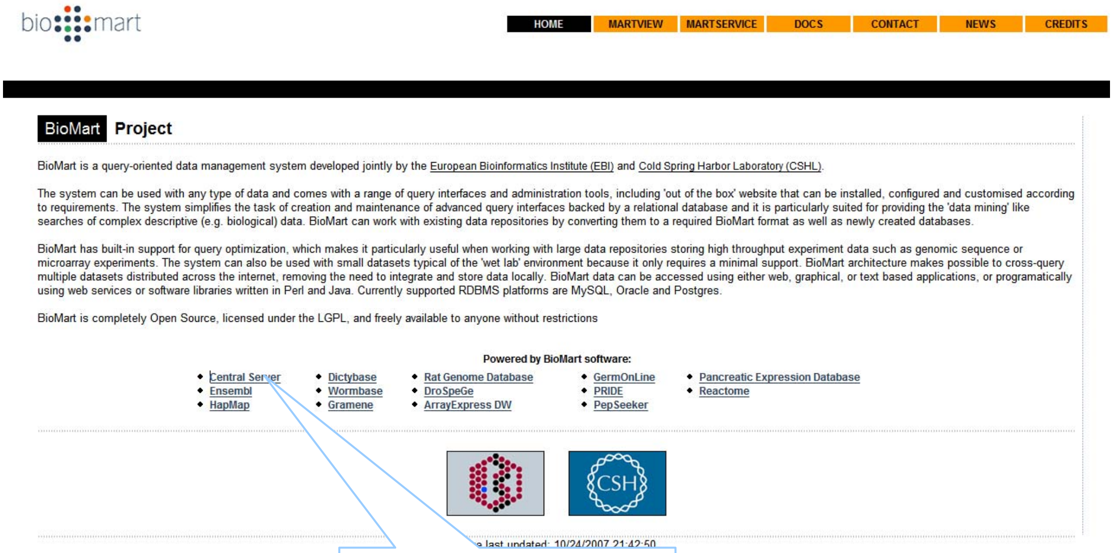
Data mining in Ensembl with BioMart
Worked Example – Demonstrating the Linked Dataset

BioMart can federate (join together) databases, in this example we will join two different datasets, Ensembl genes and RGD (the Rat Genome Database) to identify all Ensembl genes involved in carbohydrate metabolism in rat. First, we will limit our search to genes involved in the *carbohydrate metabolic process.* By linking the RGD and Ensembl databases, we ask for only genes in both databases (the intersection of the two sets). The *RGD ID*, *Ensembl gene* and *transcript ID*, along with the '*Disease Ontology*' *term* from RGD are all selected as output columns.

STEP 1:
Go to the BioMart Central server page
www.biomart.org



BioMart Project

BioMart is a query-oriented data management system developed jointly by the European Bioinformatics Institute (EBI) and Cold Spring Harbor Laboratory (CSHL).

The system can be used with any type of data and comes with a range of query interfaces and administration tools, including 'out of the box' website that can be installed, configured and customised according to requirements. The system simplifies the task of creation and maintenance of advanced query interfaces backed by a relational database and it is particularly suited for providing the 'data mining' like searches of complex descriptive (e.g. biological) data. BioMart can work with existing data repositories by converting them to a required BioMart format as well as newly created databases.

BioMart has built-in support for query optimization, which makes it particularly useful when working with large data repositories storing high throughput experiment data such as genomic sequence or microarray experiments. The system can also be used with small datasets typical of the 'wet lab' environment because it only requires a minimal support. BioMart architecture makes possible to cross-query multiple datasets distributed across the internet, removing the need to integrate and store data locally. BioMart data can be accessed using either web, graphical, or text based applications, or programatically using web services or software libraries written in Perl and Java. Currently supported RDBMS platforms are MySQL, Oracle and Postgres.

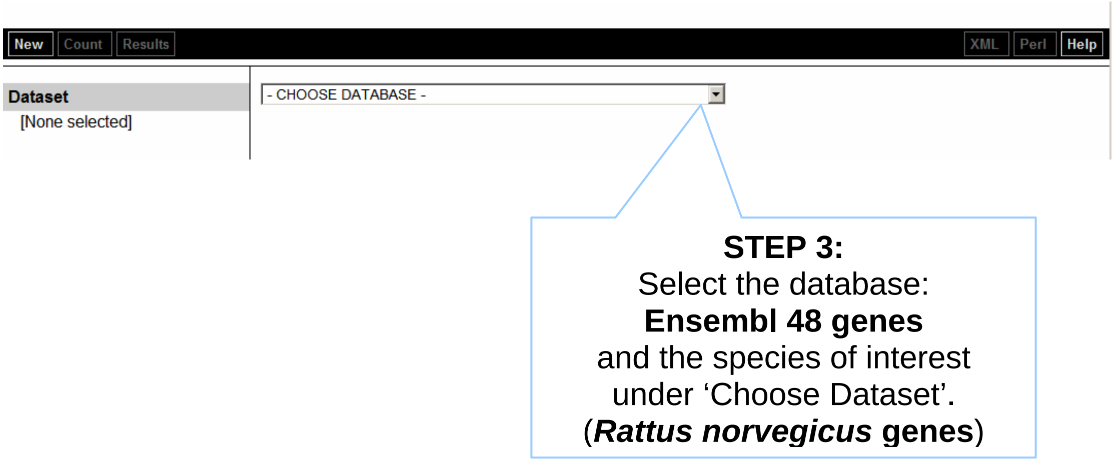
BioMart is completely Open Source, licensed under the LGPL, and freely available to anyone without restrictions

Powered by BioMart software:

- Central Server
- Ensembl
- HapMap
- Dictybase
- Wormbase
- Gramene
- Rat Genome Database
- DroSpeGe
- ArrayExpress DW
- GermOnLine
- PRIDE
- PepSeeker
- Pancreatic Expression Database
- Reactome

e last updated: 10/24/2007 21:42:50

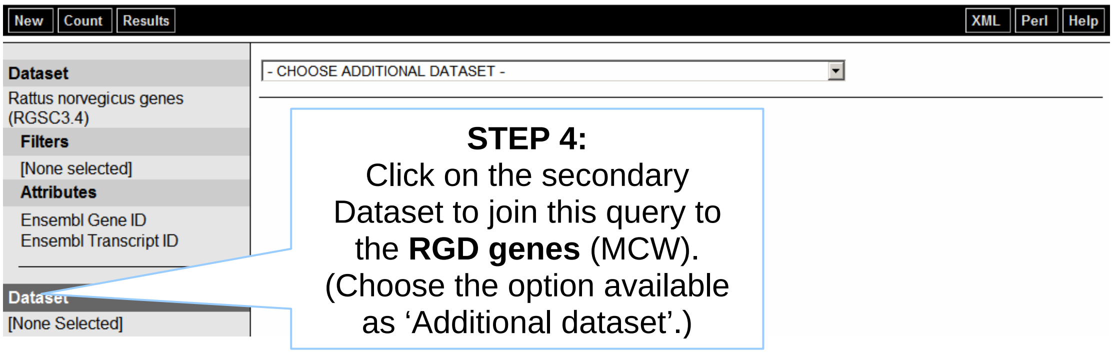
STEP 2:
Click on 'Central server'



Dataset
[None selected]

- CHOOSE DATABASE -

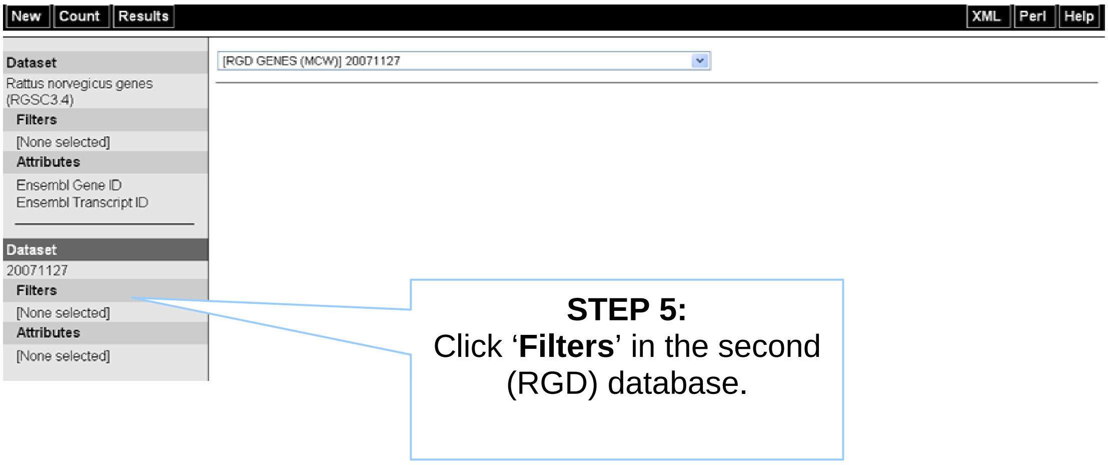
STEP 3:
Select the database:
Ensembl 48 genes
and the species of interest
under ‘Choose Dataset’.
(***Rattus norvegicus* genes**)



Dataset
Rattus norvegicus genes (RGSC3.4)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Dataset
[None Selected]

- CHOOSE ADDITIONAL DATASET -

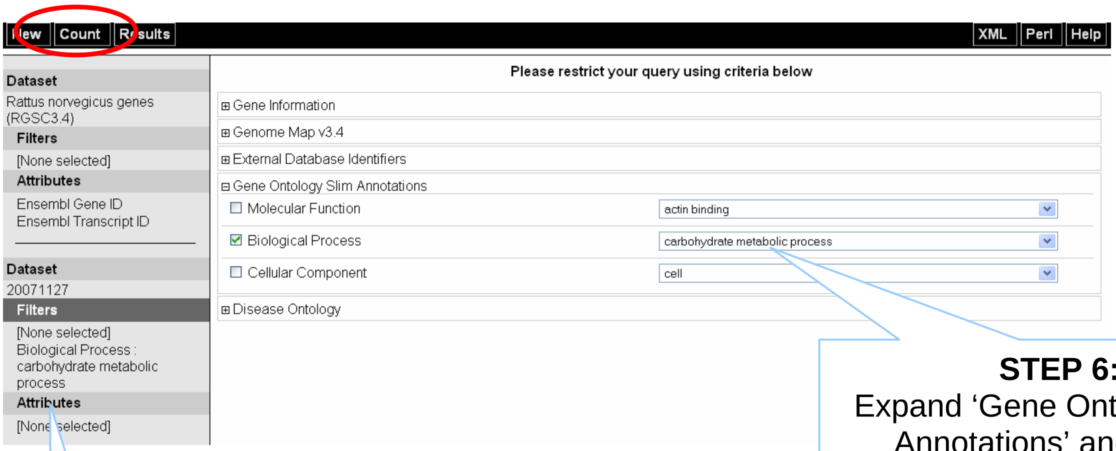
STEP 4:
Click on the secondary
Dataset to join this query to
the **RGD genes** (MCW).
(Choose the option available
as ‘Additional dataset’.)



Dataset
Rattus norvegicus genes (RGSC3.4)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Dataset
20071127
Filters
[None selected]
Attributes
[None selected]

[RGD GENES (MCW)] 20071127

STEP 5:
Click ‘**Filters**’ in the second
(RGD) database.



Dataset
Rattus norvegicus genes (RGSC3.4)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset
20071127
Filters
[None selected]
Biological Process : carbohydrate metabolic process
Attributes
[None selected]

Please restrict your query using criteria below

⊞ Gene Information
⊞ Genome Map v3.4
⊞ External Database Identifiers
⊟ Gene Ontology Slim Annotations

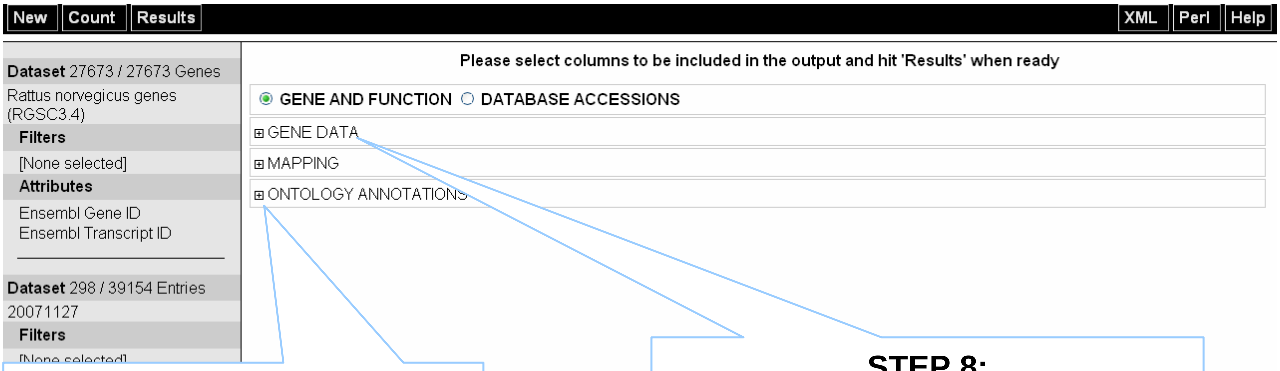
□ Molecular Function	actin binding
☑ Biological Process	carbohydrate metabolic process
□ Cellular Component	cell

⊞ Disease Ontology

STEP 6:
Expand ‘Gene Ontology Slim Annotations’ and select ‘Biological Process’ as ‘**carbohydrate metabolic process**’

The filters have determined our gene set. Click **‘Count’** (at the top) to see how many genes have passed these filters.

STEP 7:
Click on ‘**Attributes**’



Dataset 27673 / 27673 Genes
Rattus norvegicus genes (RGSC3.4)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset 298 / 39154 Entries
20071127
Filters
[None selected]

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

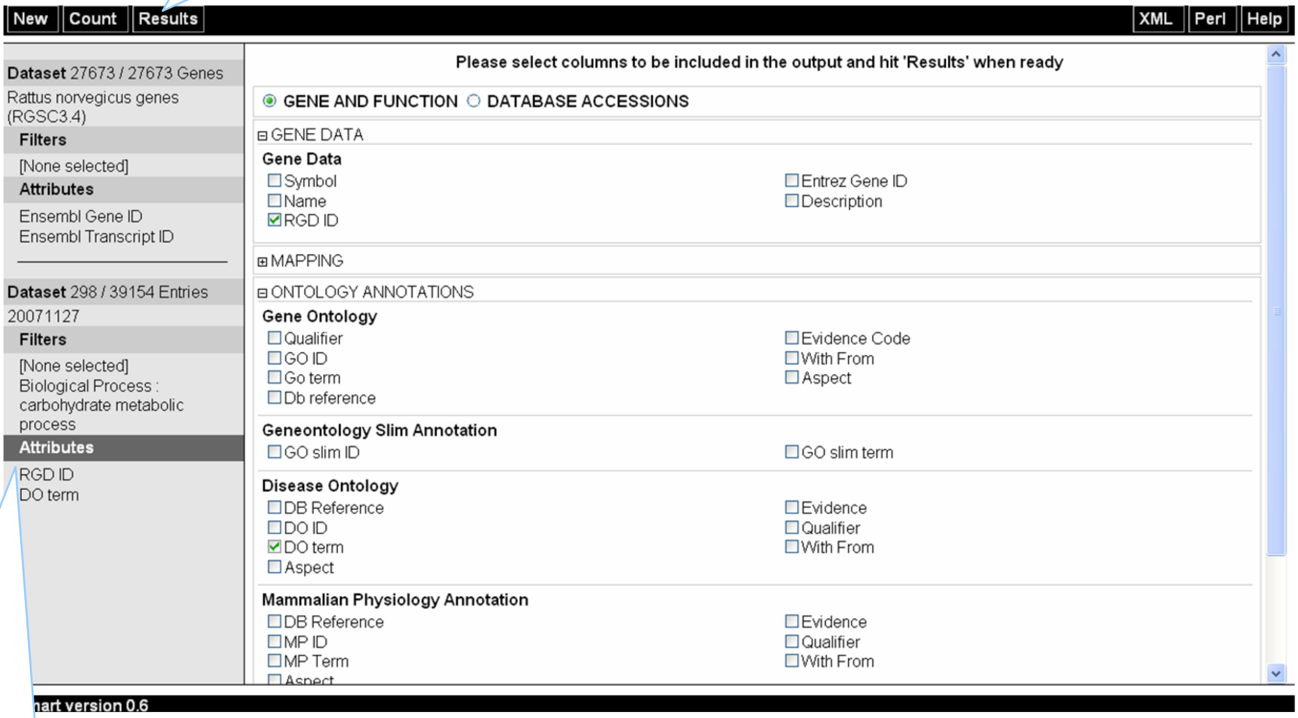
● GENE AND FUNCTION ○ DATABASE ACCESSIONS

⊞ GENE DATA
⊞ MAPPING
⊞ ONTOLOGY ANNOTATIONS

STEP 9:
Expand the ‘ONTOLOGY ANNOTATIONS’ to select **Disease Ontology**.
(**DO term**)

STEP 8:
Expand the ‘GENE DATA’ panel, and select ‘**RGD ID**’.

STEP 10:
Click **RESULTS** at the top to preview the output



Dataset 27673 / 27673 Genes
Rattus norvegicus genes (RGSC3.4)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset 298 / 39154 Entries
20071127
Filters
[None selected]
Biological Process : carbohydrate metabolic process
Attributes
RGD ID
DO term

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

● **GENE AND FUNCTION** ○ **DATABASE ACCESSIONS**

⊟ GENE DATA

Gene Data

□ Symbol	□ Entrez Gene ID
□ Name	□ Description
☑ RGD ID	

⊞ MAPPING

⊟ ONTOLOGY ANNOTATIONS

Gene Ontology

□ Qualifier	□ Evidence Code
□ GO ID	□ With From
□ Go term	□ Aspect
□ Db reference	

Geneontology Slim Annotation

□ GO slim ID	□ GO slim term

Disease Ontology

□ DB Reference	□ Evidence
□ DO ID	□ Qualifier
☑ DO term	□ With From
□ Aspect	

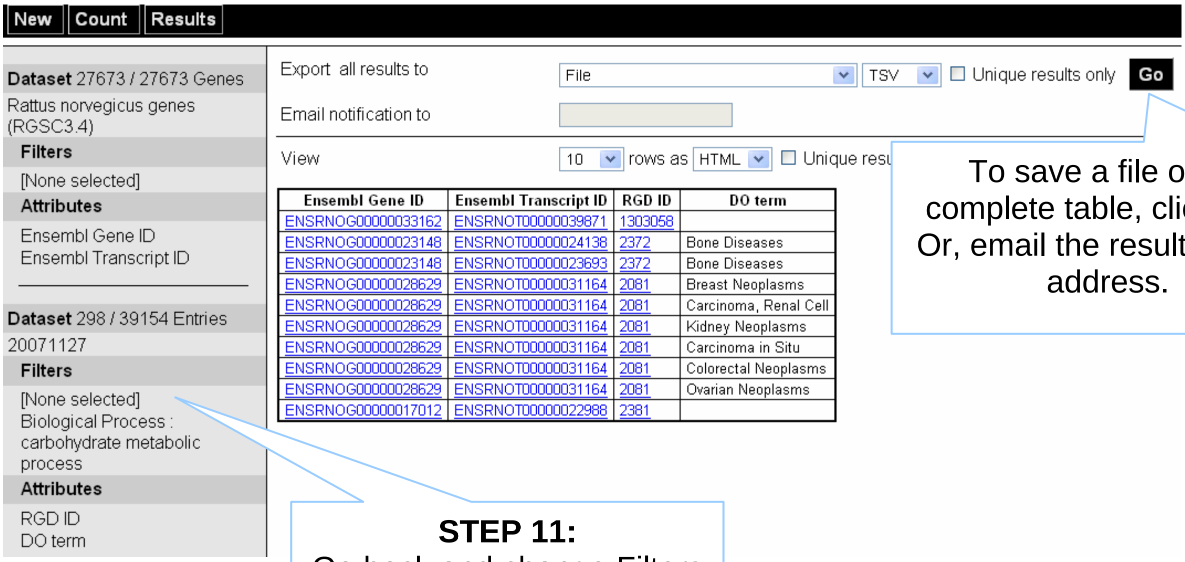
Mammalian Physiology Annotation

□ DB Reference	□ Evidence
□ MP ID	□ Qualifier
□ MP Term	□ With From
□ Aspect	

nart version 0.6

Note the summary of selected options.

The order of attributes determines the order of columns in the result table.



Dataset 27673 / 27673 Genes

Rattus norvegicus genes (RGSC3.4)

Filters

[None selected]

Attributes

Ensembl Gene ID
Ensembl Transcript ID

Dataset 298 / 39154 Entries

20071127

Filters

[None selected]
Biological Process : carbohydrate metabolic process

Attributes

RGD ID
DO term

Export all results to: File, TSV, □ Unique results only, Go

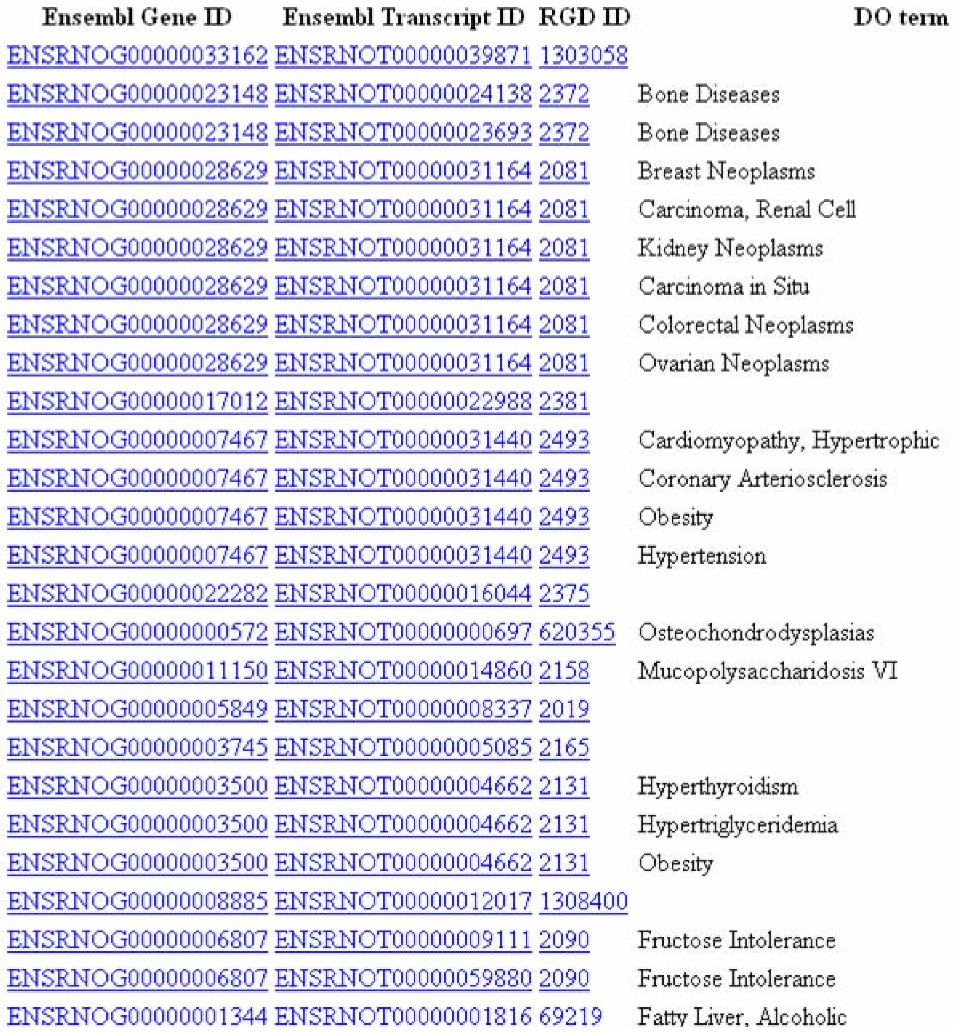
Email notification to

View: 10 rows as HTML, □ Unique resu

Ensembl Gene ID	Ensembl Transcript ID	RGD ID	DO term
ENSRNOG00000033162	ENSRNOT00000039871	1303058	
ENSRNOG00000023148	ENSRNOT00000024138	2372	Bone Diseases
ENSRNOG00000023148	ENSRNOT00000023693	2372	Bone Diseases
ENSRNOG00000028629	ENSRNOT00000031164	2081	Breast Neoplasms
ENSRNOG00000028629	ENSRNOT00000031164	2081	Carcinoma, Renal Cell
ENSRNOG00000028629	ENSRNOT00000031164	2081	Kidney Neoplasms
ENSRNOG00000028629	ENSRNOT00000031164	2081	Carcinoma in Situ
ENSRNOG00000028629	ENSRNOT00000031164	2081	Colorectal Neoplasms
ENSRNOG00000028629	ENSRNOT00000031164	2081	Ovarian Neoplasms
ENSRNOG00000017012	ENSRNOT00000022988	2381	

To save a file of the complete table, click ‘**Go**’. Or, email the results to any address.

STEP 11:
Go back and change Filters or Attributes if desired.
Or, **View ‘ALL’ as HTML**…



Ensembl Gene ID	Ensembl Transcript ID	RGD ID	DO term
ENSRNOG00000033162	ENSRNOT00000039871	1303058	
ENSRNOG00000023148	ENSRNOT00000024138	2372	Bone Diseases
ENSRNOG00000023148	ENSRNOT00000023693	2372	Bone Diseases
ENSRNOG00000028629	ENSRNOT00000031164	2081	Breast Neoplasms
ENSRNOG00000028629	ENSRNOT00000031164	2081	Carcinoma, Renal Cell
ENSRNOG00000028629	ENSRNOT00000031164	2081	Kidney Neoplasms
ENSRNOG00000028629	ENSRNOT00000031164	2081	Carcinoma in Situ
ENSRNOG00000028629	ENSRNOT00000031164	2081	Colorectal Neoplasms
ENSRNOG00000028629	ENSRNOT00000031164	2081	Ovarian Neoplasms
ENSRNOG00000017012	ENSRNOT00000022988	2381	
ENSRNOG00000007467	ENSRNOT00000031440	2493	Cardiomyopathy, Hypertrophic
ENSRNOG00000007467	ENSRNOT00000031440	2493	Coronary Arteriosclerosis
ENSRNOG00000007467	ENSRNOT00000031440	2493	Obesity
ENSRNOG00000007467	ENSRNOT00000031440	2493	Hypertension
ENSRNOG00000022282	ENSRNOT00000016044	2375	
ENSRNOG00000000572	ENSRNOT00000000697	620355	Osteochondrodysplasias
ENSRNOG00000011150	ENSRNOT00000014860	2158	Mucopolysaccharidosis VI
ENSRNOG00000005849	ENSRNOT00000008337	2019	
ENSRNOG00000003745	ENSRNOT00000005085	2165	
ENSRNOG00000003500	ENSRNOT00000004662	2131	Hyperthyroidism
ENSRNOG00000003500	ENSRNOT00000004662	2131	Hypertriglyceridemia
ENSRNOG00000003500	ENSRNOT00000004662	2131	Obesity
ENSRNOG00000008885	ENSRNOT00000012017	1308400	
ENSRNOG00000006807	ENSRNOT00000009111	2090	Fructose Intolerance
ENSRNOG00000006807	ENSRNOT00000059880	2090	Fructose Intolerance
ENSRNOG00000001344	ENSRNOT00000001816	69219	Fatty Liver, Alcoholic

END OF WORKED EXAMPLE