BioMart 0.8 offers new tools, more interfaces, and increased flexibility through plugins

Elena Rivkin

August 19, 2011

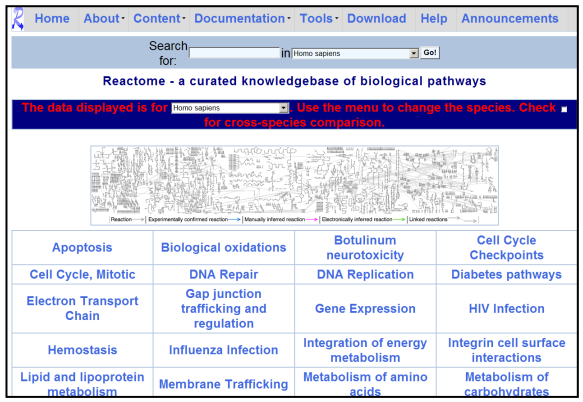
What is BioMart?

- Free, open-source federated data management system widely used by dozens of biological databases and international consortia
- Data-agnostic and platform independent, so existing databases can be easily converted to the BioMart format. An automated data conversion tool simplifies the process of BioMart installation.
- Allows data providers to quickly bring in-house data accessible online, and to effortlessly federate their data with existing public BioMart datasets
- Provides a uniform user interface for querying data stored in distributed databases
- Offers a variety of querying interfaces: web GUIs and programmatic access APIs (REST, SOAP, Perl, Java, SPARQL)
- Integrated with 3rd party software: biomaRt-BioConductor, Galaxy, Cytoscape, Taverna, Bioclipse, BioExtract, Gitools, Ruby API, WebLab

www.biomart.org

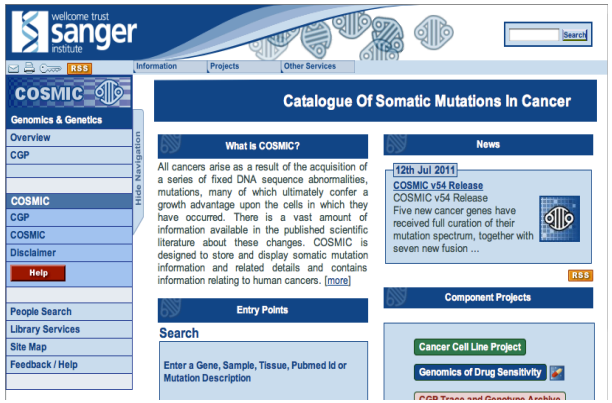
What is BioMart?

Reactome



Ensembl

COSMIC



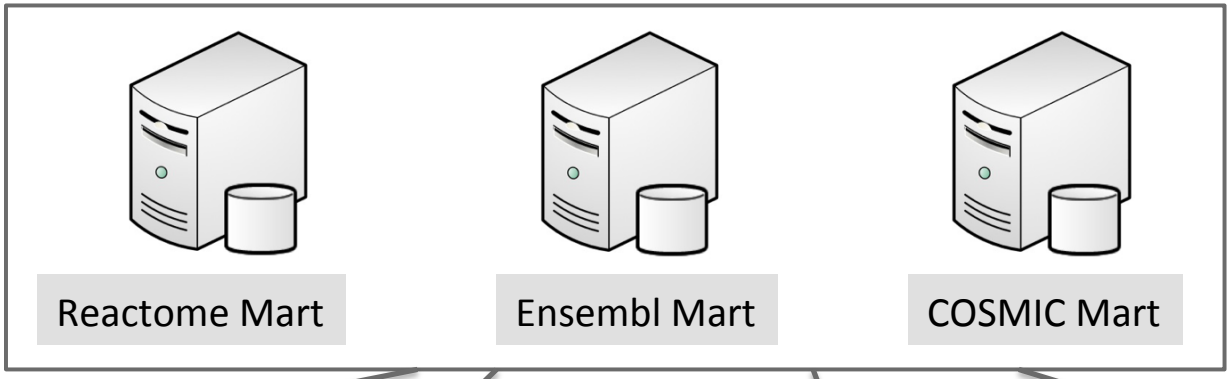
Reactome Mart

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 0.8 key features

- Integrated Java application makes it possible to build a BioMart data source, configure querying and presentation interfaces, and deploy a BioMart server from a single tool (*MartConfigurator*)
- Support more RDBMS (MS SQL Server, DB2, in addition to MySQL, PostgreSQL, and Oracle)
- Can create 'virtual mart' from 3NF normalized source database without materialization
- New diverse Web GUIs (MartForm, MartWizard, MartExplorer, MartReport, MartAnalysis) and APIs (PERL/SOAP, Java, SPARQL), provide added query flexibility
- New query optimizations: Link indexing and parallel querying optimizations
- Support several security features (HTTPS, OpenID and oAuth protocols) for managing sensitive data
- Extendable plugin framework for analysis and visualization

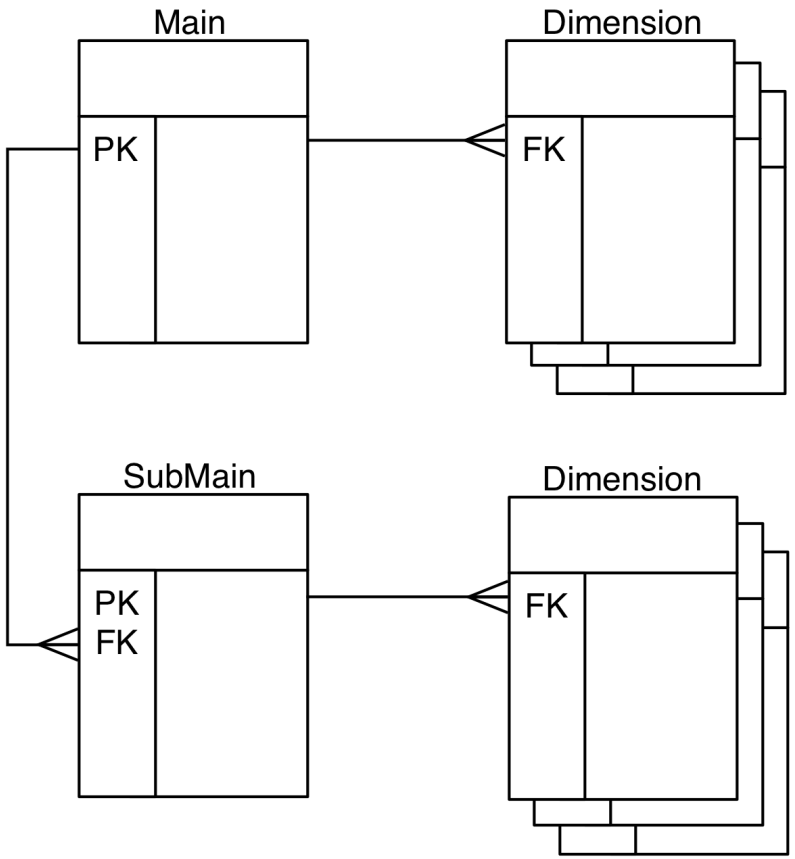
Basic BioMart Concepts – the Power of Simplicity

Building or querying a BioMart data source only requires understanding of a few basic concepts:

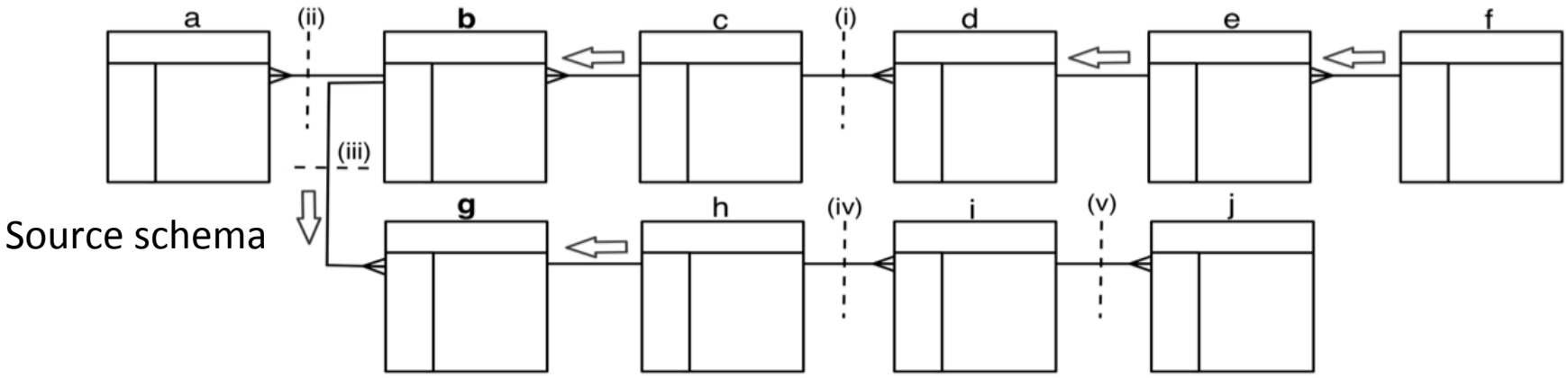
- *DataSource* – generally refers to the database server where the data is stored (e.g. Ensembl)
- *DataMart* – a collection of datasets in the BioMart format (e.g. Ensembl Genes 63)
- *DataSet* – a chunk of data stored in one or more tables, presented as a collection of filters and attributes (e.g. Homosapiens genes)
- *Attribute* – data element that can be retrieved from the dataset (Ensembl Gene ID)
- *Filter* – criteria used to restrict the query (e.g. Chromosome, gene type)
- *AccessPoint* (new) – graphical or programmatic entry to the data (e.g. Identifier Search, Analysis, Report page)
- *Analysis* (new) – query interfaces that present results from BioMart plug-in framework
- *Parameter* (new) -???
- *BioMart Portal* – BioMart server that provides access to a collection of data sources (e.g. ICGC Data Portal, BioMart Central Portal)

BioMart Schema

BioMart dataset is organized in a relational database schema called reverse star, optimized for fast retrieval of large quantities of descriptive data.

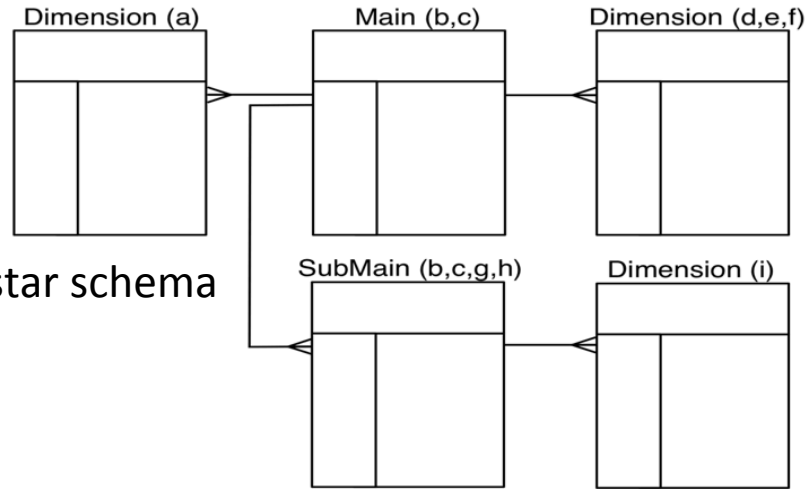


BioMart automatically converts any 3NF normalized database into the reversed star schema



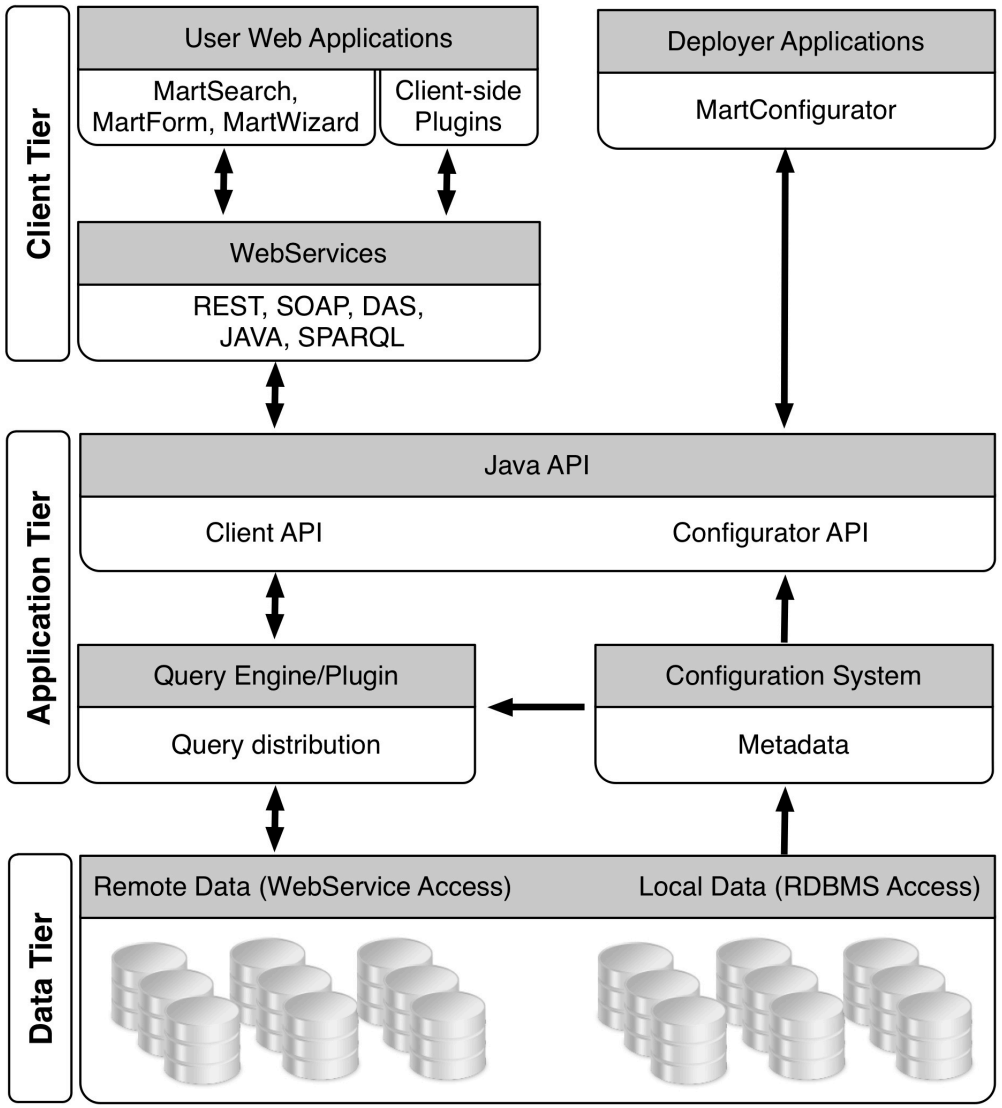
Source schema

If (b) is chosen as the candidate main table



Reverse star schema

BioMart system components



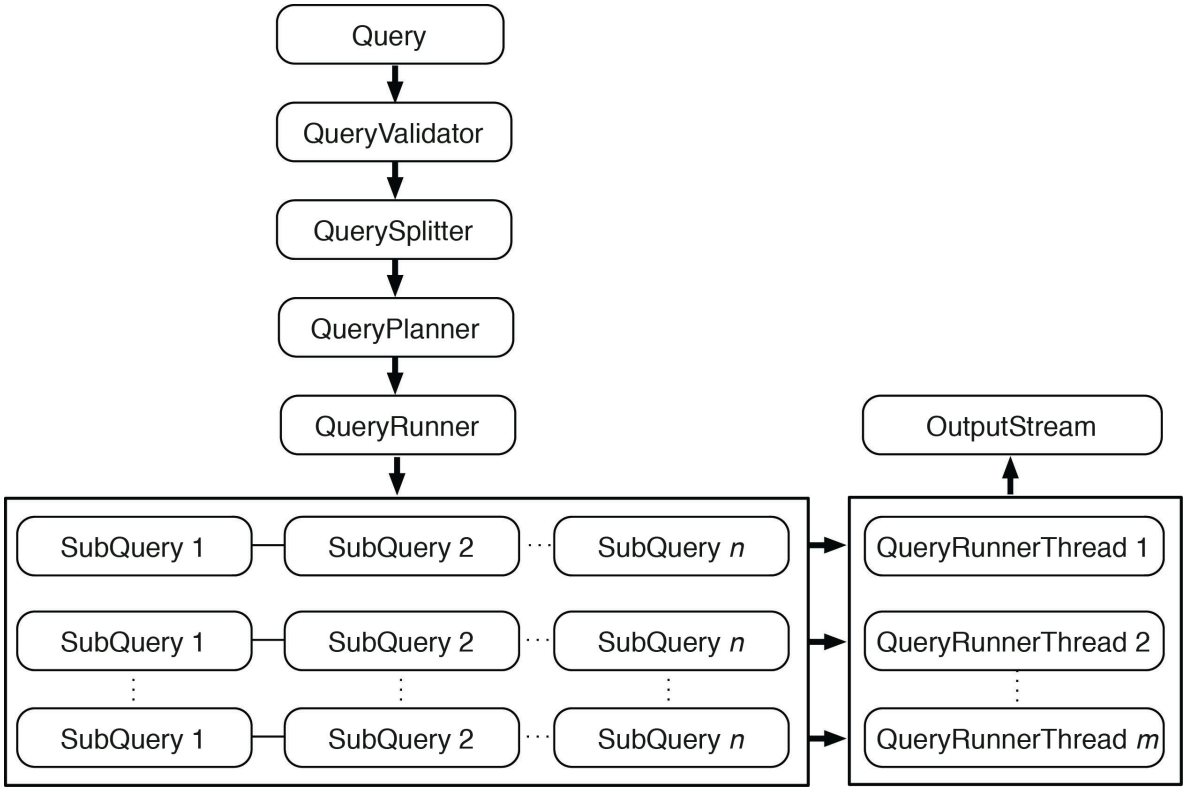
Client Tier: Deals with how the querying interface and query results are presented

Application Tier: Stores and retrieves metadata and executes queries: sends them to the appropriate destination, receives data in response, and organizes the results before sending them back to the client tier.

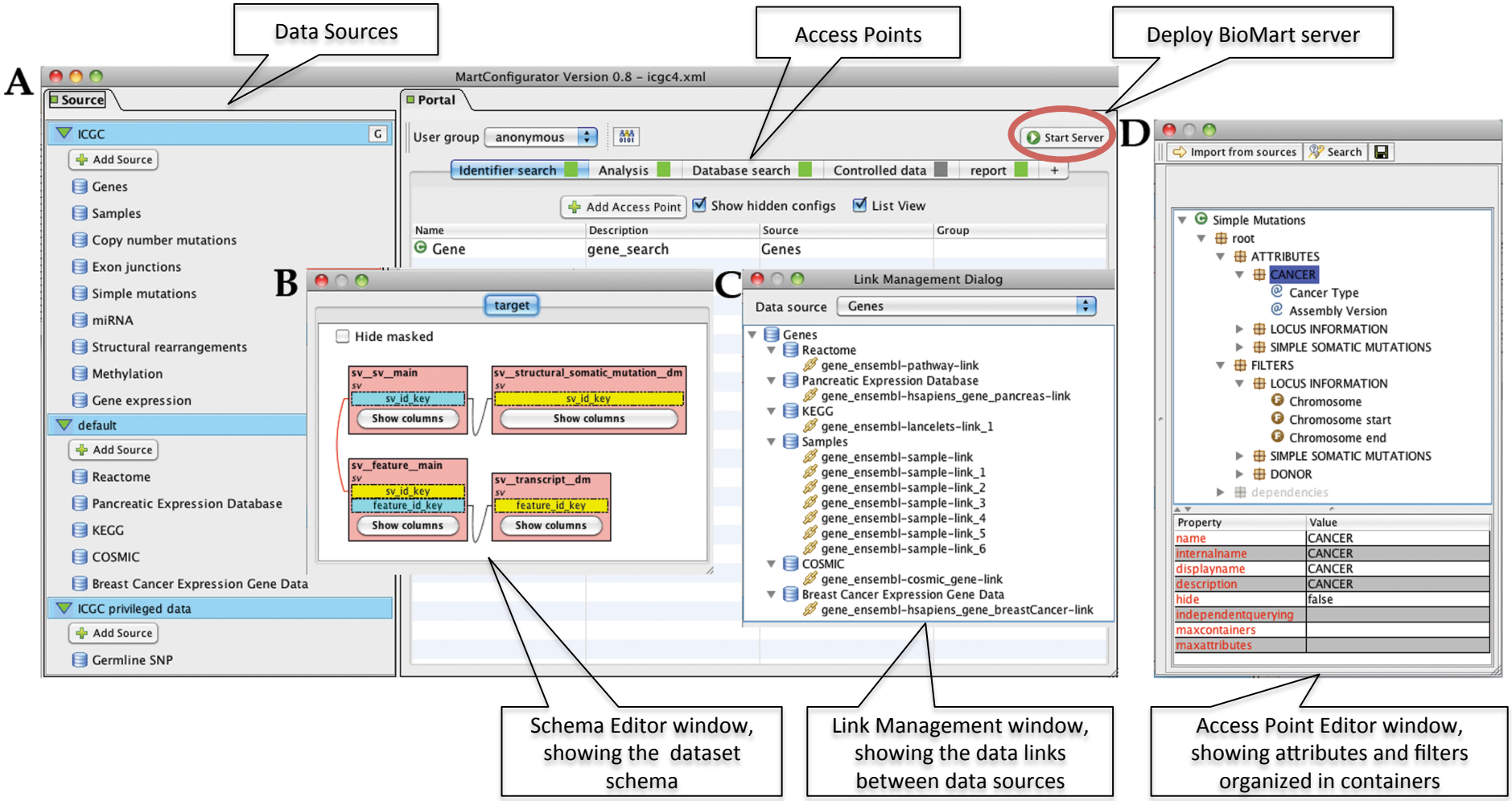
Data Tier: Deals with how data is organized, stored and managed in the relational database

BioMart Query Engine

Query engine breaks the query down to be distributed to the individual data sources. The results are returned back to the query engine, which recompiles the query into a single result set to be presented back to the user, or to be processed further using plugins. Includes a number of query optimizations to improve querying speed: optimized relational schema; support for physical partitioning of datasets, coupled with parallel query processing; optional use of pre-computed indexes (for joining multiple data sources)

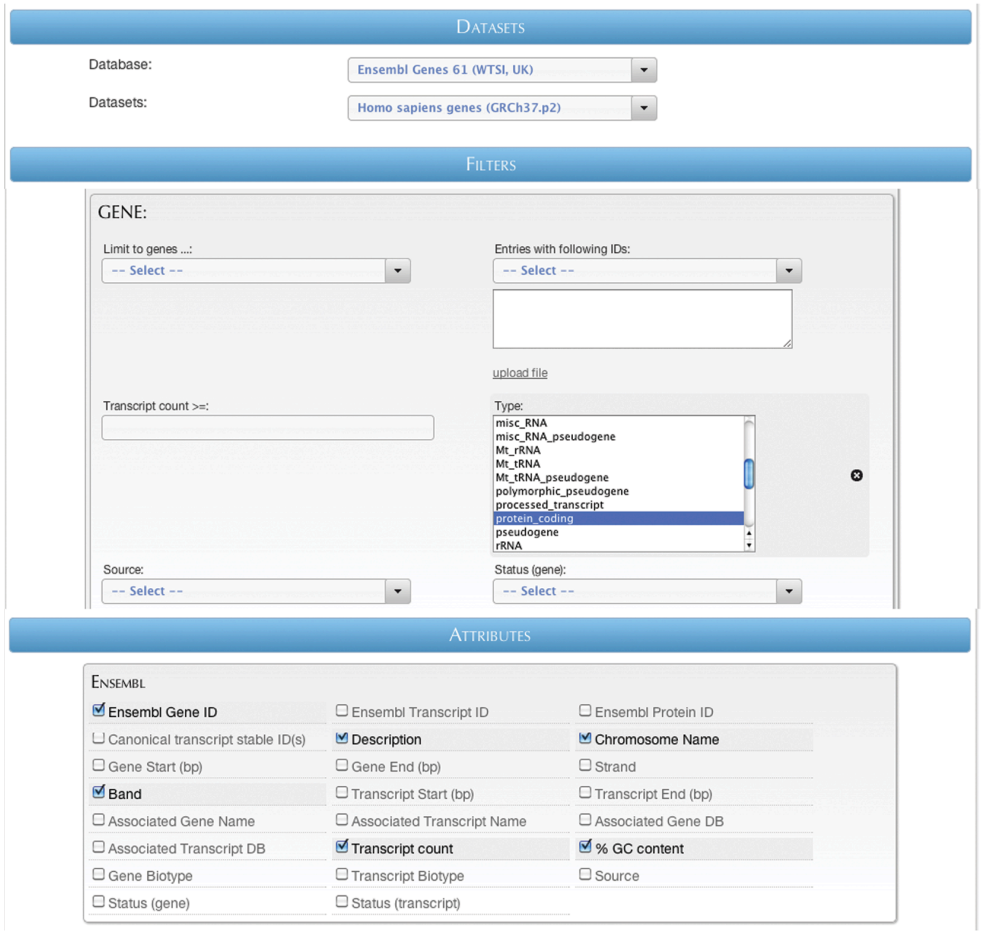


MartConfigurator – an integrated tool for setting up, configuring, and managing a BioMart server



BioMart 0.8 provides several querying web interfaces

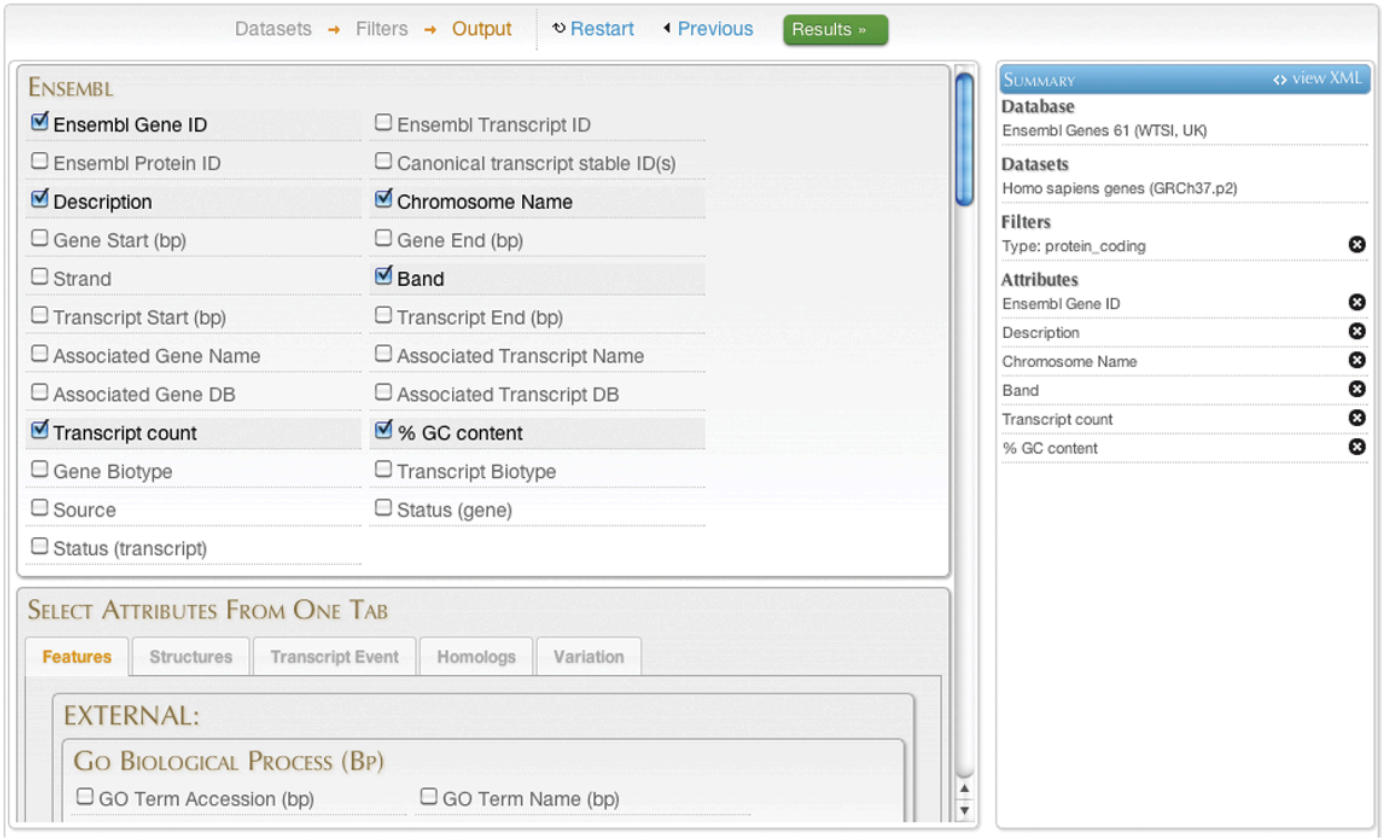
MartForm



Ideal for: Configurations with fewer Filters and Attributes; quick queries, where all selections (datasets, filters, attribute) are performed on the same page.

BioMart 0.8 provides several querying web interfaces

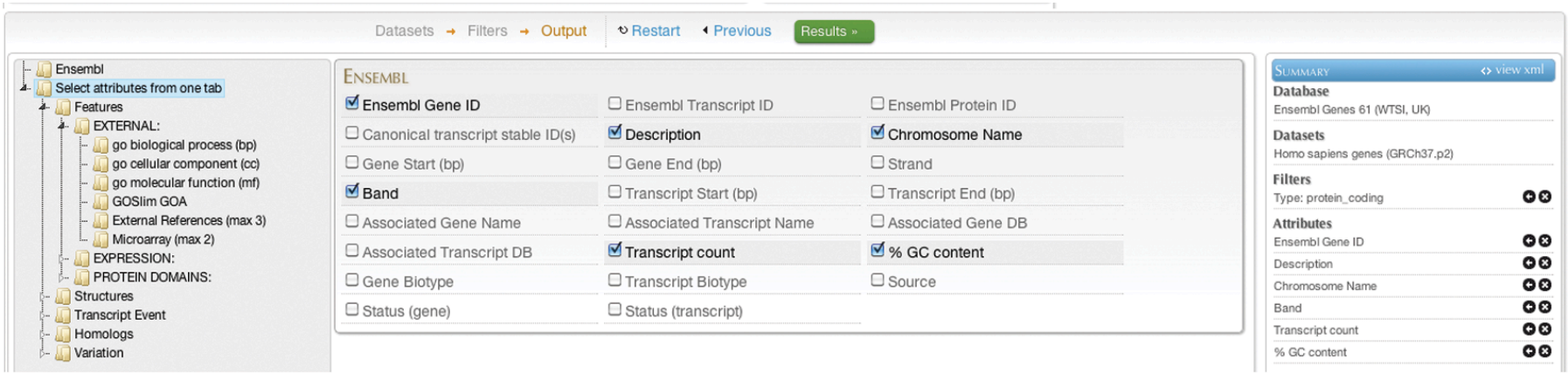
MartWizard



Ideal for: Configurations with more number of Filters and Attributes; datasets, filters are selected on separate pages. Right panel allows users to view, delete, and reorder their selections.

BioMart 0.8 provides several querying web interfaces

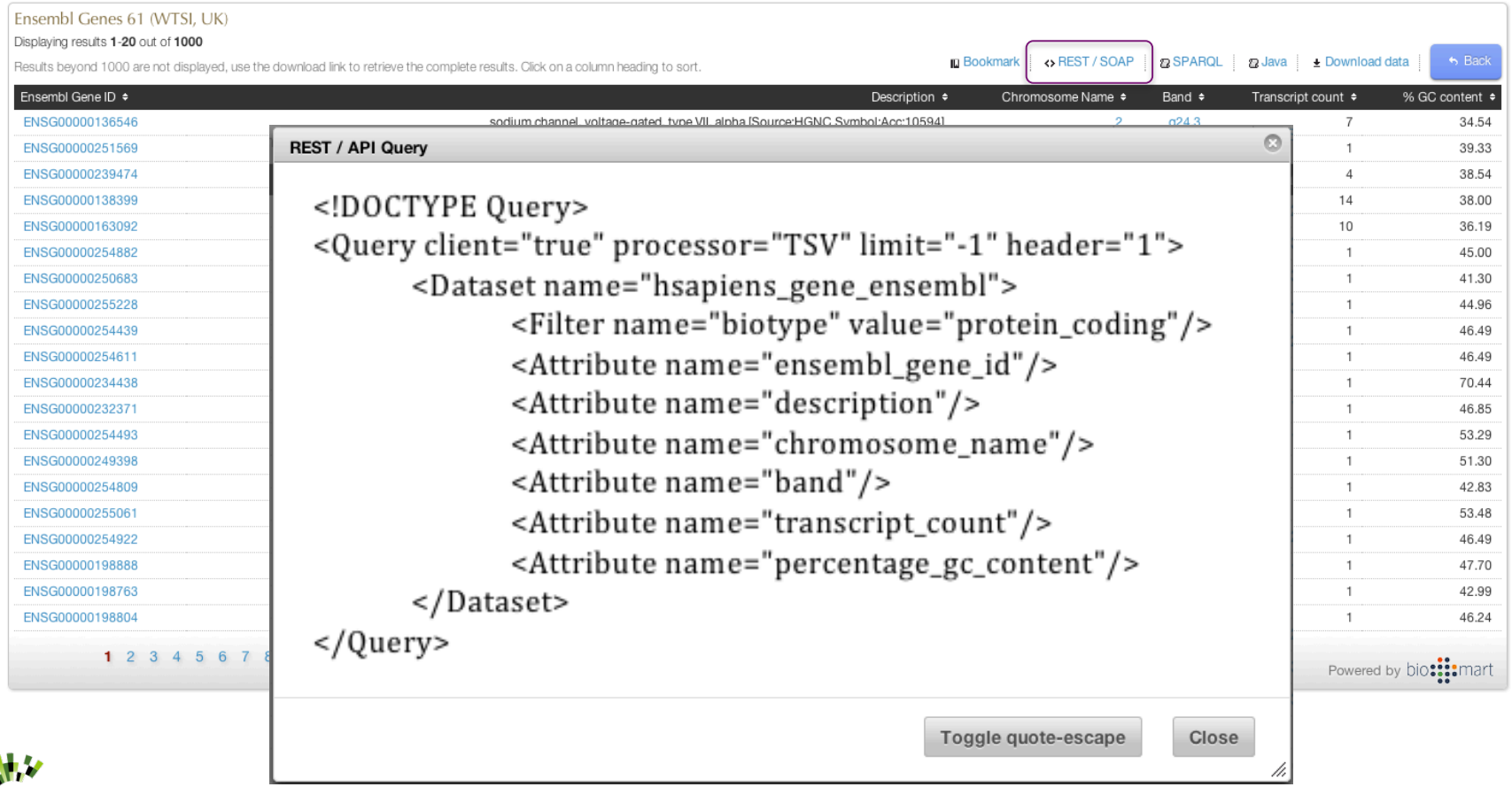
MartExplorer



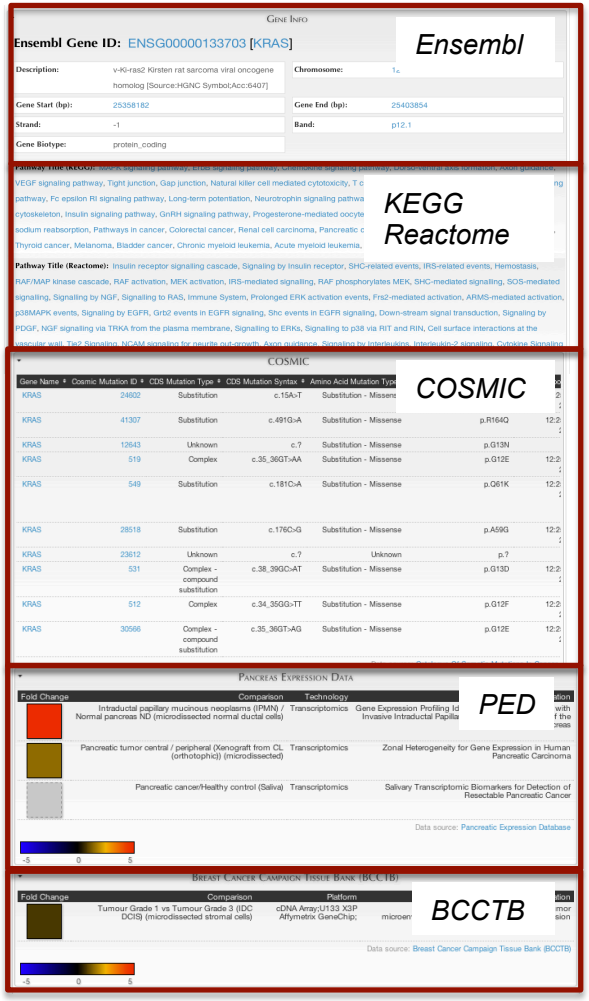
Ideal for: Datasets with large number of Filters and Attributes organized in multiple containers; datasets, filters are selected on separate pages. Right panel allows users to view, delete, and reorder their selections. Left panel allows to view and browse filter/ attribute containers.

Programmatic access - API query syntax at the click of a button

BioMart provides several programmatic interfaces: REST/SOAP API, Java API, 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. Therefore, queries can be saved, modified, and easily transferred from one format to another.

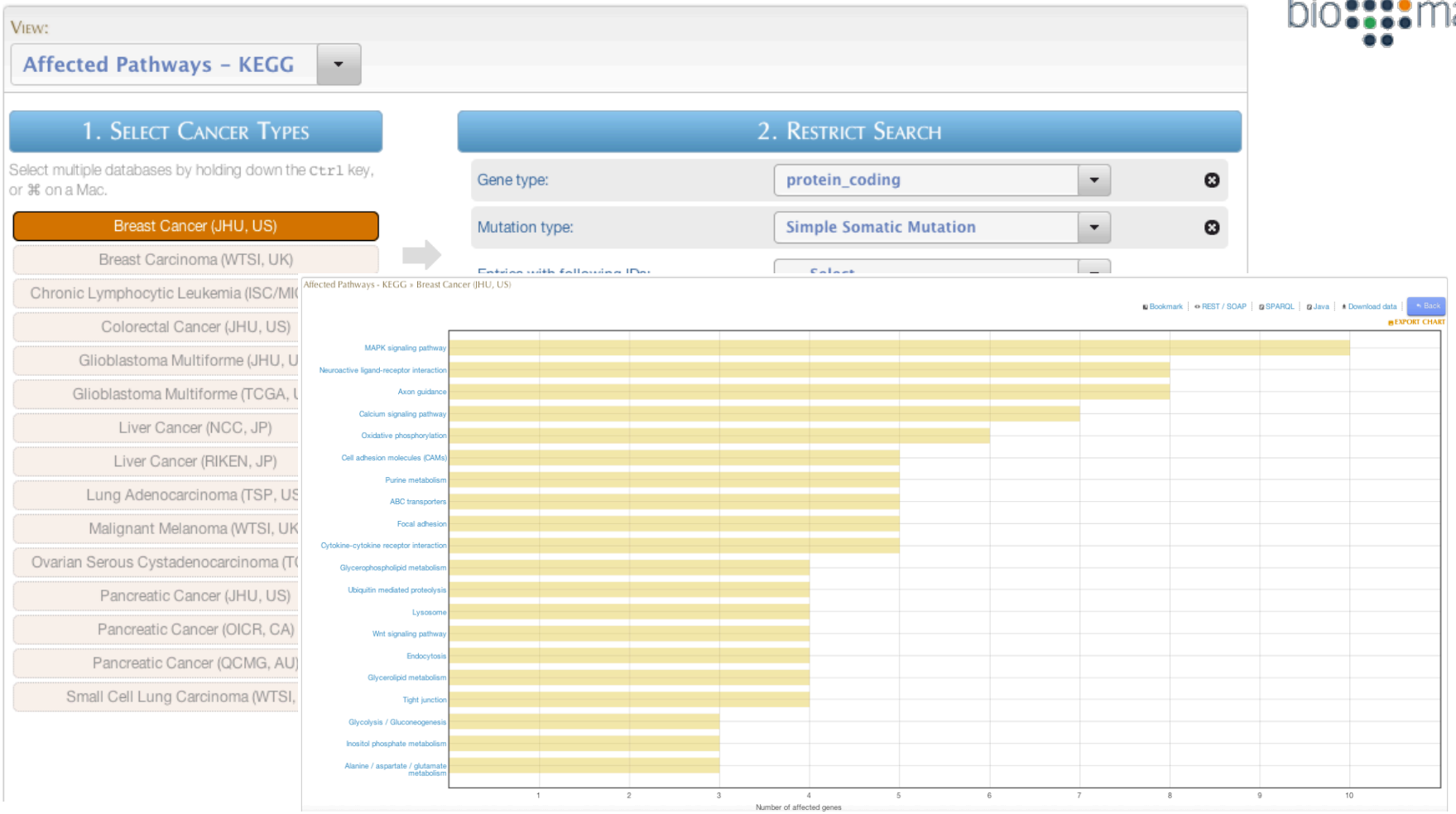


Special GUI - MartReport



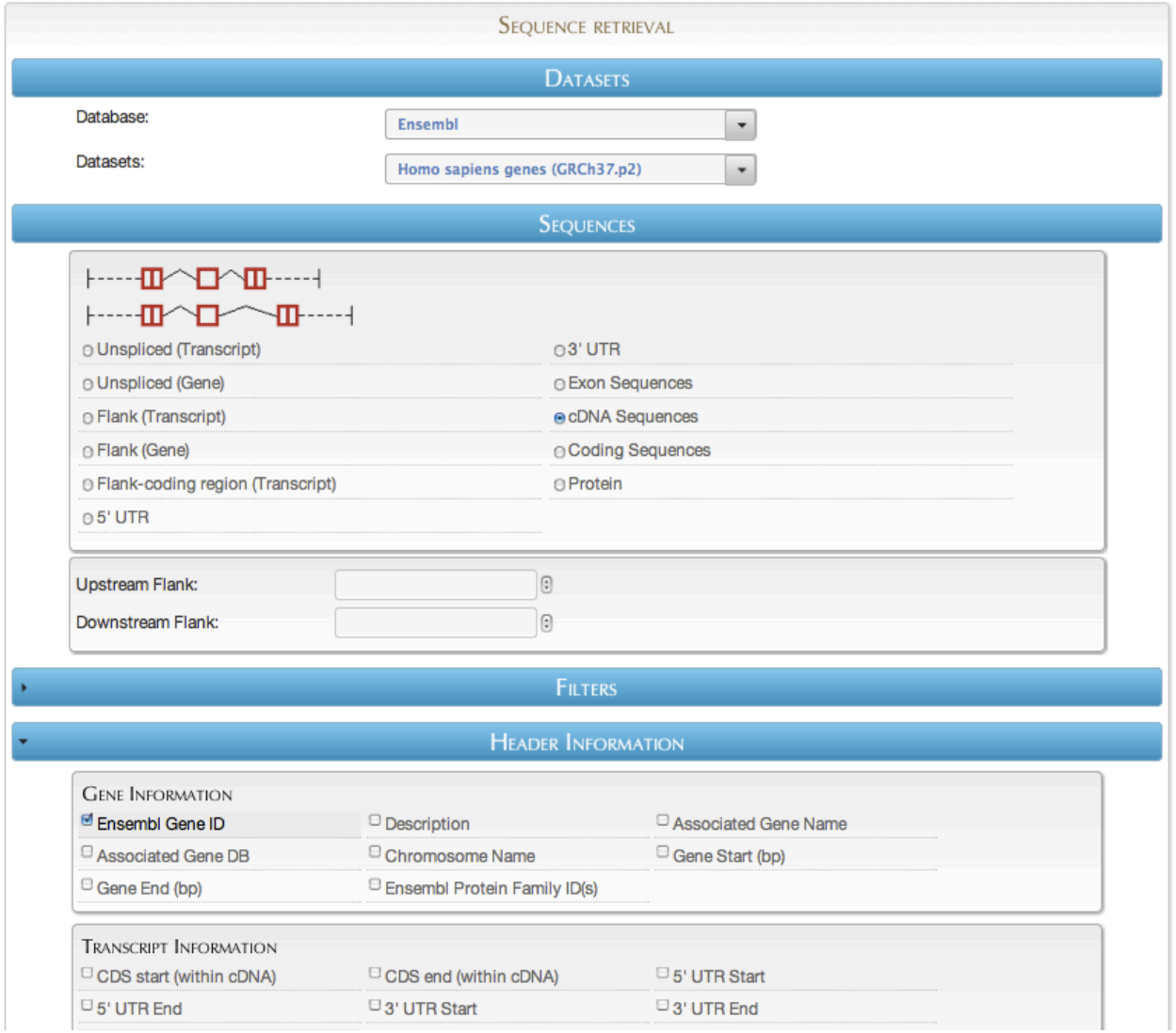
Ideal for: Displaying many attributes linked to a single identifier. Data from multiple data sources can be integrated through the BioMart data federation mechanism (e.g. Gene report)

Special GUI – MartAnalysis



Ideal for: Displaying query results that have been processed through the BioMart plugin framework (e.g. chart presenting the Most affected pathways)

Special GUI – Sequence Retrieval



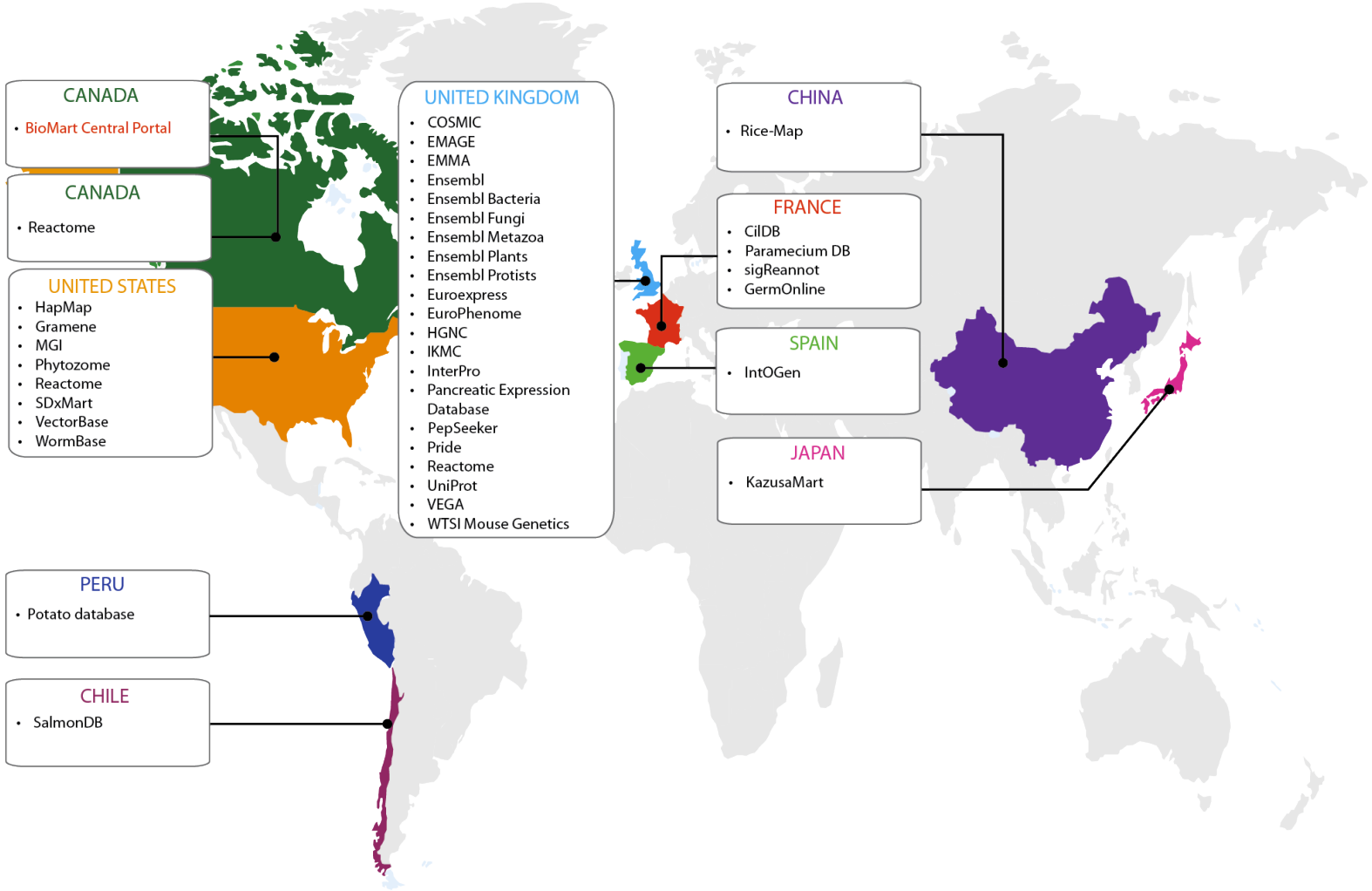
Sequence retrieval tool is implemented as server-side analysis plugin

Large collaborative projects using BioMart for data management:

- BioMart Central Portal (http://central.biomart.org) (Refer to the 'BioMart Central Portal' Presentation for more details)
- International Cancer Genome Consortium (http://dcc.icgc.org) (Refer to the 'ICGC Data Portal Presentation' for more details.
- PopCure (Private and public collaboration, controlled access)
- Digital Medicine (Canada, controlled access)

BioMart Central Portal (central.biomart.org)

First-of-its kind, community-driven effort to provide unified access to dozens of biological databases spanning genomics, proteomics, model organisms, cancer data, and more



(central.biomart.org)

biomart

Identifier Search

Go

Examples: KRAS, ENSG00000146648

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

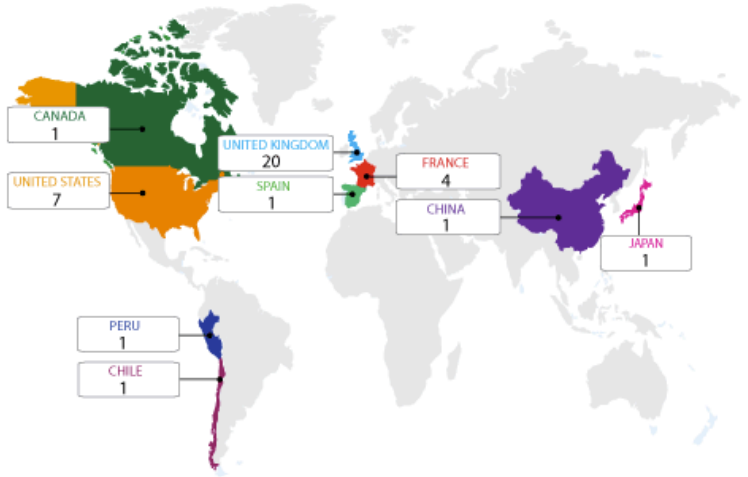
Database Search

Search by type | Search by organism

- Genome
- Gene annotation
- Protein sequence and structure
- Interaction and pathways
- Gene expression
- Cancer
- Model organism databases

BioMart Central Portal

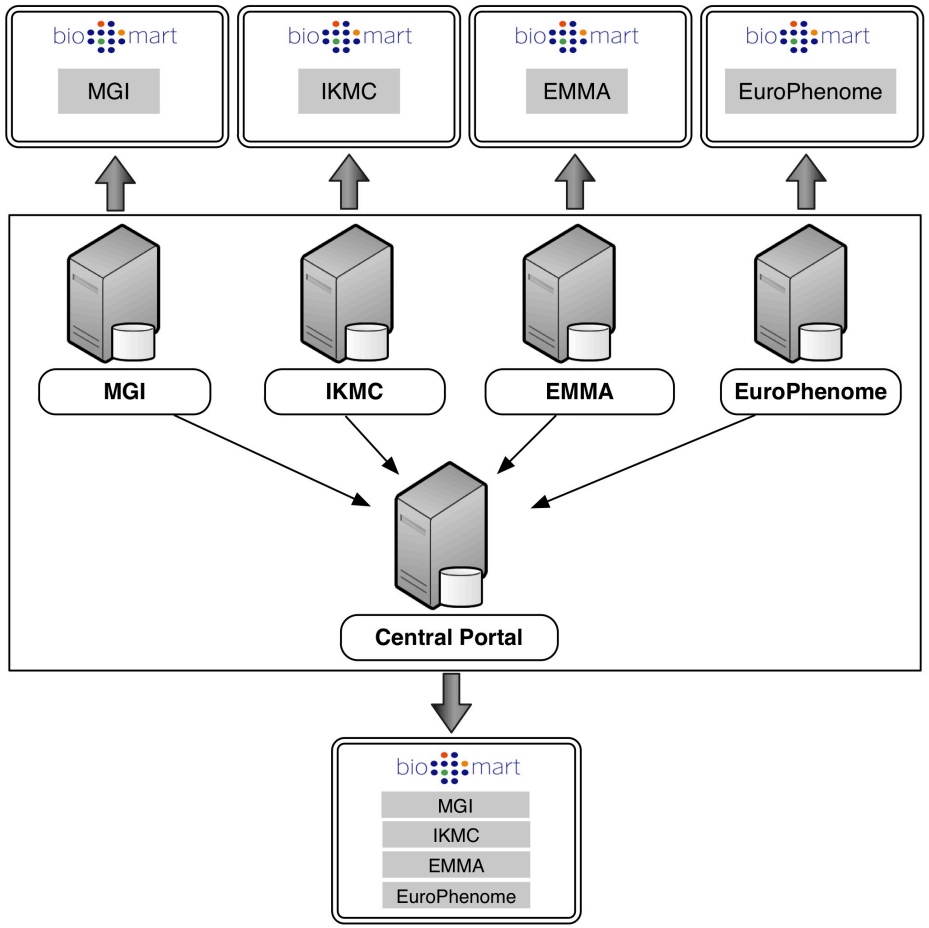
Databases: **37**



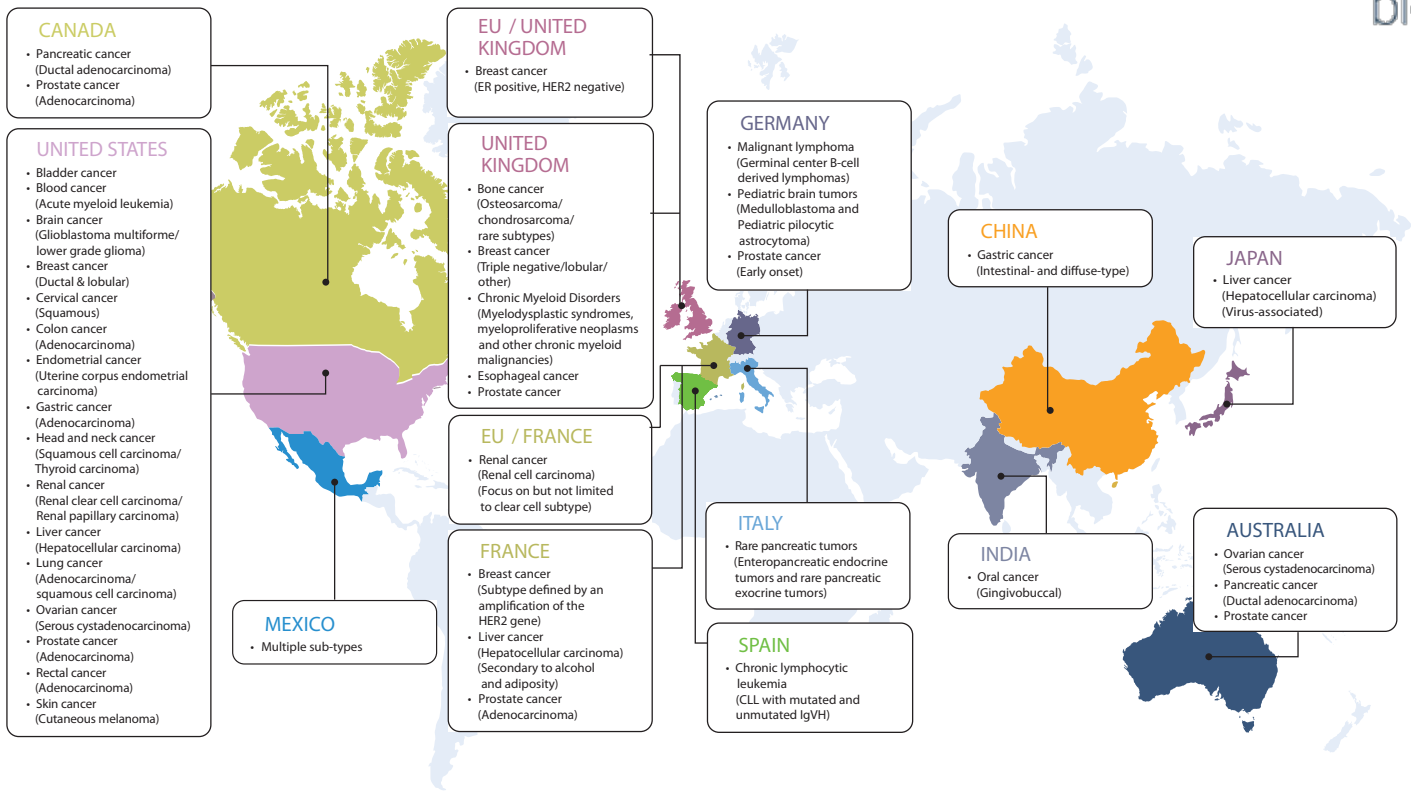
OICR Ontario Institute for Cancer Research

BioMart Central Portal Architecture

BioMart Central Portal provides access to a collection of data sources and is configured in **“Master/Slave”** like architecture, where individual servers present only their own data sources while a single ‘master’ server acts as a portal providing a unified view of all the sources.



International Cancer Genome Consortium (ICGC) Data Portal

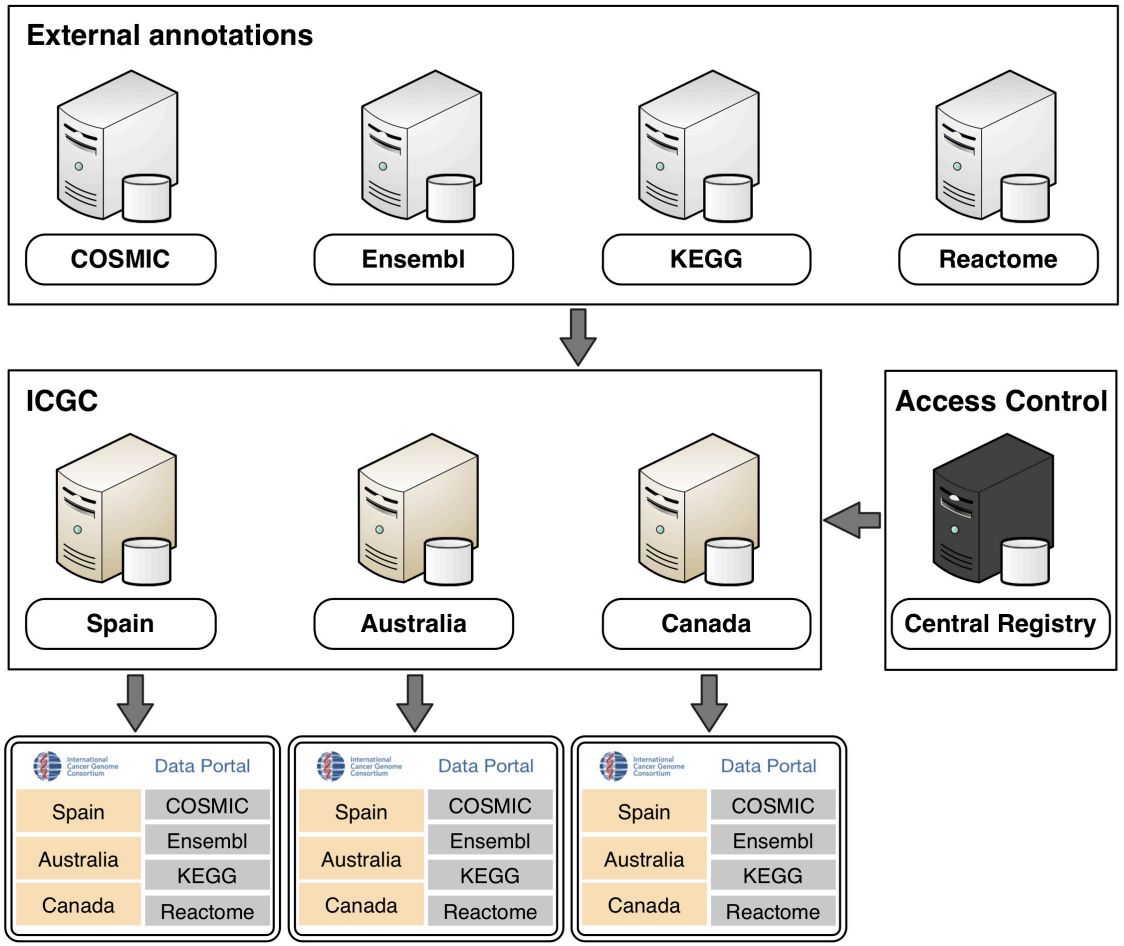


GOALS: To obtain a comprehensive description of genomic, transcriptomic, and epigenomic changes in 50 different tumor types and/or subtypes, which are of clinical and societal importance across the globe. 500 tumor and matched control samples will be analyzed per tumor type. At present, 12 countries joined ICGC. Data will be generated by institutions all over the world.

To make the data available rapidly and with minimal restrictions, to accelerate research of the causes and control of cancer.

ICGC Data Portal Architecture

ICGC Data Portal is configured in a "**Peer-to-Peer**" like architecture, where each of the individual servers links to all of the other servers in the group, such that all servers act as a portal providing access to all the data.



(dcc.icgc.org)

International Cancer Genome Consortium Data Portal

OICR Ontario Institute for Cancer Research

Home | ICGC Home | Publication Policy | Download Data | Documentation | Help

Home

Not logged in (Login) You are on the: Canada website

Identifier Search

Go

Examples: TP53, ENSG00000133703, NM_000314

Analysis

Genes | Pathway

Affected Genes

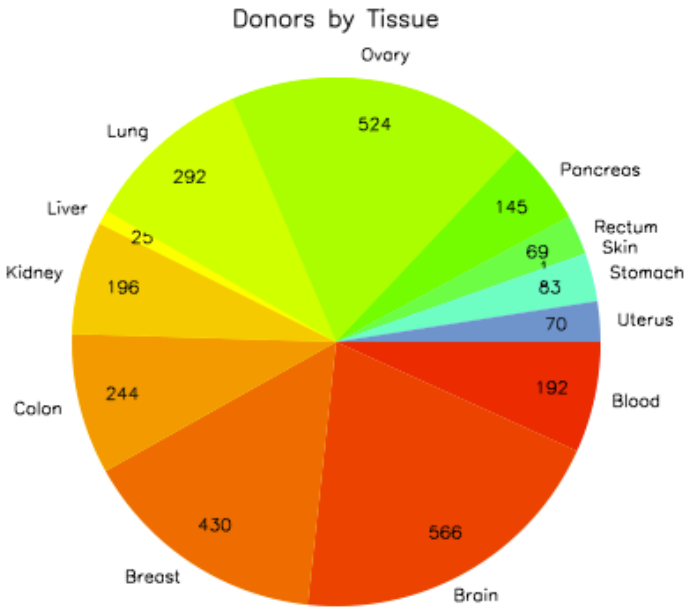
Database Search

Quick | Flexible | Advanced

Genes
Samples
Simple Mutations
Copy Number Alterations
Structural Rearrangements
Gene Expression
Methylation
miRNA
Exon Junction

ICGC Dataset Version 6 (July 7th, 2011)

Cancer Projects: **25**



Total Donors: **2,837**

Powered by biomart

OICR Ontario Institute for Cancer Research

BioMart Future Directions

- Creation of *BioMart Central Registry* to improve coordination between BioMart servers. It will be a permanent resource where BioMart data providers can register their data models, data sources and services.

- Enhancing data transformation module for building BioMart databases from non-RDBMS data sources (e.g. flat data files, XML data files etc.) with high scalability and flexibility.

- Enhancing the plugin system to allow various forms of data analysis and visualization. Third parties are encouraged to develop plugins to extend the capabilities of the system.

The BioMart team

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

For support: users@biomart.org