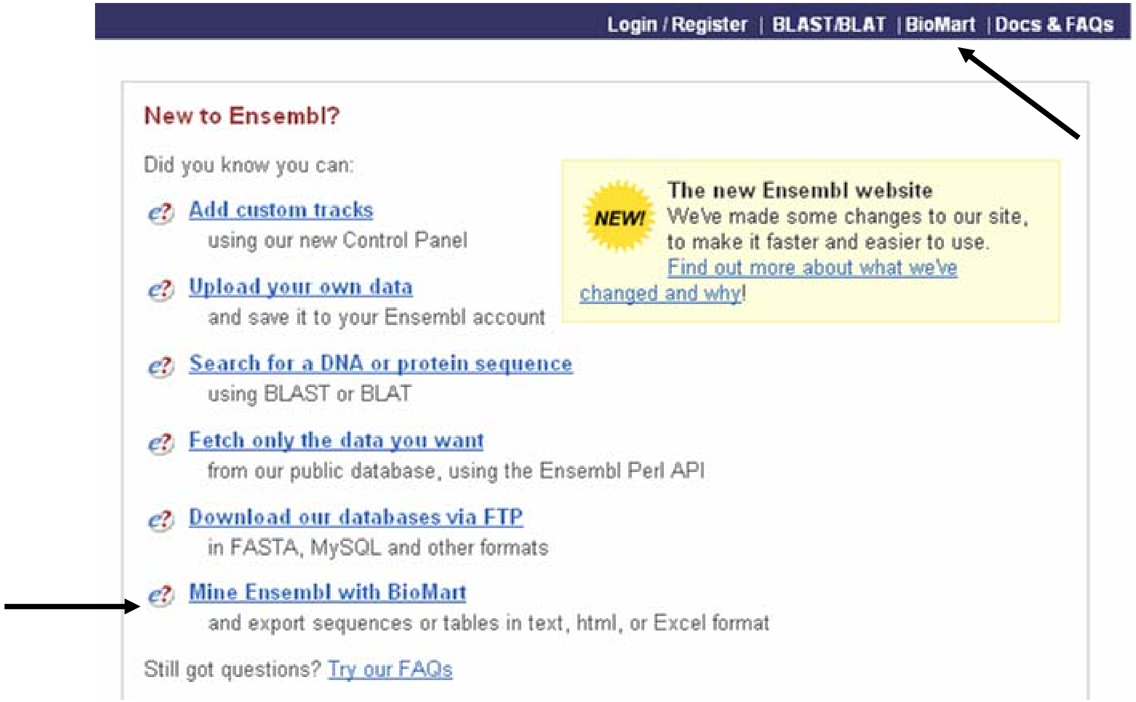
www.gramene.org

How to Get There

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

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

- Or click on ‘BioMart’ from Ensembl



Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

New to Ensembl?

Did you know you can:

Add custom tracks
using our new Control Panel

Upload your own data
and save it to your Ensembl account

Search for a DNA or protein sequence
using BLAST or BLAT

Fetch only the data you want
from our public database, using the Ensembl Perl API

Download our databases via FTP
in FASTA, MySQL and other formats

Mine Ensembl with BioMart
and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs

NEW!
The new Ensembl website
We've made some changes to our site, to make it faster and easier to use. Find out more about what we've changed and why!