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

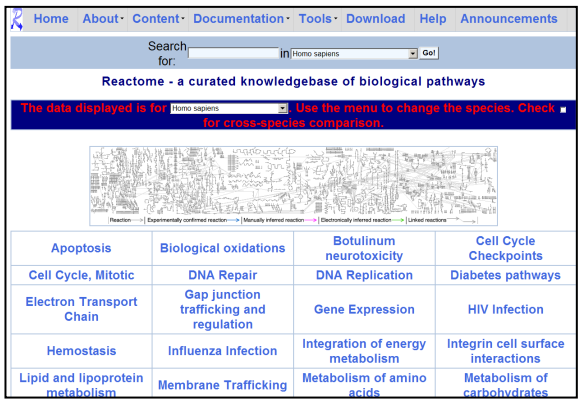
BioMart system is data agnostic and platform independent, and is adopted by dozens of public and private databases & international consortia to manage many different types of biological data

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

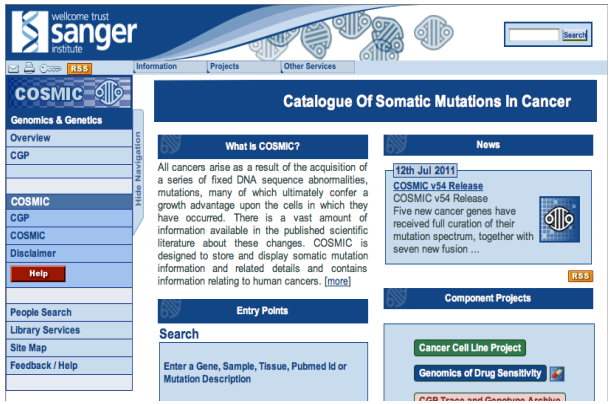
What is BioMart?

Reactome



Ensembl

COSMIC



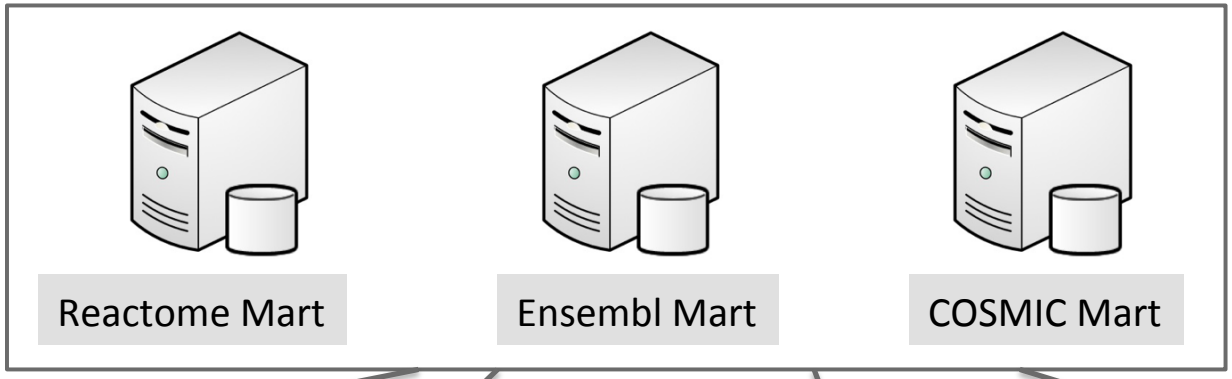
Reactome Mart

Ensembl Mart

COSMIC Mart

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

What is BioMart?



Graphical Interfaces:
MartForm
MartWizard
MartExplorer
MartReport,
MartAnalysis
Converter

Programmatic Interfaces:
Java API
SOAP
REST
SPARQL
DAS

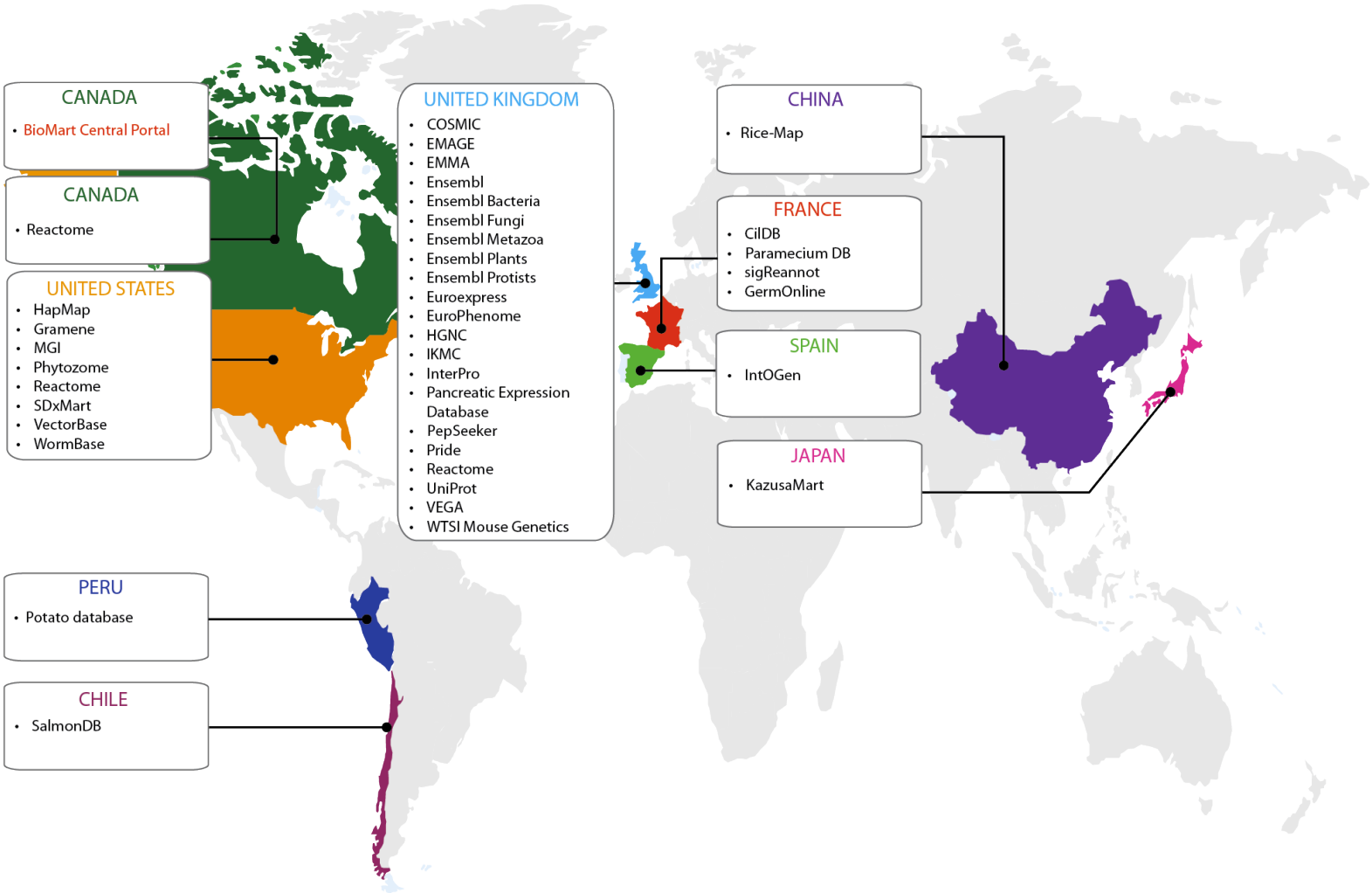
Plug-ins
Sequence retrieval
Visualization tools

3rd party software integration:
biomarRt-BioConductor
Taverna
Galaxy
Ruby API
Cytoscape
BioClipse
Bitools
WebLab

In BioMart format, data can be exposed through various BioMart interfaces, linked with data from other BioMarts, and be processed through plugins or 3rd party software.

BioMart Central Portal (central.biomart.org)

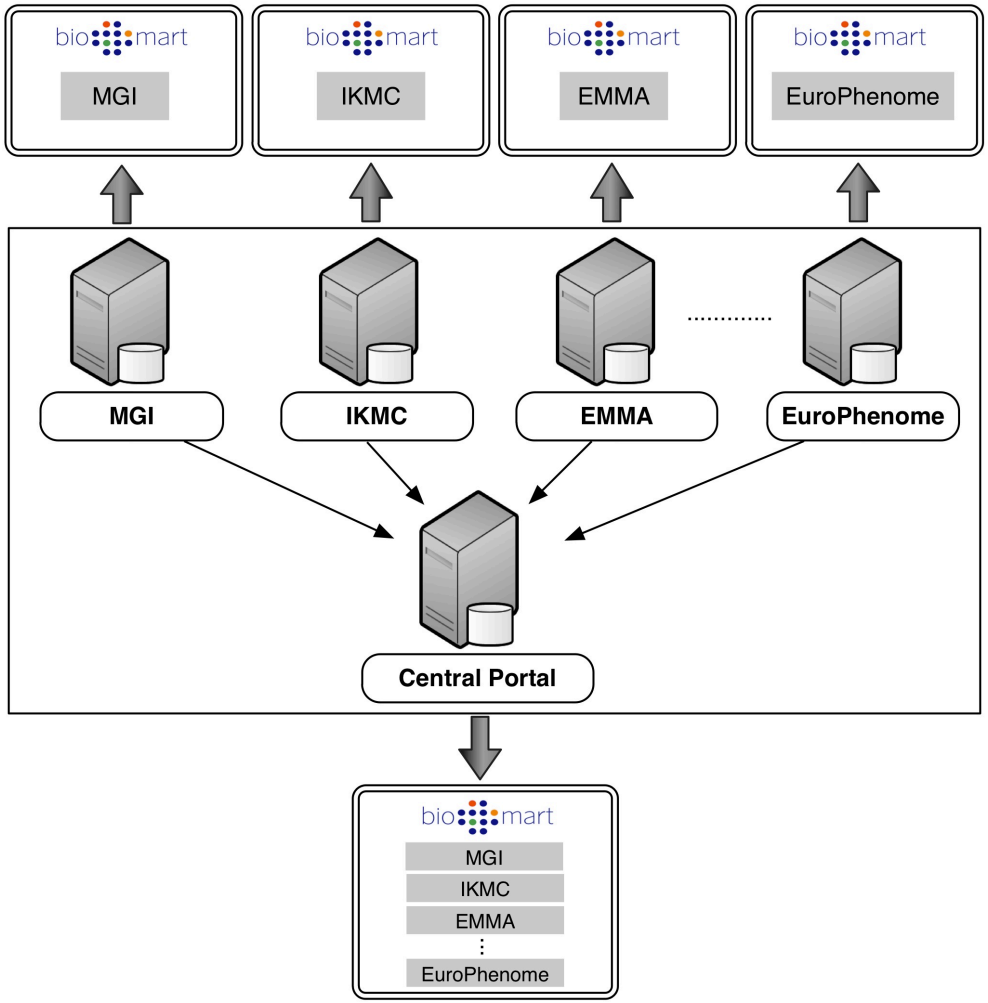
A unified access to dozens of biological databases spanning genomics, proteomics, model organisms, cancer data, and more



BioMart Central Portal

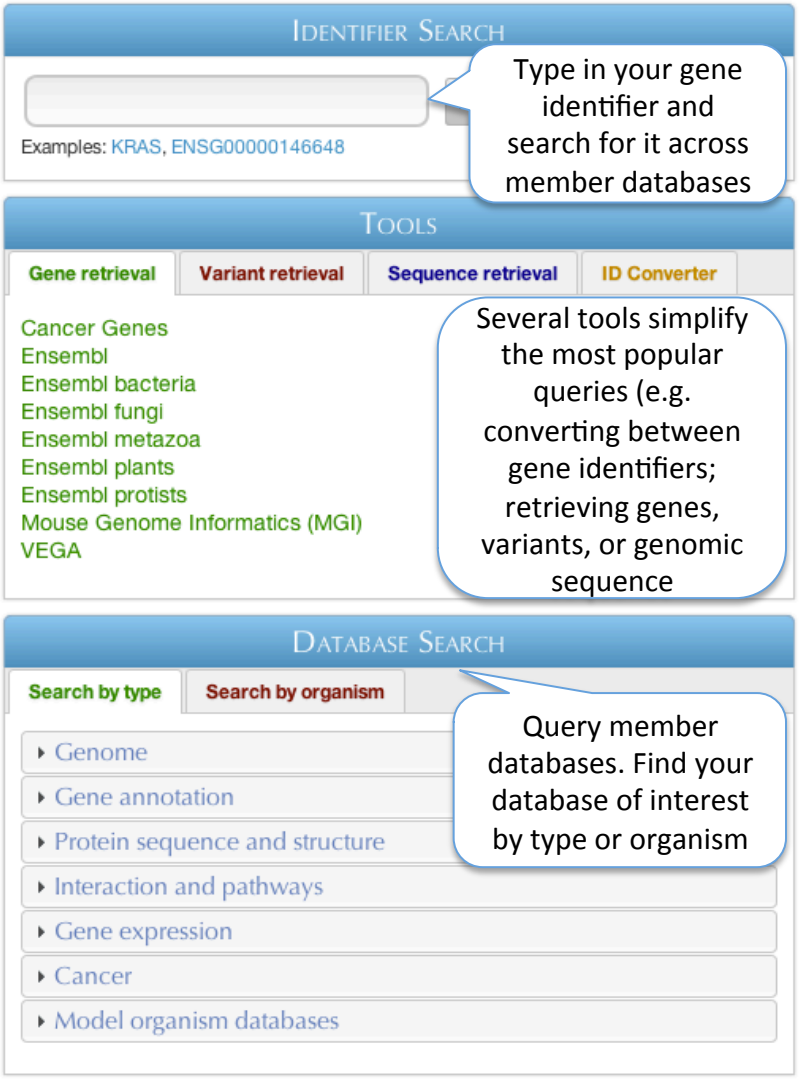
- First-of-its-kind-community-driven effort to make data from dozens of bio-medical databases accessible to the broad scientific community
- Users can query data from one data source, or integrate data from multiple sources
- Provides a rich annotation source that can be integrated with in-house data
- 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

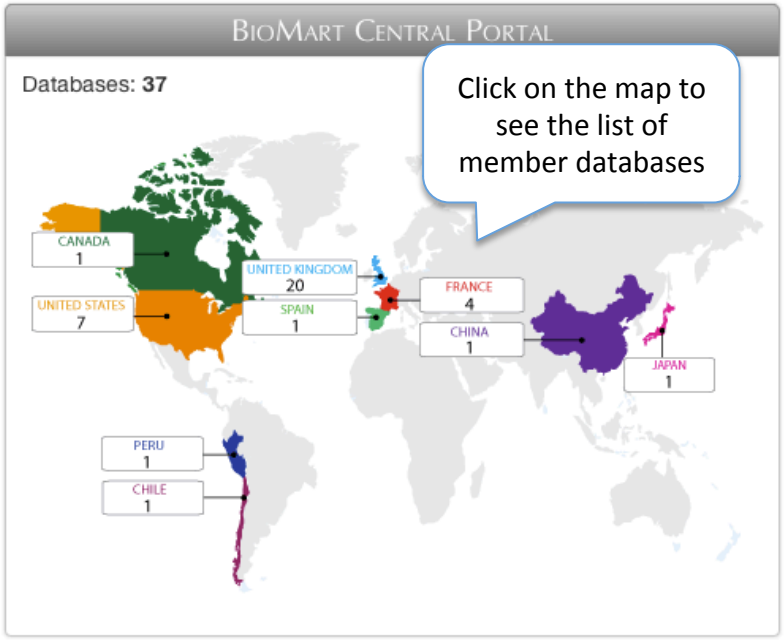
BioMart Central Portal Architecture



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)

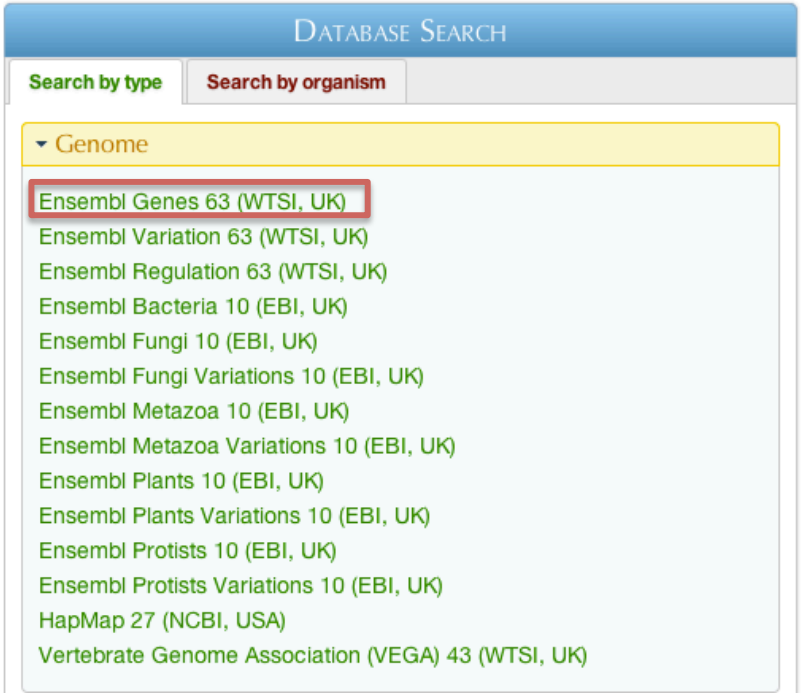




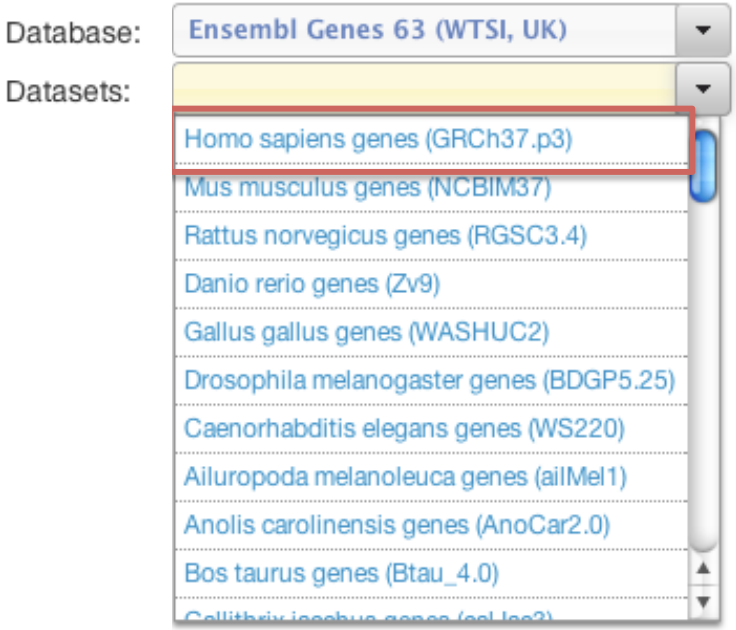
Querying BioMart Central Portal

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

1. Selecting Database



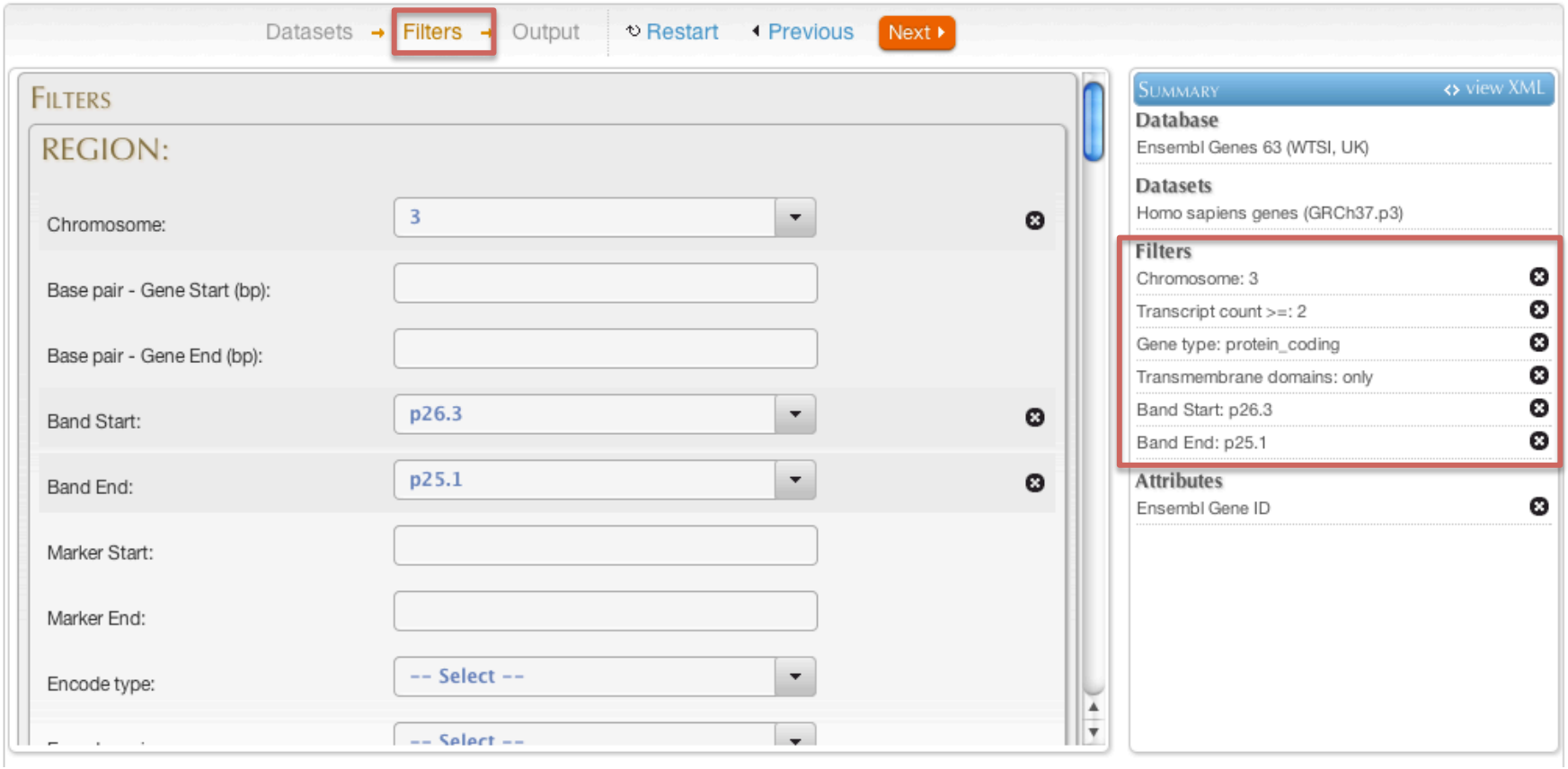
2. Selecting Dataset



Querying BioMart Central Portal

All BioMart queries are performed using the following simple steps:

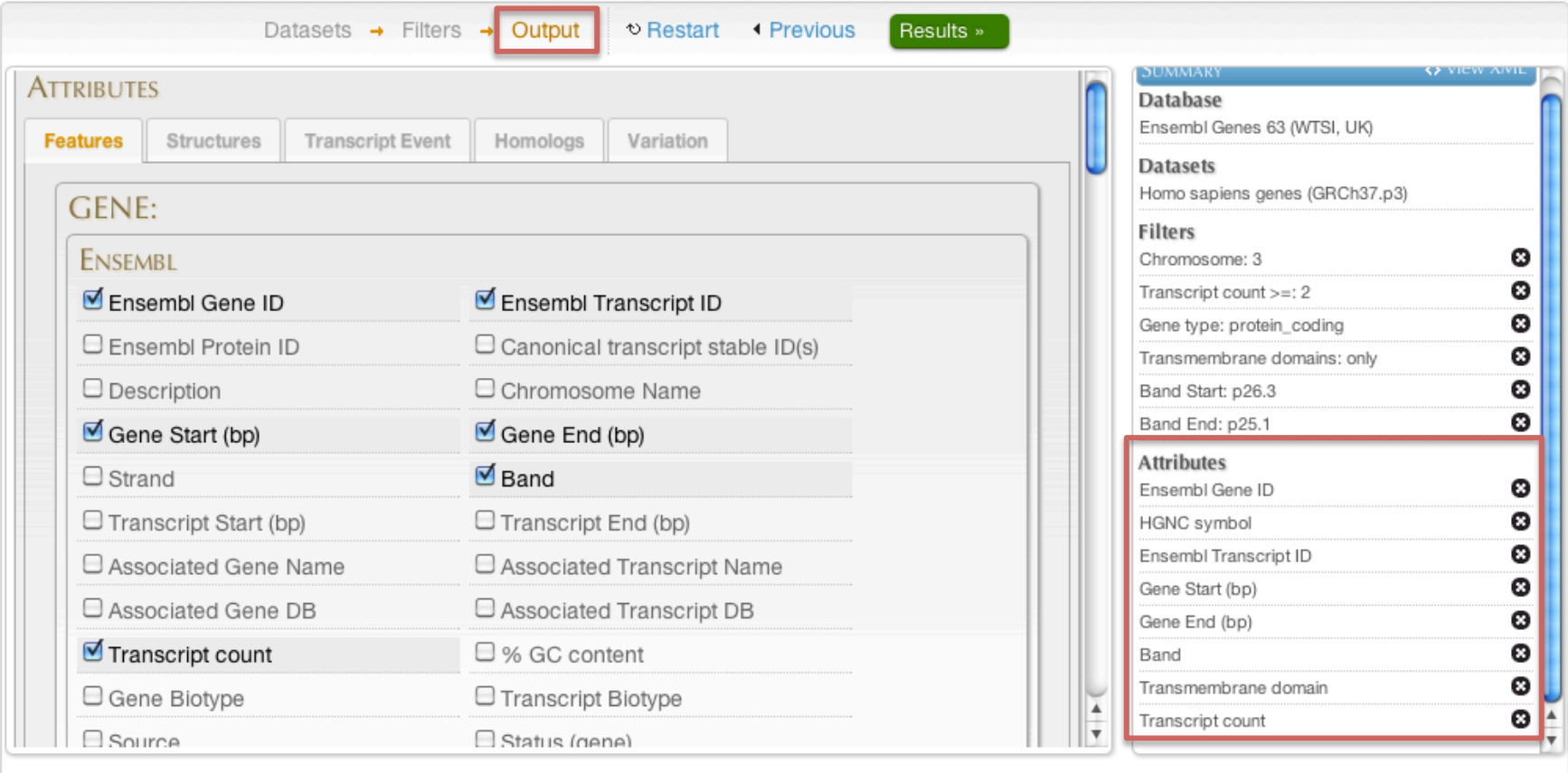
3. Selecting Filters (Optional)



Querying BioMart Central Portal

All BioMart queries are performed using the following simple steps:

4. Selecting Attributes



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

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

Queries can be bookmarked to be reused or shared

Ensembl Genes 63 (WTSI, UK)

Displaying results **1-20** out of **104**

See how many records are returned

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

Sort results by clicking on the black triangle toggle icon

Download data in tab-delimited tabular format

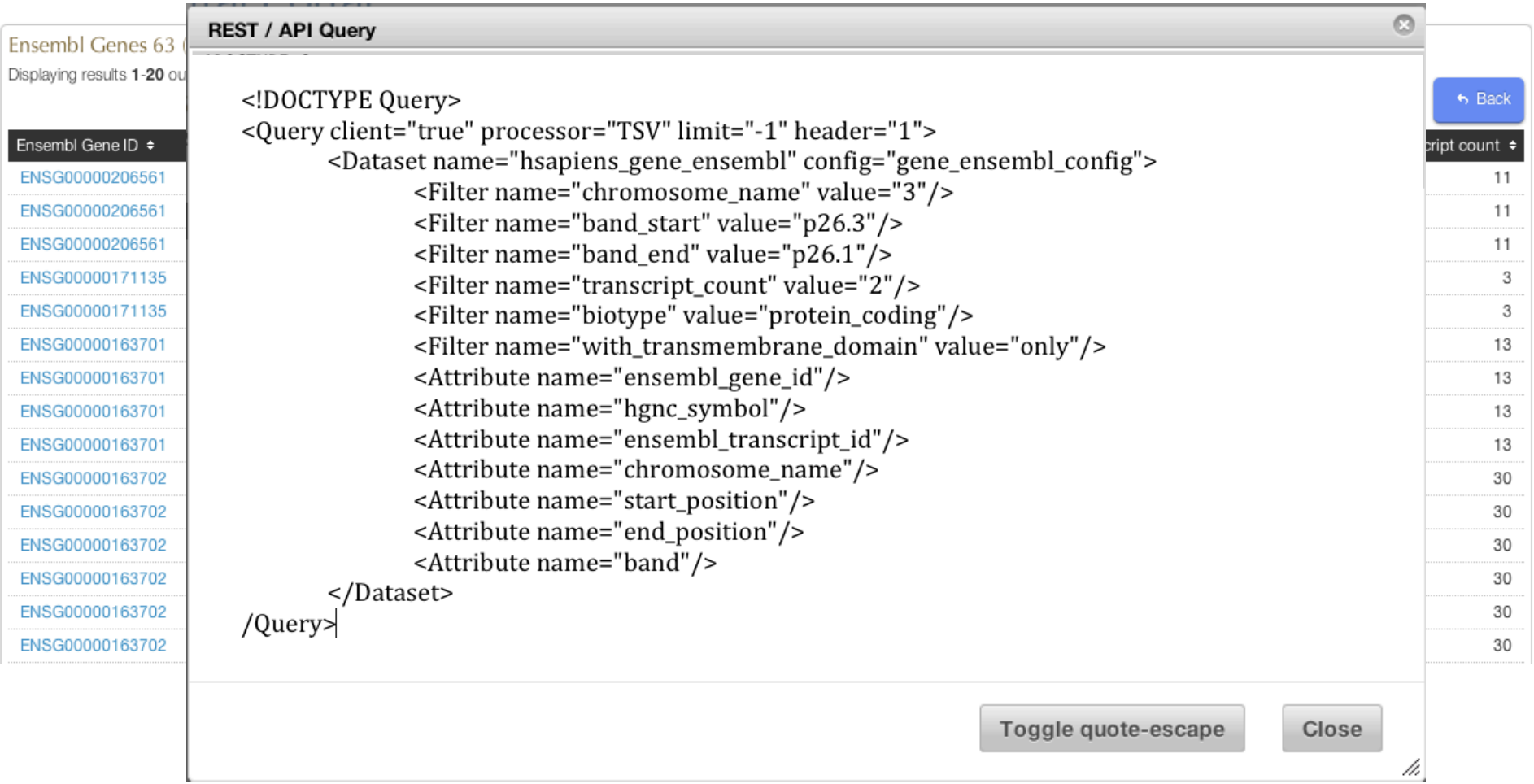
Ens	mbol	Ensembl Transcript ID	Gene Start (bp)	Gene End (bp)	Ban		Transcript count
EN	COLQ	EN	540	15563258	p2	hmm	11
EN	COLQ	EN	491640	15563258	p2	mhmm	11
ENSG00000206561	COLQ	EN	5491640	15563258	p2	mhmm	11
ENSG00000171135	JAGN1	EN	9932238	9936033	p25.3	Tmhmm	3
ENSG00000171135	JAGN1	ENST00000543379	9932238	9936033	p25.3	Tmhmm	3
ENSG00000163701	IL17RE	ENST00000421412	9944296	9958086	p25.3	Tmhmm	13
ENSG00000163701	IL17RE	ENST00000383814	9944296	9958086	p25.3	Tmhmm	13
ENSG00000163701	IL17RE	ENST00000295980	9944296	9958086	p25.3	Tmhmm	13
ENSG00000163701	IL17RE	ENST00000441648	9944296	9958086	p25.3	Tmhmm	13
ENSG00000163702	IL17RC	ENST00000383812	9958758	9975314	p25.3	Tmhmm	30
ENSG00000163702	IL17RC	ENST00000295981	9958758	9975314	p25.3	Tmhmm	30
ENSG00000163702	IL17RC	ENST00000403601	9958758	9975314	p25.3	Tmhmm	30
ENSG00000163702	IL17RC	ENST00000455057	9958758	9975314	p25.3	Tmhmm	30
ENSG00000163702	IL17RC	ENST00000413608	9958758	9975314	p25.3	Tmhmm	30
ENSG00000163702	IL17RC	ENST00000416074	9958758	9975314	p25.3	Tmhmm	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.

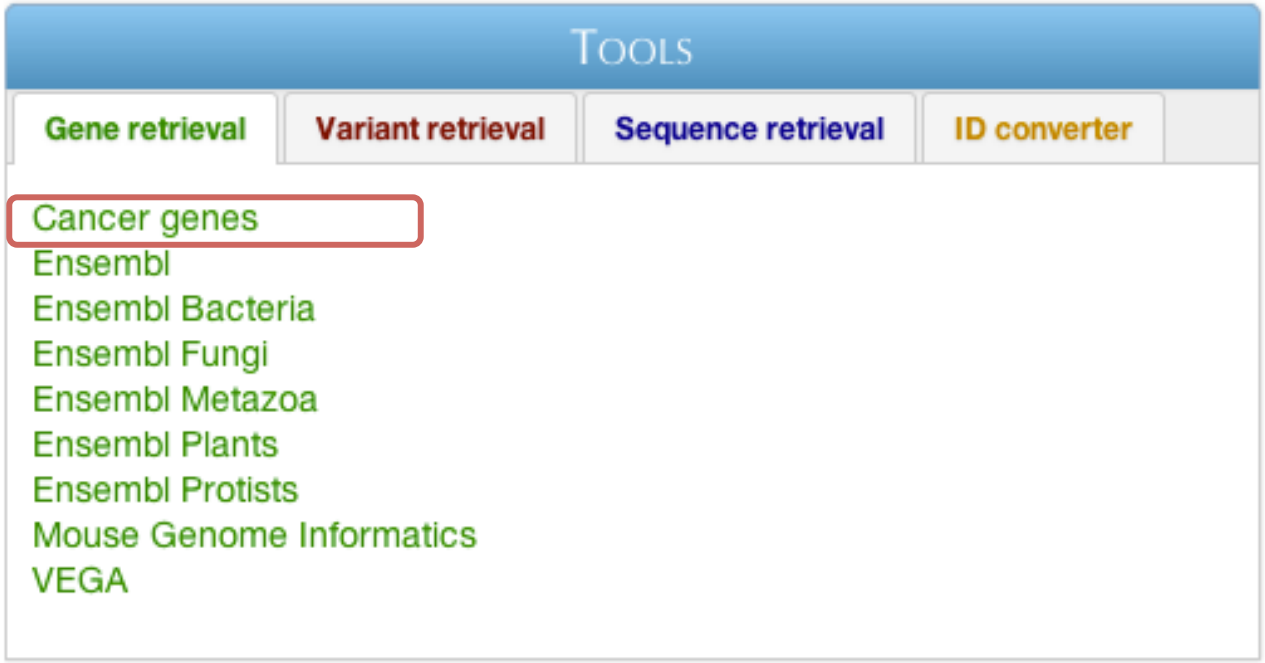


BioMart Central Portal allows cross-dataset querying

- 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

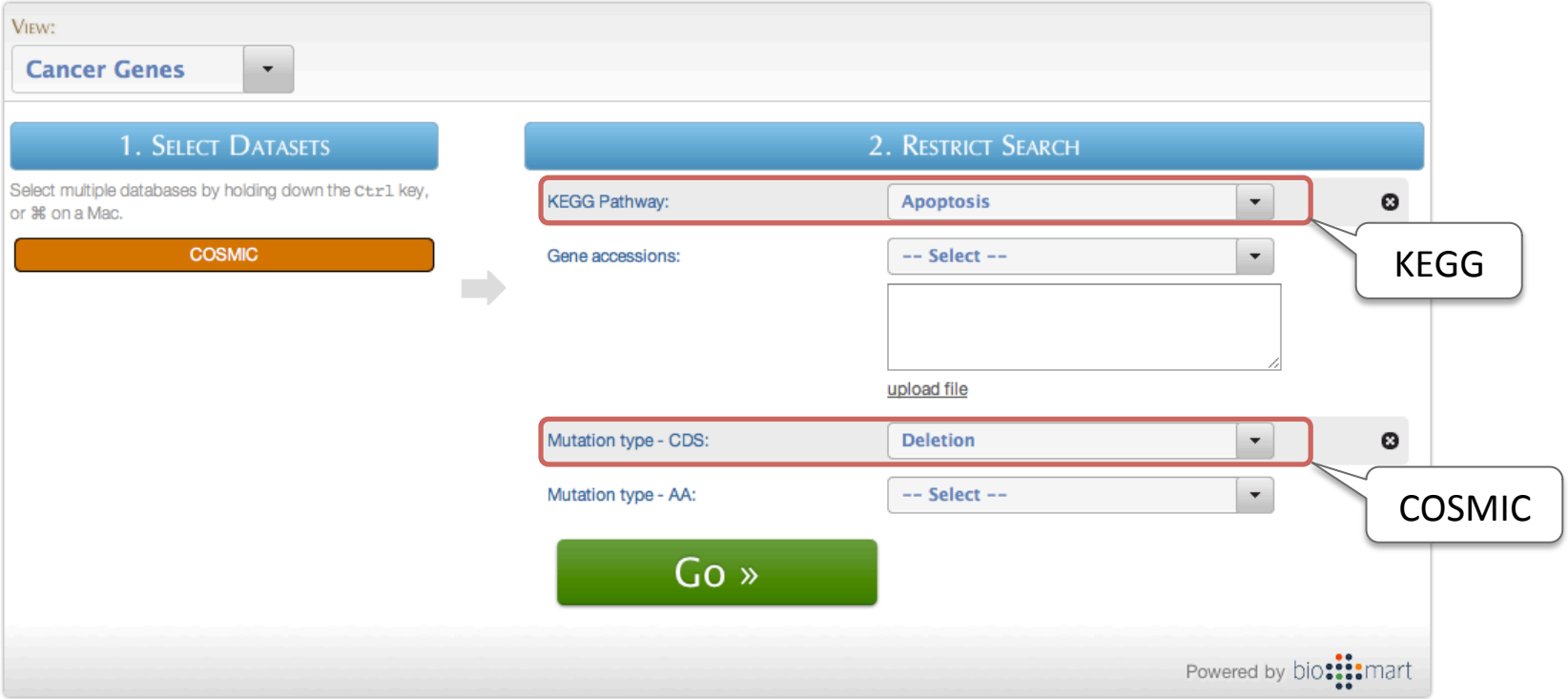
Query #1: 'Find deletion mutations in the COSMIC database that affect genes involved in Apoptosis

Step 1. Select Cancer genes under TOOLS/Gene retrieval



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



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

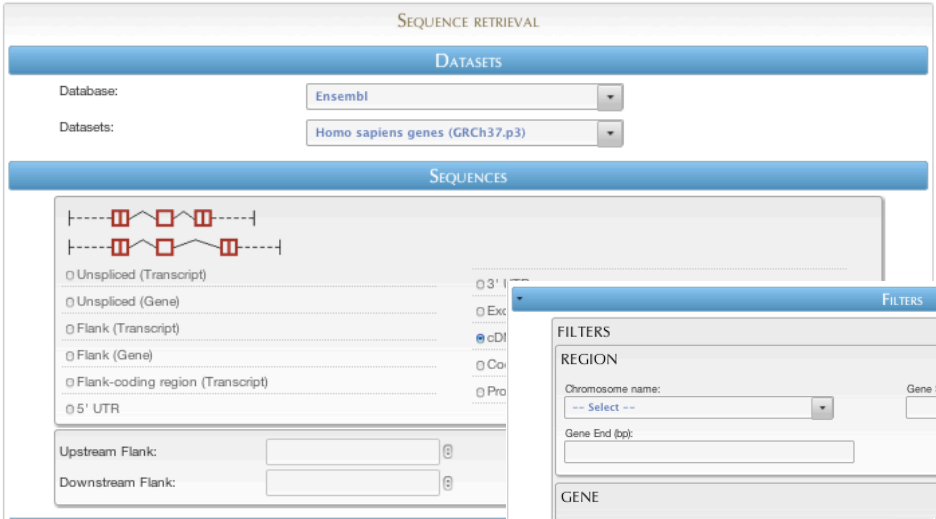
« Previous 3 4 5 6 **7** 8 9 10 11 12 Next »

Powered by biomart

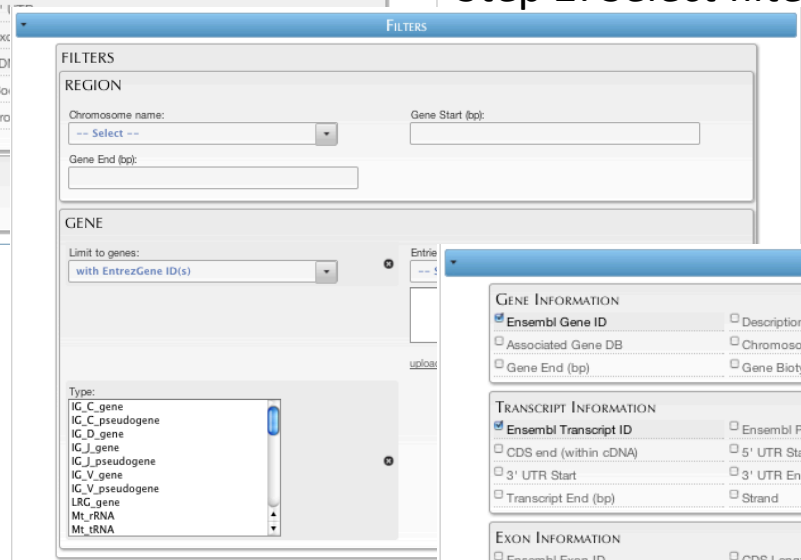
BioMart offers several plugins, e.g. TOOLS-Sequence retrieval

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
Mailing list: users@biomart.org