



# **BioMart Central Portal: An Open Data base Network for the Biological Community**

August 19, 2011

## Data management Challenge I: Data

- Large
- Heterogeneous
- Distributed
- Disconnected

## Data management Challenge II: Software

Lack of “off the shelf” solutions that have/are:

- Interactive querying interfaces
- Broadly applicable
- Scalable
- Secure
- Support data federation

## What is BioMart?



Free, open-source federated data management system that makes it possible to make distributed biological data accessible to the research community through a unified user interface

BioMart system is data agnostic and platform independent, and is adopted by dozens of public and private databases & international consortia to manage many different types of biological data

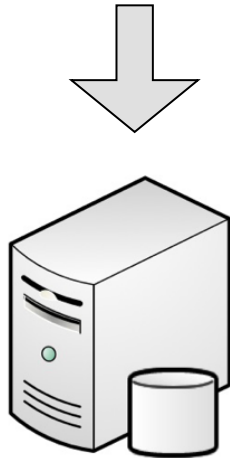
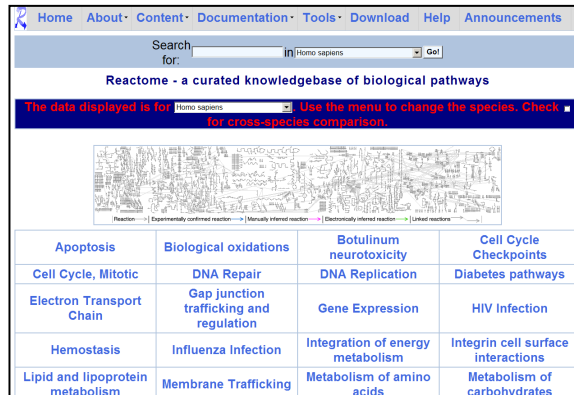
### **Key features:**

- Easy installation
- Supports data federation
- Optimized for large-scale data retrieval
- Variety of graphical and programmatic query interfaces
- Enterprise level security features to manage sensitive data
- Robust plug-in framework for data analysis and visualization

# What is BioMart?

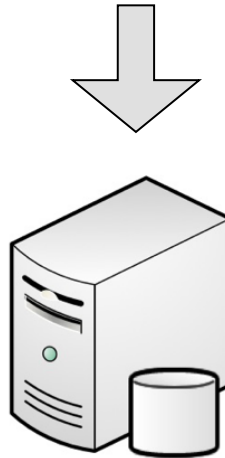


## Reactome



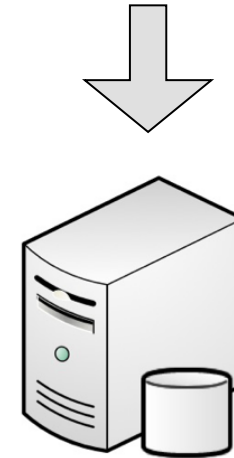
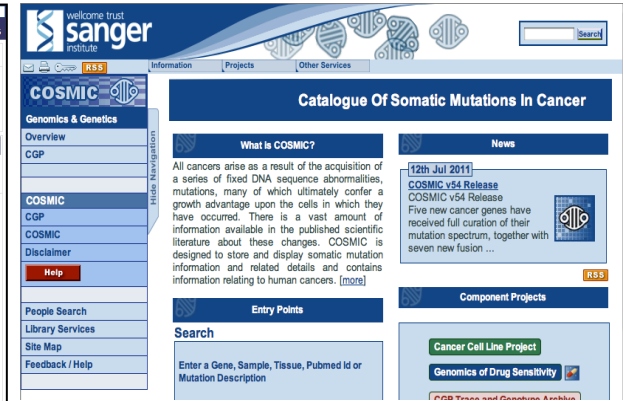
Reactome Mart

## Ensembl



Ensembl Mart

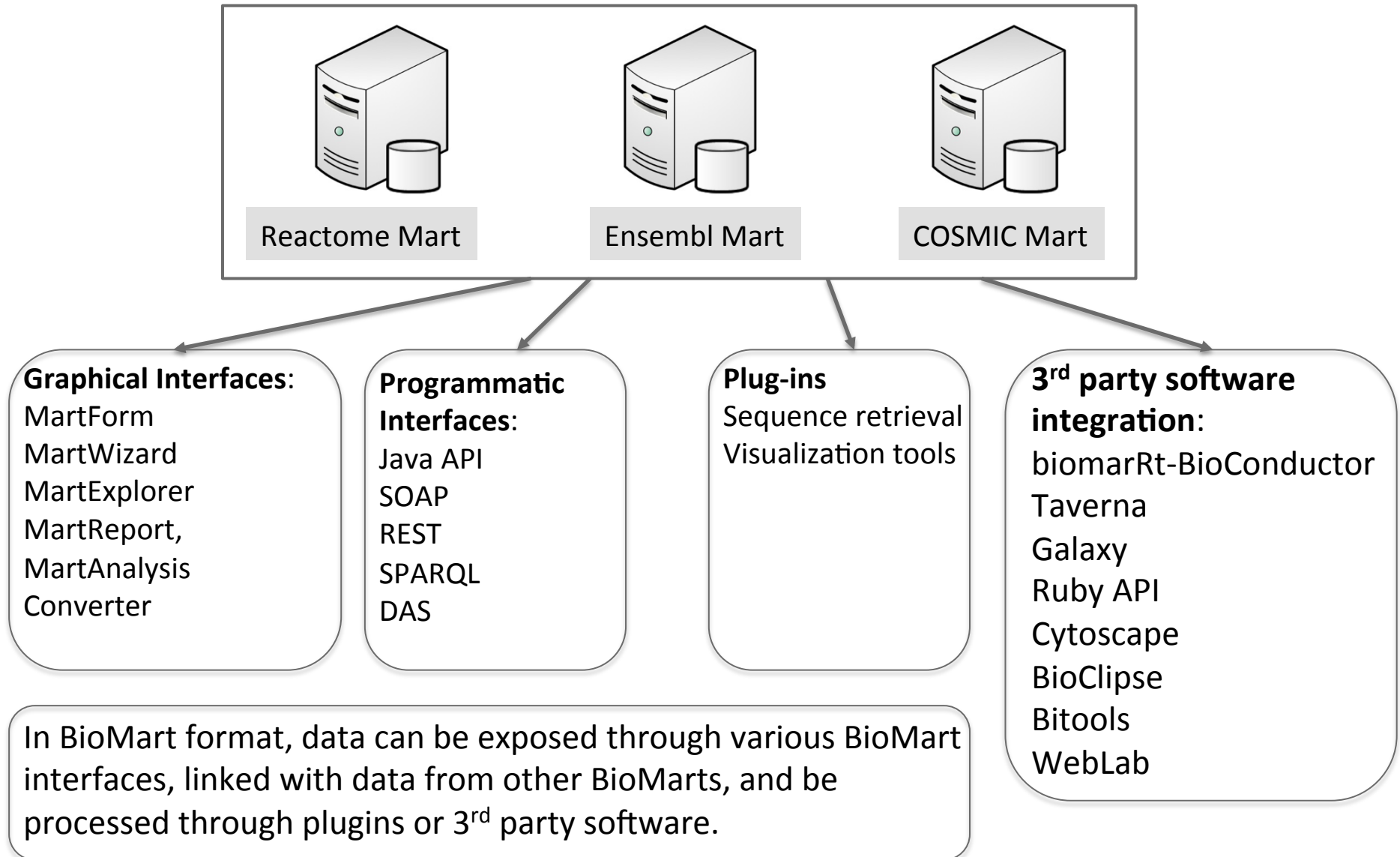
## COSMIC



COSMIC Mart

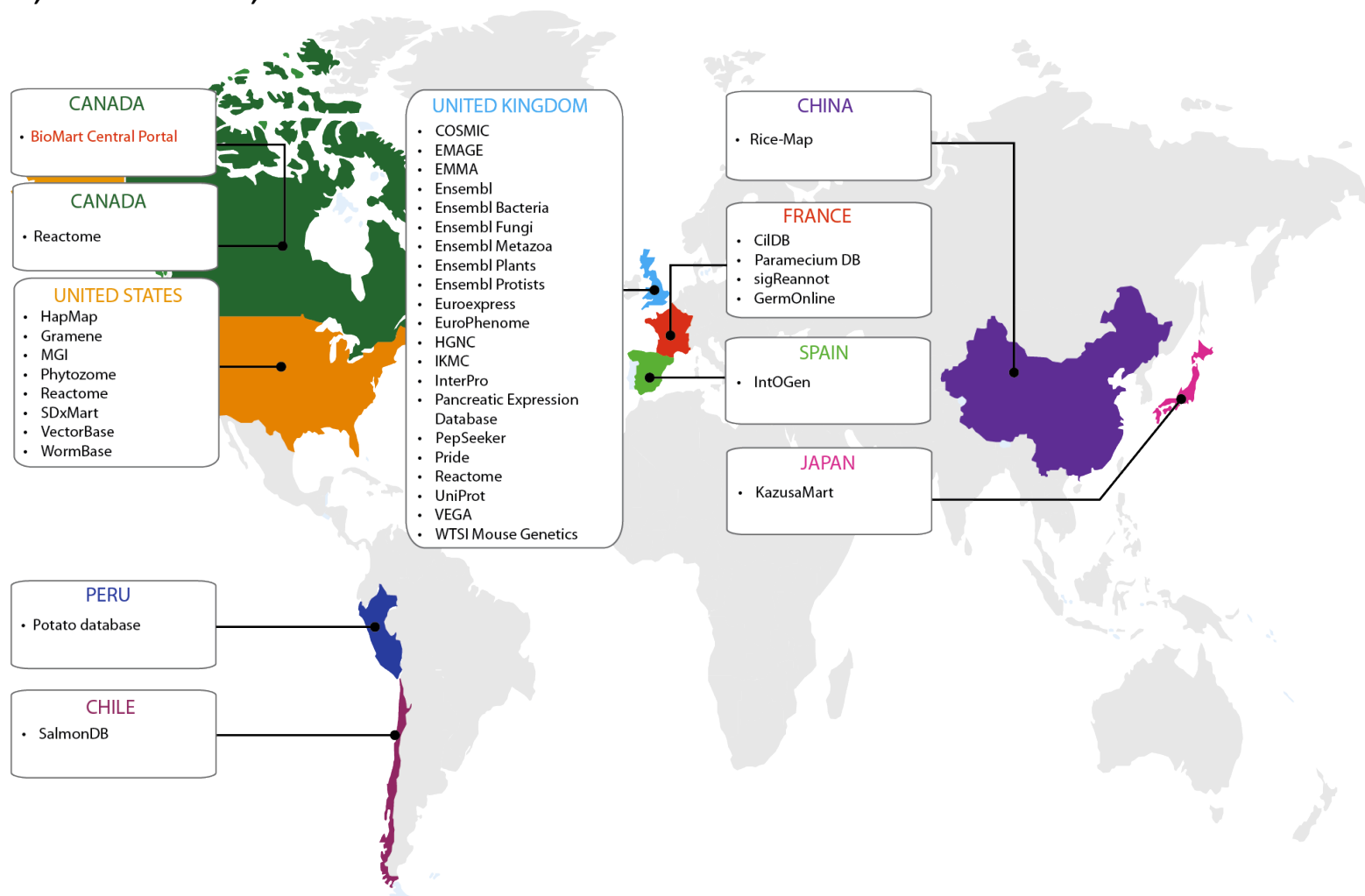
Data from individual data sources is transformed to the BioMart (warehouse) format

## What is BioMart?



## BioMart Central Portal ([central.biomart.org](http://central.biomart.org))

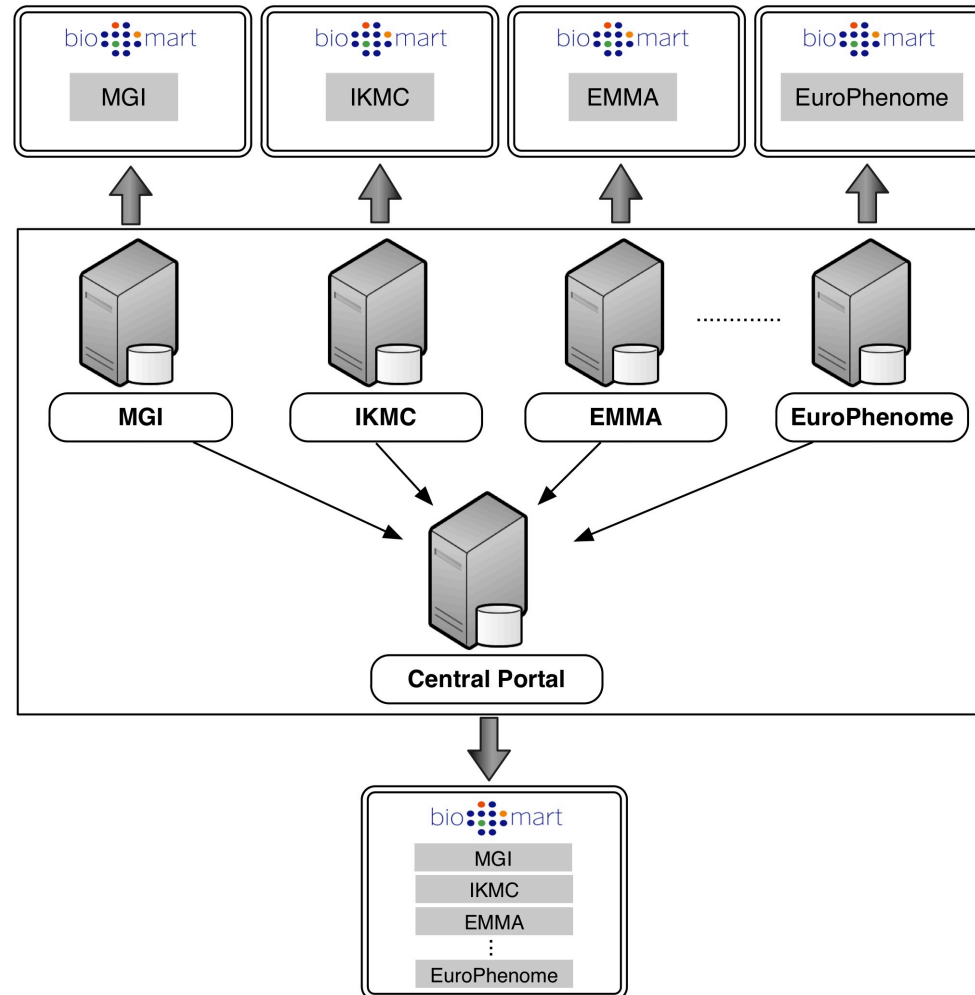
A unified access to dozens of biological databases spanning genomics, proteomics, model organisms, cancer data, and more



## BioMart Central Portal

- First-of-its-kind-community-driven effort to make data from dozens of bio-medical databases accessible to the broad scientific community
- Users can query data from one data source, or integrate data from multiple sources
- Provides a rich annotation source that can be integrated with in-house data
- Anybody can include their data source in the Central Portal
- All databases are maintained independently by data providers, allowing 'dynamic' data sharing and coordination without the need for data aggregation
- Several graphical and programmatic interfaces provide a range of query options

## BioMart Central Portal Architecture



BioMart Central Portal is configured in a Master/Slave like architecture, where each of the individual BioMart servers present only their own data sources, while a single “master” server acts as a portal providing a unified view over all the sources.



# BioMart Central Portal – Home page (central.biomart.org)



### IDENTIFIER SEARCH

Examples: KRAS, ENSG00000146648

Type in your gene identifier and search for it across member databases

### TOOLS

**Gene retrieval** Variant retrieval Sequence retrieval ID Converter

Cancer Genes

Ensembl

Ensembl bacteria

Ensembl fungi

Ensembl metazoa

Ensembl plants

Ensembl protists

Mouse Genome Informatics (MGI)

VEGA

Several tools simplify the most popular queries (e.g. converting between gene identifiers; retrieving genes, variants, or genomic sequence)

### DATABASE SEARCH

**Search by type** Search by organism

▸ Genome

▸ Gene annotation

▸ Protein sequence and structure

▸ Interaction and pathways

▸ Gene expression

▸ Cancer

▸ Model organism databases

Query member databases. Find your database of interest by type or organism

### BioMart CENTRAL PORTAL


Databases: 37

Click on the map to see the list of member databases

# Querying BioMart Central Portal

All BioMart Central Portal queries are performed using the following simple steps:

## 1. Selecting Database



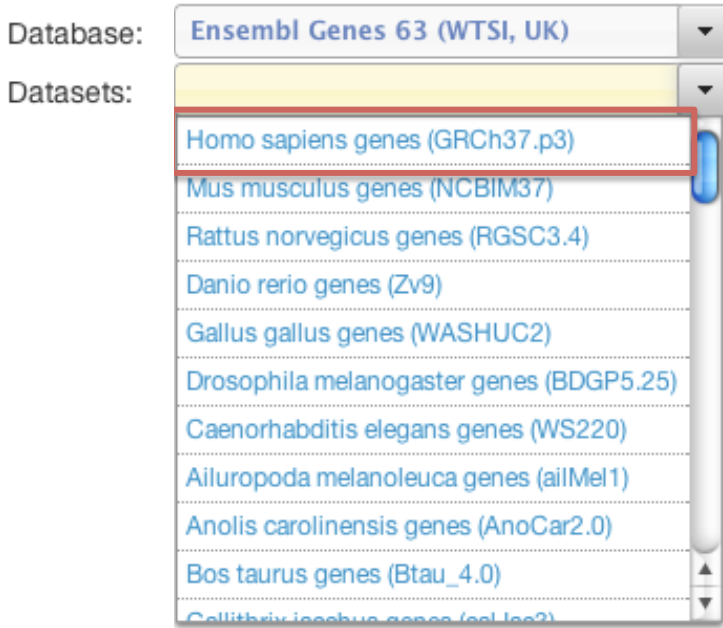
**DATABASE SEARCH**

Search by type | Search by organism

▼ Genome

- Ensembl Genes 63 (WTSI, UK)**
- Ensembl Variation 63 (WTSI, UK)
- Ensembl Regulation 63 (WTSI, UK)
- Ensembl Bacteria 10 (EBI, UK)
- Ensembl Fungi 10 (EBI, UK)
- Ensembl Fungi Variations 10 (EBI, UK)
- Ensembl Metazoa 10 (EBI, UK)
- Ensembl Metazoa Variations 10 (EBI, UK)
- Ensembl Plants 10 (EBI, UK)
- Ensembl Plants Variations 10 (EBI, UK)
- Ensembl Protists 10 (EBI, UK)
- Ensembl Protists Variations 10 (EBI, UK)
- HapMap 27 (NCBI, USA)
- Vertebrate Genome Association (VEGA) 43 (WTSI, UK)

## 2. Selecting Dataset



Database: **Ensembl Genes 63 (WTSI, UK)**

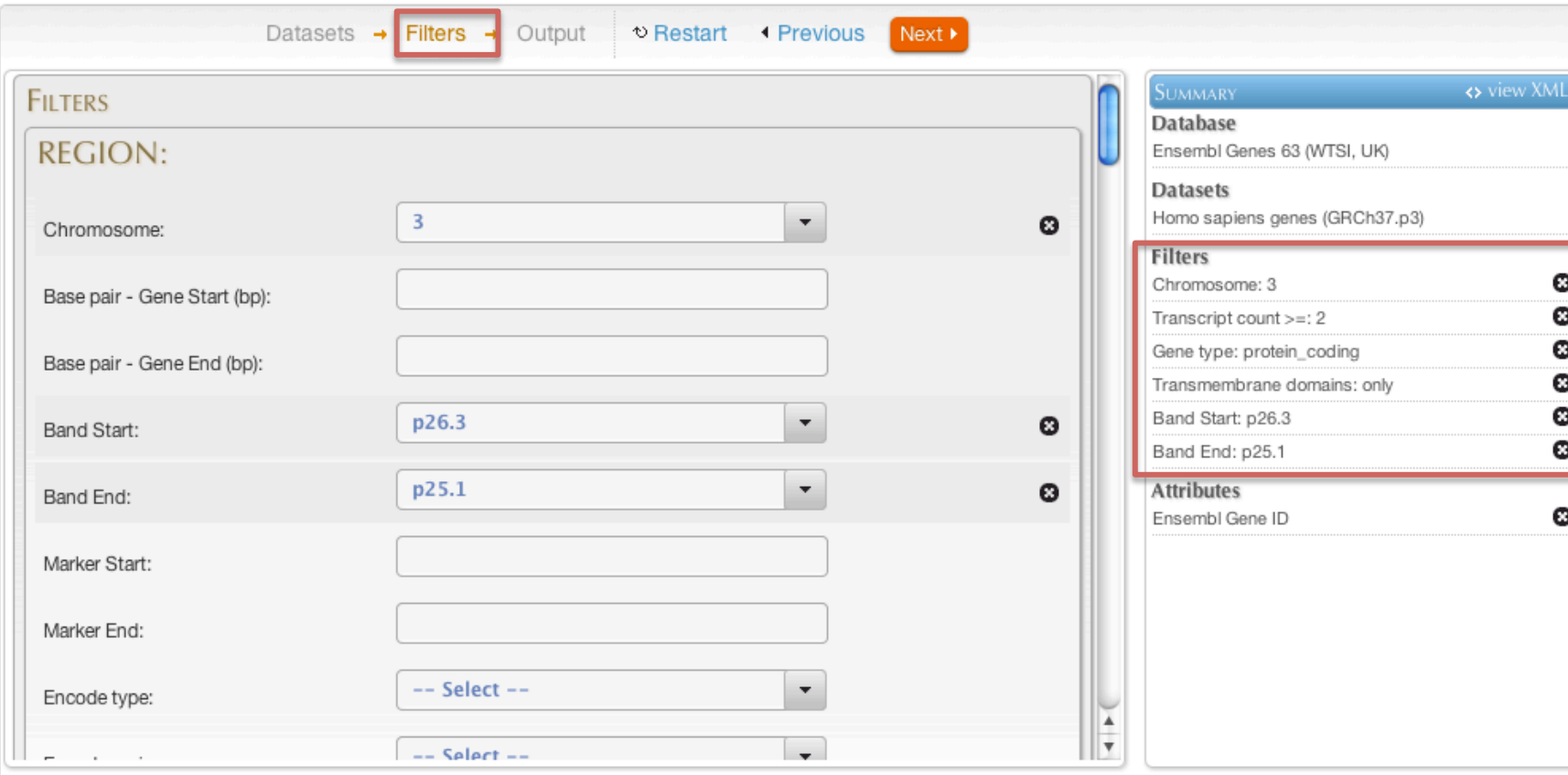
Datasets:

- Homo sapiens genes (GRCh37.p3)**
- Mus musculus genes (NCBIM37)
- Rattus norvegicus genes (RGSC3.4)
- Danio rerio genes (Zv9)
- Gallus gallus genes (WASHUC2)
- Drosophila melanogaster genes (BDGP5.25)
- Caenorhabditis elegans genes (WS220)
- Ailuropoda melanoleuca genes (ailMel1)
- Anolis carolinensis genes (AnoCar2.0)
- Bos taurus genes (Btau\_4.0)
- Callithrix jacchus genes (celJac2)

## Querying BioMart Central Portal

All BioMart queries are performed using the following simple steps:

### 3. Selecting Filters (Optional)



The screenshot displays the BioMart Central Portal interface during the 'Filters' step. The top navigation bar includes 'Datasets', 'Filters' (highlighted), and 'Output', along with 'Restart', 'Previous', and 'Next' buttons. The main panel is titled 'FILTERS' and contains a 'REGION' section with the following filters:

- Chromosome: 3
- Base pair - Gene Start (bp):
- Base pair - Gene End (bp):
- Band Start: p26.3
- Band End: p25.1
- Marker Start:
- Marker End:
- Encode type: -- Select --

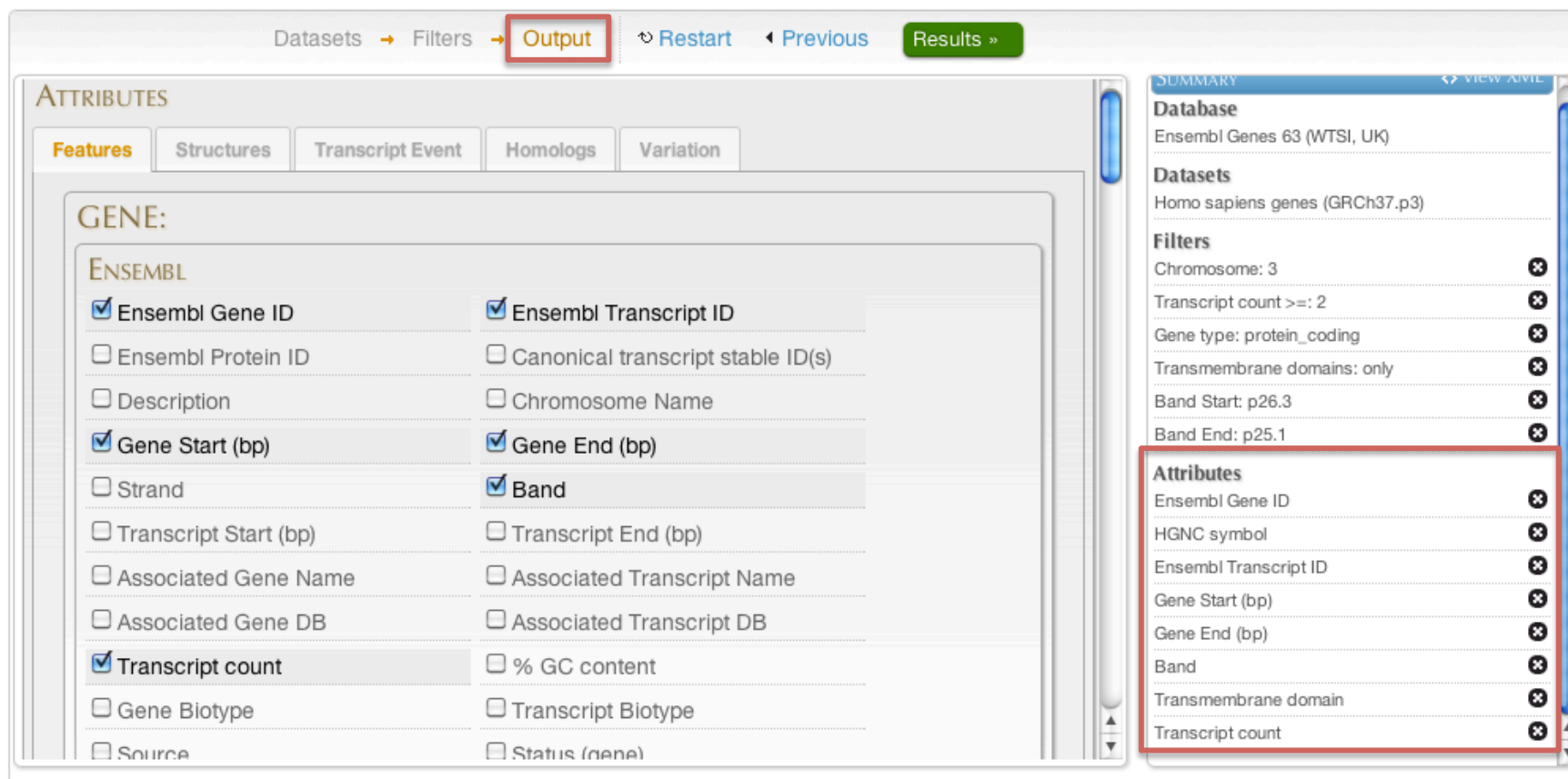
On the right, a 'SUMMARY' panel shows the query details:

- Database:** Ensembl Genes 63 (WTSI, UK)
- Datasets:** Homo sapiens genes (GRCh37.p3)
- Filters:**
  - Chromosome: 3
  - Transcript count >= 2
  - Gene type: protein\_coding
  - Transmembrane domains: only
  - Band Start: p26.3
  - Band End: p25.1
- Attributes:** Ensembl Gene ID

## Querying BioMart Central Portal

All BioMart queries are performed using the following simple steps:

### 4. Selecting Attributes



The screenshot shows the BioMart Central Portal interface. At the top, there are navigation tabs: 'Datasets', 'Filters', and 'Output' (which is highlighted with a red box). Below these are buttons for 'Restart', 'Previous', and 'Results'. The main area is titled 'ATTRIBUTES' and has sub-tabs for 'Features', 'Structures', 'Transcript Event', 'Homologs', and 'Variation'. The 'Features' tab is selected, showing a list of attributes under the 'GENE:' section. The 'ENSEMBL' section is expanded, showing a list of attributes with checkboxes. The 'Output' tab is also highlighted with a red box, showing a list of attributes with checkboxes. The 'Attributes' section is expanded, showing a list of attributes with checkboxes. The 'Attributes' section is highlighted with a red box, showing a list of attributes with checkboxes.

Note: The order of attributes determines the order of the column in the results. Drag and drop attributes to reorder; click on the X to remove filters or attributes.

# Querying the BioMart Central Portal

## Results Page

Ensembl Genes 63 (WTSI, UK)  
Displaying results 1-20 out of 104

Bookmark | REST / SOAP | SPARQL | Java | Download data | Back

See how many records are returned

Sort results by clicking on the black triangle toggle icon

Queries can be bookmarked to be reused or shared

Get query syntax in any of API formats available (e.g. next slide)

Download data in tab-delimited tabular format

Ensembl Gene ID	Symbol	Ensembl Transcript ID	Gene Start (bp)	Gene End (bp)	Band	Transcript count
ENSG00000206561	COLQ	ENST00000206561	15491640	15563258	p25.3	11
ENSG00000171135	JAGN1	ENST00000171135	9932238	9936033	p25.3	3
ENSG00000163701	IL17RE	ENST00000163701	9944296	9958086	p25.3	13
ENSG00000163702	IL17RC	ENST00000163702	9958758	9975314	p25.3	30

1 2 Next »

Powered by bioMart

## Central Portal offers programmatic access for automated querying



XML querying via REST or SOAP request, full Java API, and RDF querying via SPARQL. Queries constructed in the web GUI can be converted to any of the API formats by clicking on the appropriate button on the results page; in this way, queries can be saved, modified and easily transferred from one format to another.

Ensembl Genes 63 (100,000)

Displaying results 1-20 out of 100,000

Ensembl Gene ID
ENSG00000206561
ENSG00000206561
ENSG00000206561
ENSG00000171135
ENSG00000171135
ENSG00000163701
ENSG00000163701
ENSG00000163701
ENSG00000163701
ENSG00000163702
ENSG00000163702
ENSG00000163702
ENSG00000163702
ENSG00000163702
ENSG00000163702
ENSG00000163702
ENSG00000163702

REST / API Query

```
<!DOCTYPE Query>
<Query client="true" processor="TSV" limit="-1" header="1">
  <Dataset name="hsapiens_gene_ensembl" config="gene_ensembl_config">
    <Filter name="chromosome_name" value="3"/>
    <Filter name="band_start" value="p26.3"/>
    <Filter name="band_end" value="p26.1"/>
    <Filter name="transcript_count" value="2"/>
    <Filter name="biotype" value="protein_coding"/>
    <Filter name="with_transmembrane_domain" value="only"/>
    <Attribute name="ensembl_gene_id"/>
    <Attribute name="hgnc_symbol"/>
    <Attribute name="ensembl_transcript_id"/>
    <Attribute name="chromosome_name"/>
    <Attribute name="start_position"/>
    <Attribute name="end_position"/>
    <Attribute name="band"/>
  </Dataset>
</Query>
```

Toggle quote-escape

Close

Back

Transcript count
11
11
11
3
3
13
13
13
13
30
30
30
30
30
30
30

## BioMart Central Portal allows cross-dataset querying



- Cross-dataset query takes advantage of **inter-linked** public BioMart databases to get answers for questions that would not be possible if data sources were disconnected

Query #1: 'Find deletion mutations in the COSMIC database that affect genes involved in Apoptosis'

Step 1. Select Cancer genes under TOOLS/Gene retrieval

The screenshot shows the 'TOOLS' section of the BioMart interface. The 'Gene retrieval' tab is selected, and a dropdown menu is open, listing several databases. 'Cancer genes' is highlighted with a red border. The other databases listed are Ensembl, Ensembl Bacteria, Ensembl Fungi, Ensembl Metazoa, Ensembl Plants, Ensembl Protists, Mouse Genome Informatics, and VEGA.

TOOLS			
Gene retrieval	Variant retrieval	Sequence retrieval	ID converter
<div>Cancer genes</div> <div>Ensembl</div> <div>Ensembl Bacteria</div> <div>Ensembl Fungi</div> <div>Ensembl Metazoa</div> <div>Ensembl Plants</div> <div>Ensembl Protists</div> <div>Mouse Genome Informatics</div> <div>VEGA</div>			

Query #1: 'Find deletion mutations in the COSMIC database that affect genes involved in Apoptosis

Step 2. Select pathway filter from KEGG, and mutation filter from COSMIC

VIEW: Cancer Genes

**1. SELECT DATASETS**  
Select multiple databases by holding down the Ctrl key, or ⌘ on a Mac.

COSMIC

**2. RESTRICT SEARCH**

KEGG Pathway: Apoptosis

Gene accessions: -- Select --

upload file

Mutation type - CDS: Deletion

Mutation type - AA: -- Select --

Go »

Powered by bio:mart

KEGG

COSMIC

By linking the COSMIC and KEGG databases, we ask for only genes in both datasets (intersection of the two datasets)



# Query results

Cancer Genes » COSMIC

Displaying results 121-140 out of 883

[Bookmark](#)
[REST / SOAP](#)
[SPARQL](#)
[Java](#)
[Download data](#)
[Back](#)

Gene symbol	Pathway title	Ensembl gene ID	Entrez gene ID	Swissprot ID	COSMIC mutation ID	Mutation type - CDS	Mutation type - AA	Primary sites	Number of samples
TP53	Apoptosis	ENSG00000141510	7157	P04637	13416	Deletion	Deletion - In frame	upper_aerodigestive_tract	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	44688	Deletion	Deletion - Frameshift	breast	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	44231	Deletion	Deletion - Frameshift	large_intestine / upper_aerodigestive_tract	2
TP53	Apoptosis	ENSG00000141510	7157	P04637	18610	Deletion	Deletion - Frameshift	endometrium / ovary	2
TP53	Apoptosis	ENSG00000141510	7157	P04637	46093	Deletion	Deletion - In frame	lung	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	45370	Deletion	Deletion - Frameshift	lung	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	45579	Deletion	Deletion - Frameshift	upper_aerodigestive_tract	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	45564	Deletion	Deletion - In frame	lung	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	44651	Deletion	Deletion - Frameshift	liver	2
TP53	Apoptosis	ENSG00000141510	7157	P04637	44314	Deletion	Deletion - In frame	skin	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	44913	Deletion	Deletion - Frameshift	adrenal_gland	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	44134	Deletion	Deletion - Frameshift	ovary / skin / soft_tissue	3
TP53	Apoptosis	ENSG00000141510	7157	P04637	69085	Deletion	Deletion - Frameshift	ovary	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	44871	Deletion	Deletion - Frameshift	large_intestine / ovary	2
TP53	Apoptosis	ENSG00000141510	7157	P04637	53256	Deletion	Deletion - Frameshift	large_intestine	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	46238	Deletion	Deletion - Frameshift	upper_aerodigestive_tract	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	45723	Deletion	Deletion - In frame	breast	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	18597	Deletion	Deletion - Frameshift	central_nervous_system	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	44862	Deletion	Complex - deletion inframe	breast	1
TP53	Apoptosis	ENSG00000141510	7157	P04637	45039	Deletion	Deletion - Frameshift	breast	1

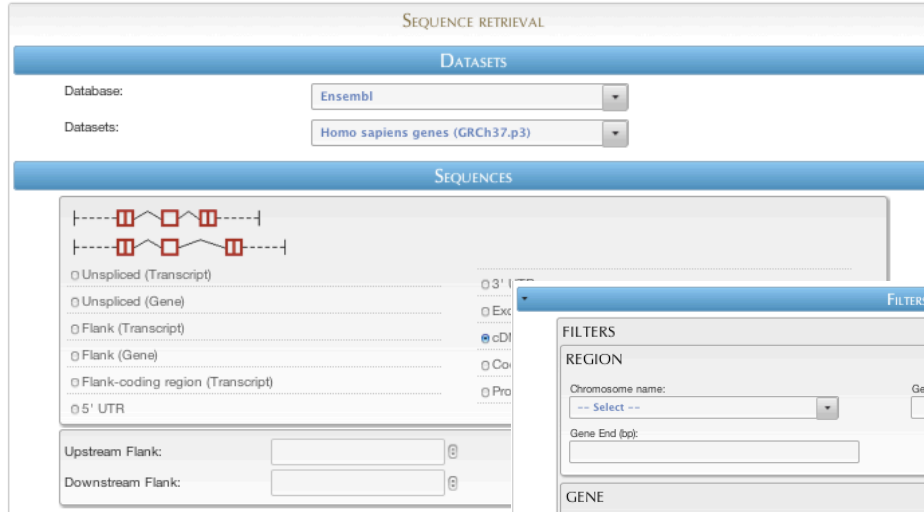
« Previous 3 4 5 6 7 8 9 10 11 12 Next »

Powered by bio::mart

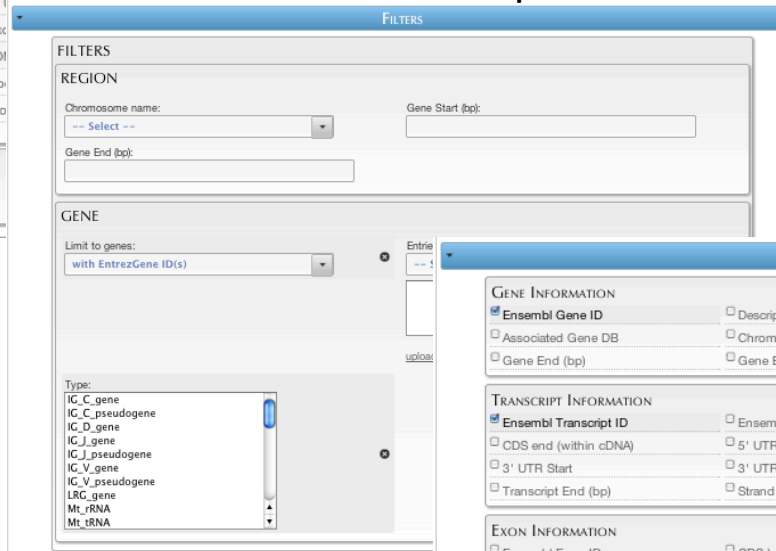
## BioMart offers several plugins, e.g. TOOLS-Sequence retrieval

Query #2: Retrieve cDNA sequences of protein coding human genes that have EntrezGene ID

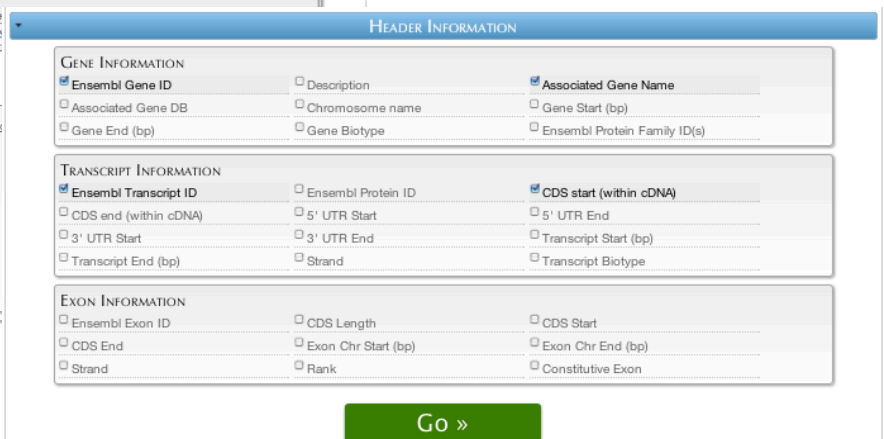
### Step 1: Select sequence type (e.g. cDNA)



### Step 2: Select filters



### Step 3: Select header info



*Sequence retrieval tool is implemented as server-side analysis plugin*

## Conclusions

- BioMart Central Portal provide access to over 30 biological databases spanning genomics, proteomics, model organisms, gene expression, cancer, and more.
- Data federation increases scalability and flexibility, and enables cross-dataset querying
- Query results are retrieved directly from the data source, hence the results are always up-to-date
- BioMart Central Portal has both “biologist-friendly” and “bioinformatician-friendly” query interfaces enabling a range of query options

## Acknowledgements

Joachim Baran  
Anthony Cros  
Jonathan Guberman  
Syed Haider  
Jack Hsu  
Yong Liang  
Long Yao  
Elena Rivkin  
Brett Whitty  
Marie Wong-Erasmus  
Zhang Junjun  
Arek Kasprzyk

*Website: [www.biomart.org](http://www.biomart.org)  
Mailing list: [users@biomart.org](mailto:users@biomart.org)*



MINISTRY OF RESEARCH AND INNOVATION