



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

What is BioMart?



Reactome

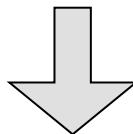
[Home](#) [About](#) [Content](#) [Documentation](#) [Tools](#) [Download](#) [Help](#) [Announcements](#)

Search for: in Homo sapiens Go!

Reactome - a curated knowledgebase of biological pathways

The data displayed is for Homo sapiens. Use the menu to change the species. Check for cross-species comparison.

Apoptosis	Biological oxidations	Botulinum neurotoxicity	Cell Cycle Checkpoints
Cell Cycle, Mitotic	DNA Repair	DNA Replication	Diabetes pathways
Electron Transport Chain	Gap junction trafficking and regulation	Gene Expression	HIV Infection
Hemostasis	Influenza Infection	Integration of energy metabolism	Integrin cell surface interactions
Lipid and lipoprotein metabolism	Membrane Trafficking	Metabolism of amino acids	Metabolism of carbohydrates



Reactome Mart

Ensembl

[Ensembl Home](#) [Login / Register](#) [BLAST/BLAT](#) [BioMart](#) [Docs & FAQs](#)

Search Ensembl Human

Search for: e.g. gene BRCA2 or AL032821.2.1.143563 or muscular dystrophy

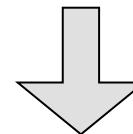
[Assembly and Genebuild](#) [Description](#)

Assembly

This release is based on the NCBI 36 assembly of the [human genome](#) [November 2005]. The data consists of a reference assembly of the complete genome plus the Celera WGS and a number of alternative assemblies of individual haplotypic chromosomes or regions. [Full list of assemblies](#)

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page



Ensembl Mart

COSMIC

[wellcome trust sanger institute](#) [Information](#) [Projects](#) [Other Services](#) [Search](#)

Catalogue Of Somatic Mutations In Cancer

What is COSMIC?

All cancers arise as a result of the acquisition of a series of fixed DNA sequence abnormalities, mutations, many of which ultimately confer a growth advantage upon the cells in which they have occurred. There is a vast amount of information available in the published scientific literature about these changes. COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers. [\[more\]](#)

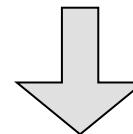
News

[12th Jul 2011] COSMIC v54 Release COSMIC v54 Release Five new cancer genes have received full curation of their mutation spectrum, together with seven new fusion ... [\[RSS\]](#)

Component Projects

Search

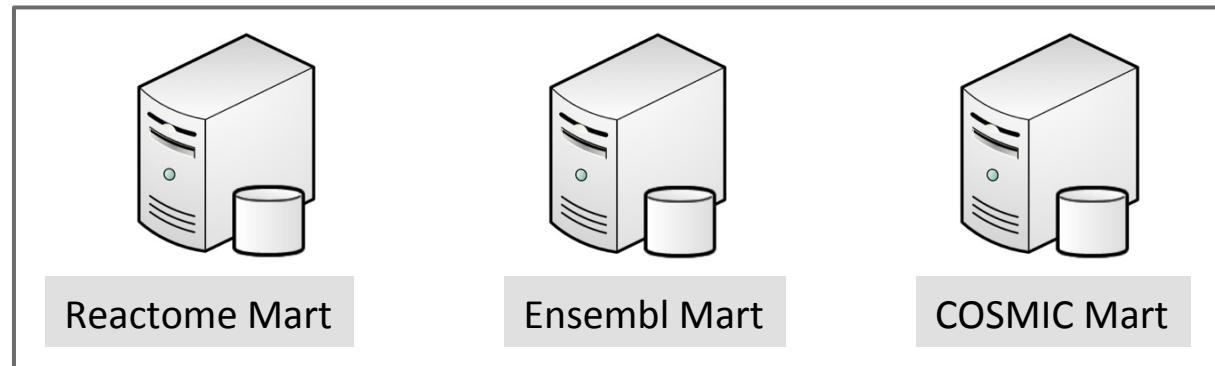
Enter a Gene, Sample, Tissue, Pubmed Id or Mutation Description



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



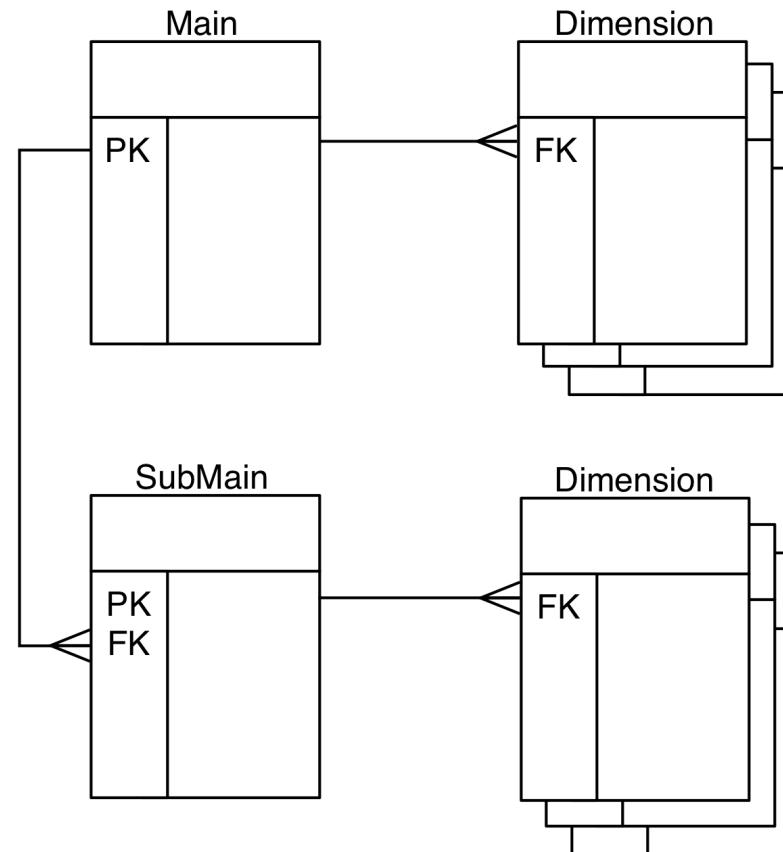
Building or querying a BioMart data source only requires understanding of a few basic concepts:

- *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)

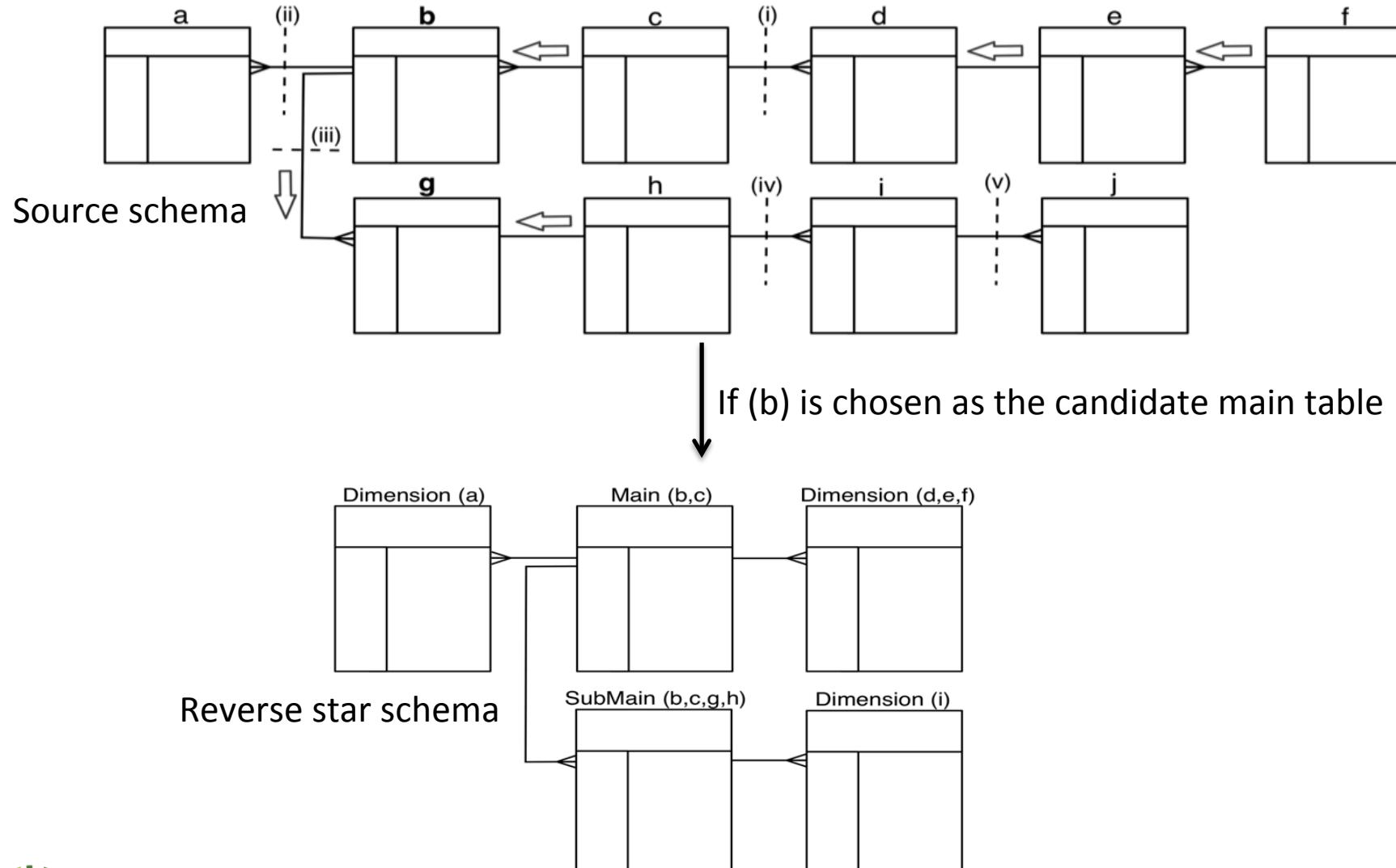
BioMart Schema



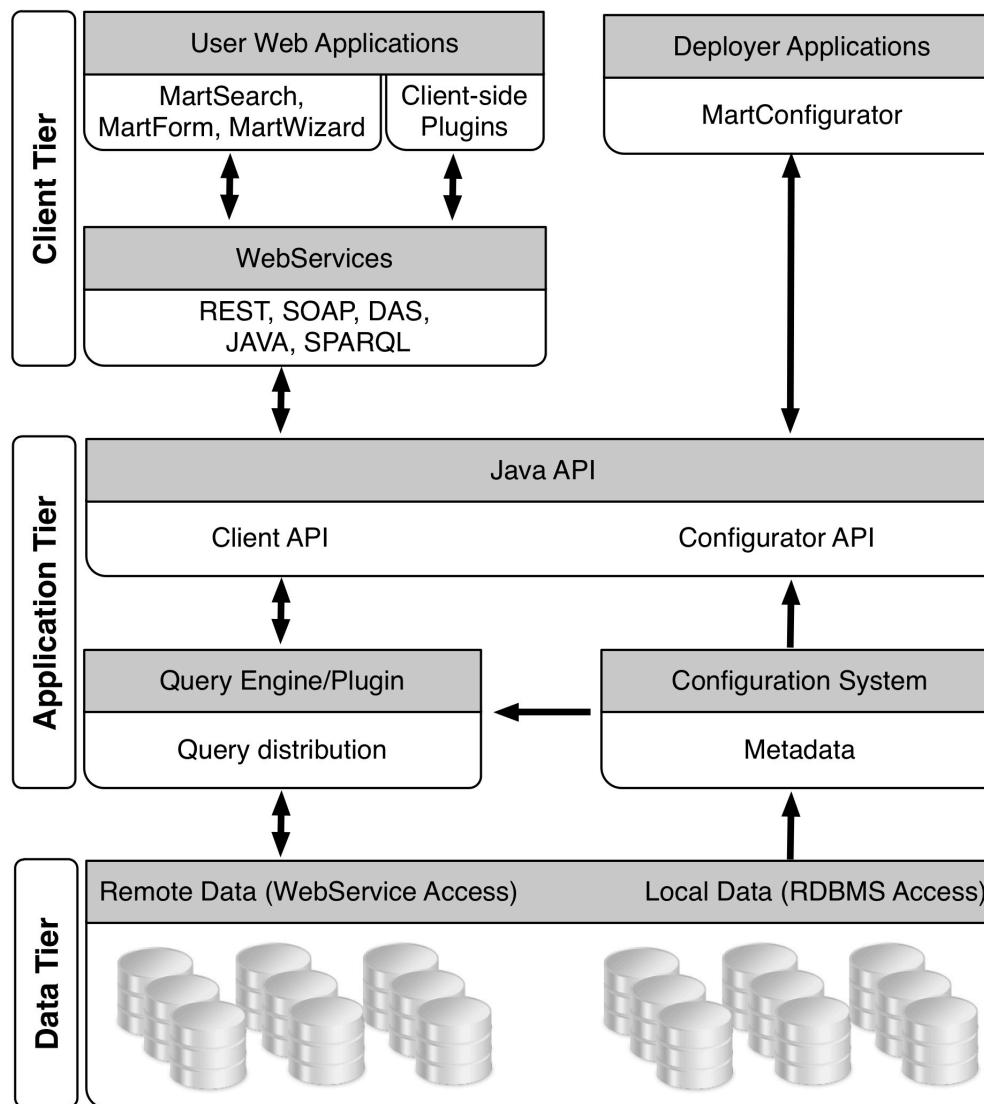
BioMart dataset is organized in a relational database schema called reverse star, optimized for fast retrieval of large quantities of descriptive data.



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



BioMart system components



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

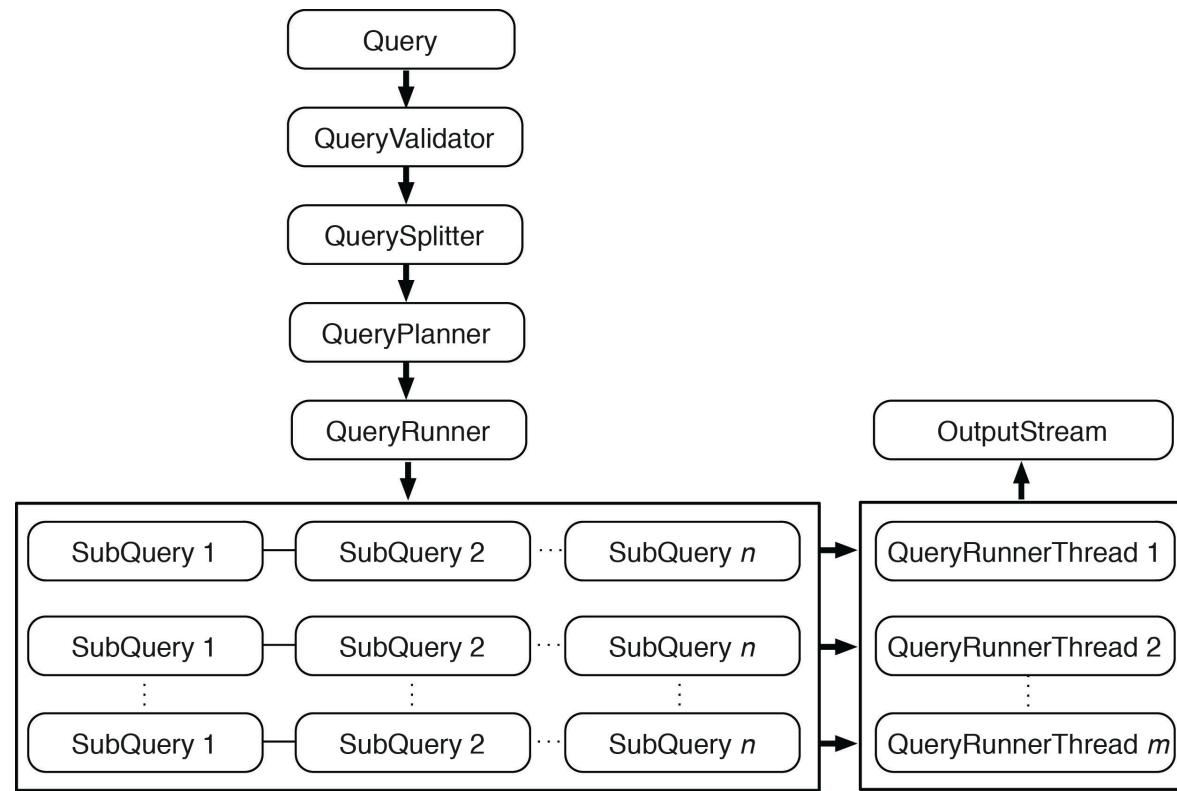
Application Tier: Stores and retrieves metadata and executes queries: sends them to the appropriate destination, receives data in response, and organizes the results before sending them back to the client tier.

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



A

Data Sources

B

Access Points

C

Link Management Dialog

D

Deploy BioMart server

Schema Editor window, showing the dataset schema

Link Management window, showing the data links between data sources

Access Point Editor window, showing attributes and filters organized in containers

BioMart 0.8 provides several querying web interfaces



MartForm

The screenshot shows the BioMart MartForm interface. At the top, there are dropdown menus for 'Database' (Ensembl Genes 61 (WTSI, UK)) and 'Datasets' (Homo sapiens genes (GRCh37.p2)). Below this is a 'FILTERS' section with fields for 'Limit to genes ...', 'Entries with following IDs ...', 'Transcript count >=:' (with a dropdown menu), 'upload file', 'Type' (a dropdown menu listing various RNA types like misc_RNA, misc_RNA_pseudogene, Mt_rRNA, Mt_tRNA, Mt_tRNA_pseudogene, polymorphic_pseudogene, processed_transcript, protein_coding, pseudogene, rRNA), 'Source' (dropdown menu), and 'Status (gene)' (dropdown menu). At the bottom is an 'ATTRIBUTES' section titled 'ENSEMBL' containing a grid of checkboxes for selecting attributes. Checked attributes include Ensembl Gene ID, Band, Description, Chromosome Name, Transcript count, and % GC content. Other options like Ensembl Transcript ID, Gene Start (bp), Strand, Transcript Start (bp), Transcript End (bp), Associated Gene DB, Associated Transcript Name, Associated Transcript DB, Transcript Biotype, Source, and Status (transcript) are also listed.

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

The screenshot shows the MartWizard interface with the following sections:

- ENSEMBL** (Left Panel): A list of checkboxes for selecting attributes from the ENSEMBL dataset. Checked items include: Ensembl Gene ID, Description, Transcript count, and % GC content. Unchecked items include: Ensembl Transcript ID, Ensembl Protein ID, Strand, Transcript Start (bp), Associated Gene Name, Associated Gene DB, Gene Biotype, Source, Status (transcript), Canonical transcript stable ID(s), Gene End (bp), Band, Transcript End (bp), Associated Transcript Name, Associated Transcript DB, Transcript Biotype, and Status (gene).
- SELECT ATTRIBUTES FROM ONE TAB** (Bottom Left Panel): A tabbed interface for selecting attributes from one tab. The "Features" tab is selected, showing sub-options: EXTERNAL, GO BIOLOGICAL PROCESS (Bp), GO Term Accession (bp), and GO Term Name (bp).
- SUMMARY** (Right Panel): A summary of the current query configuration:
 - Database:** Ensembl Genes 61 (WTSI, UK)
 - Datasets:** Homo sapiens genes (GRCh37.p2)
 - Filters:** Type: protein_coding
 - Attributes:** Ensembl Gene ID, Description, Chromosome Name, Band, Transcript count, and % GC content.

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

The screenshot shows the MartExplorer interface for the Ensembl dataset. The top navigation bar includes links for Datasets, Filters, Output, Restart, Previous, and Results. The main area is divided into three panels: a left sidebar with a tree view of filter categories, a central filter selection grid, and a right panel for viewing and managing selected filters.

Left Sidebar (Ensembl dataset structure):

- Ensembl
- Select attributes from one tab
- Features
 - EXTERNAL:
 - go biological process (bp)
 - go cellular component (cc)
 - go molecular function (mf)
 - GOSlim GOA
 - External References (max 3)
 - Microarray (max 2)
- EXPRESSION:
- PROTEIN DOMAINS:
 - Structures
 - Transcript Event
 - Homologs
 - Variation

Central Filter Selection Grid (ENSEMBL dataset):

Filter Type	Available Options	Selected Options
Ensembl Gene ID	<input type="checkbox"/> Ensembl Transcript ID	<input checked="" type="checkbox"/> Ensembl Protein ID
EXTERNAL	<input type="checkbox"/> Canonical transcript stable ID(s)	<input checked="" type="checkbox"/> Description
Band	<input type="checkbox"/> Gene Start (bp)	<input type="checkbox"/> Gene End (bp)
PROTEIN DOMAINS	<input type="checkbox"/> Associated Gene Name	<input type="checkbox"/> Transcript Start (bp)
Structures	<input type="checkbox"/> Associated Transcript DB	<input type="checkbox"/> Associated Transcript Name
Transcript Event	<input type="checkbox"/> Gene Biotype	<input checked="" type="checkbox"/> Transcript count
Homologs	<input type="checkbox"/> Status (gene)	<input type="checkbox"/> Transcript Biotype
Variation		<input checked="" type="checkbox"/> % GC content
		<input type="checkbox"/> Source
		<input type="checkbox"/> Status (transcript)

Right Panel (SUMMARY):

Category	Value
Database	Ensembl Genes 61 (WTSI, UK)
Datasets	Homo sapiens genes (GRCh37.p2)
Filters	Type: protein_coding
Attributes	Ensembl Gene ID Description Chromosome Name Band Transcript count % GC content

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.

Ensembl Genes 61 (WTSI, UK)
Displaying results 1-20 out of 1000
Results beyond 1000 are not displayed, use the download link to retrieve the complete results. Click on a column heading to sort.

Ensembl Gene ID ↑ Description ↑ Chromosome Name ↑ Band ↑ Transcript count ↑ % GC content ↑

Ensembl Gene ID	Description	Chromosome Name	Band	Transcript count	% GC content
ENSG00000136546	sodium_channel_voltage-gated_type_VII_alpha [Source:HGNC Symbol;Acc:10594]	2	p24.3	7	34.54
ENSG00000251569				1	39.33
ENSG00000239474				4	38.54
ENSG00000138399				14	38.00
ENSG00000163092				10	36.19
ENSG00000254882				1	45.00
ENSG00000250683				1	41.30
ENSG00000255228				1	44.96
ENSG00000254439				1	46.49
ENSG00000254611				1	46.49
ENSG00000234438				1	70.44
ENSG00000232371				1	46.85
ENSG00000254493				1	53.29
ENSG00000249398				1	51.30
ENSG00000254809				1	42.83
ENSG00000255061				1	53.48
ENSG00000254922				1	46.49
ENSG00000198888				1	47.70
ENSG00000198763				1	42.99
ENSG00000198804				1	46.24

1 2 3 4 5 6 7 8

REST / API Query

```
<!DOCTYPE Query>
<Query client="true" processor="TSV" limit="-1" header="1">
    <Dataset name="hsapiens_gene_ensembl">
        <Filter name="biotype" value="protein_coding"/>
        <Attribute name="ensembl_gene_id"/>
        <Attribute name="description"/>
        <Attribute name="chromosome_name"/>
        <Attribute name="band"/>
        <Attribute name="transcript_count"/>
        <Attribute name="percentage_gc_content"/>
    </Dataset>
</Query>
```

Powered by bio**mart**

Toggle quote-escape Close

Special GUI - MartReport



Ensembl Gene Info

Ensembl Gene ID: ENSG00000133703 [KRAS]

Description: v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog [Source:HGNC Symbol;Acc:6407]

Chromosome: 1q

Gene Start (bp): 25358182 Gene End (bp): 25403854

Strand: -1 Band: p12.1

Gene Biotype: protein_coding

KEGG Reactome

Pathway Title (Reactome): Insulin receptor signaling cascade, Signaling by Insulin receptor, SHC-related events, IRS-related events, Hemostasis, RAF/MAP kinase cascade, RAF activation, MEK activation, IRS-mediated signaling, RAF phosphorylates MEK, SHC-mediated signaling, SOS-mediated Signaling, Signaling by NGF, Signaling to RAS, Immune System, Prolonged ERK activation events, Frz2-mediated activation, ARMS-mediated activation, p38MAPK events, Signaling by EGFR, Grb2 events in EGFR signaling, Shc events in EGFR signaling, Down-stream signal transduction, Signaling by PDGF, NGF signaling via TrkA from the plasma membrane, Signaling to ERKs, Signaling to p38 via RIT and RIN, Cell surface interactions at the vascular wall, Tie2 Receptor NCAM signaling for neurite-induced growth, Axon guidance, Signaling by intermediate, Interleukin-2 signaling, Cytokine Signaling

COSMIC

Gene Name	Cosmic Mutation ID	CDS Mutation Type	CDS Mutation Syntax	Amino Acid Mutation Type	Substitution	Mutant
KRAS	24602	Substitution	c.154T>A	Substitution - Missense	p.R164Q	12.8
KRAS	41307	Substitution	c.491G>A	Substitution - Missense	p.R164Q	12.8
KRAS	12643	Unknown	c.7	Substitution - Missense	p.G15N	12.8
KRAS	519	Complex	c.35_36GT>AA	Substitution - Missense	p.G12E	12.8
KRAS	549	Substitution	c.181G>A	Substitution - Missense	p.Q61K	12.8
KRAS	28518	Substitution	c.176G>G	Substitution - Missense	p.A59G	12.8
KRAS	23612	Unknown	c.7	Unknown	p.?	12.8
KRAS	531	Complex - compound substitution	c.38_39GC>AT	Substitution - Missense	p.G13D	12.8
KRAS	512	Complex	c.34_35GG>TT	Substitution - Missense	p.G12F	12.8
KRAS	30566	Complex - compound substitution	c.35_36GT>AG	Substitution - Missense	p.G12E	12.8

PANCREAS EXPRESSION DATA

Fold Change Comparison Technology

Comparison	Technology
Intraductal papillary mucinous neoplasms (IPMN) / Normal pancreas ND (microdissected normal ductal cells)	Transcriptomics
Pancreatic tumor central / peripher (Xenograft from CL (orthotopic) (microdissected))	Transcriptomics
Pancreatic cancer/healthy control (Baliva) (Saliva)	Transcriptomics

PED

Data source: Pancreatic Expression Database

BREAST CANCER CAMPAIGN TISSUE BANK (BCCTB)

Fold Change Comparison Platform

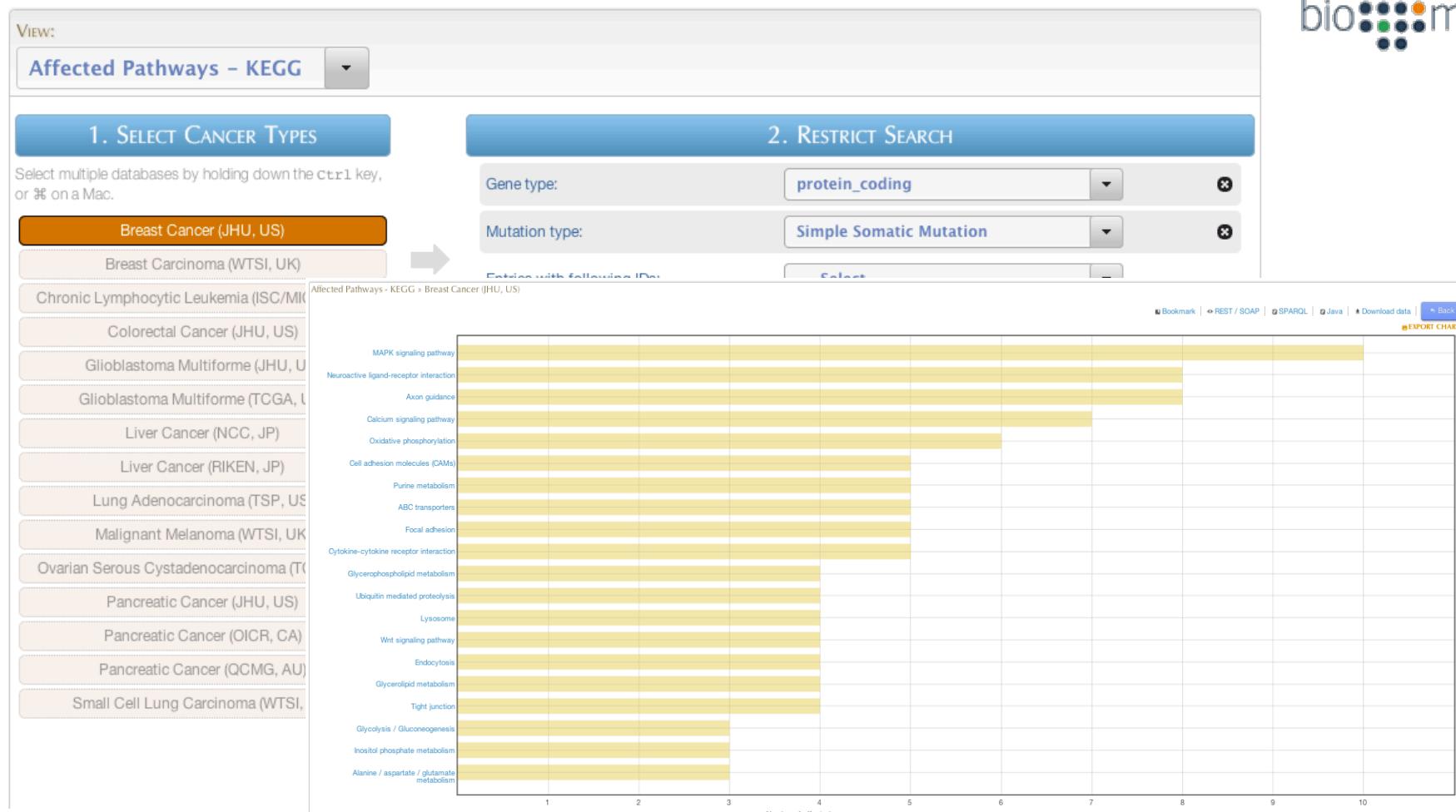
Comparison	Platform
Tumour Grade 1 vs Tumour Grade 3 (IDC DCIS) (microdissected stromal cells)	cDNA Array U133 XSP Affymetrix GeneChip; microarray

BCCTB

Data source: Breast Cancer Campaign Tissue Bank (BCCTB)

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

DATASETS

Database: Ensembl

Datasets: Homo sapiens genes (GRCh37.p2)

SEQUENCES



- Unspliced (Transcript)
- Unspliced (Gene)
- Flank (Transcript)
- Flank (Gene)
- Flank-coding region (Transcript)
- 5' UTR
- 3' UTR
- Exon Sequences
- cDNA Sequences
- Coding Sequences
- Protein

Upstream Flank: []

Downstream Flank: []

FILTERS

HEADER INFORMATION

GENE INFORMATION

Ensembl Gene ID Description Associated Gene Name
 Associated Gene DB Chromosome Name Gene Start (bp)
 Gene End (bp) Ensembl Protein Family ID(s)

TRANSCRIPT INFORMATION

CDS start (within cDNA) CDS end (within cDNA) 5' UTR Start
 5' UTR End 3' UTR Start 3' UTR End

Sequence retrieval tool is implemented as server-side analysis plugin

Large collaborative projects using BioMart for data management:

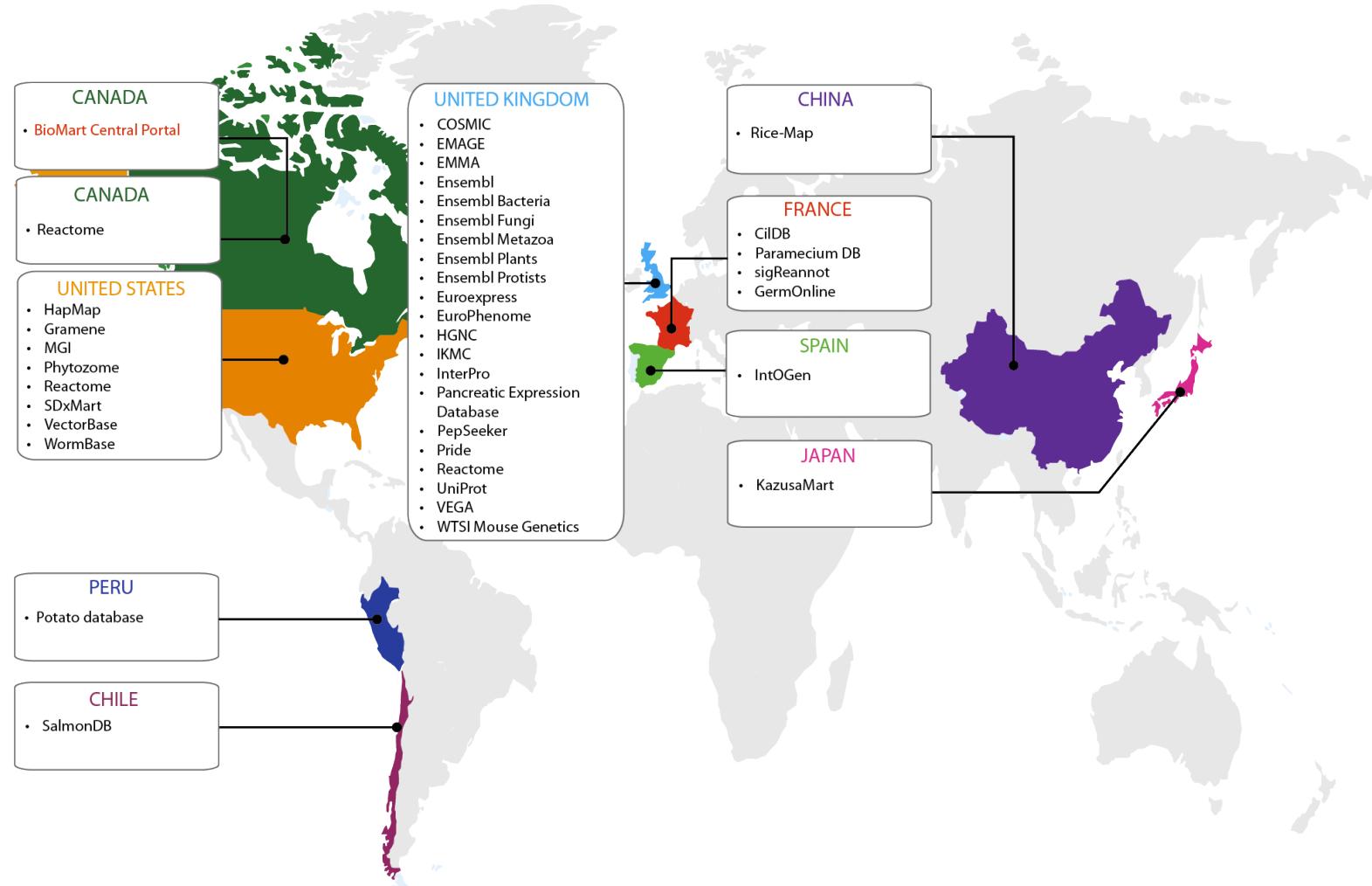


- BioMart Central Portal (<http://central.biomart.org>) (Refer to the 'BioMart Central Portal' Presentation for more details)
- International Cancer Genome Consortium (<http://dcc.icgc.org>) (Refer to the 'ICGC Data Portal Presentation' for more details.)
- PopCure (Private and public collaboration,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



IDENTIFIER SEARCH

Examples: KRAS, ENSG00000146648

TOOLS

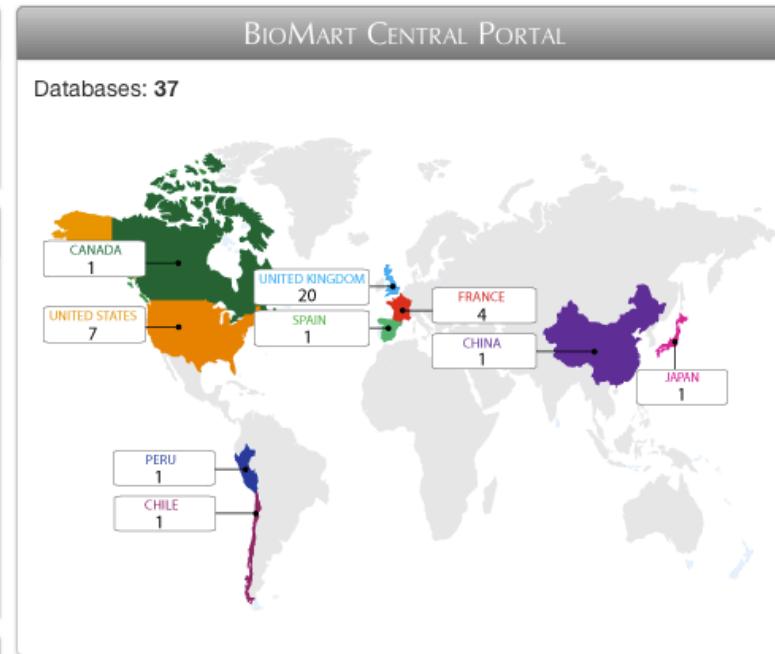
Gene retrieval **Variant retrieval** **Sequence retrieval** **ID Converter**

Cancer Genes
Ensembl
Ensembl bacteria
Ensembl fungi
Ensembl metazoa
Ensembl plants
Ensembl protists
Mouse Genome Informatics (MGI)
VEGA

DATABASE SEARCH

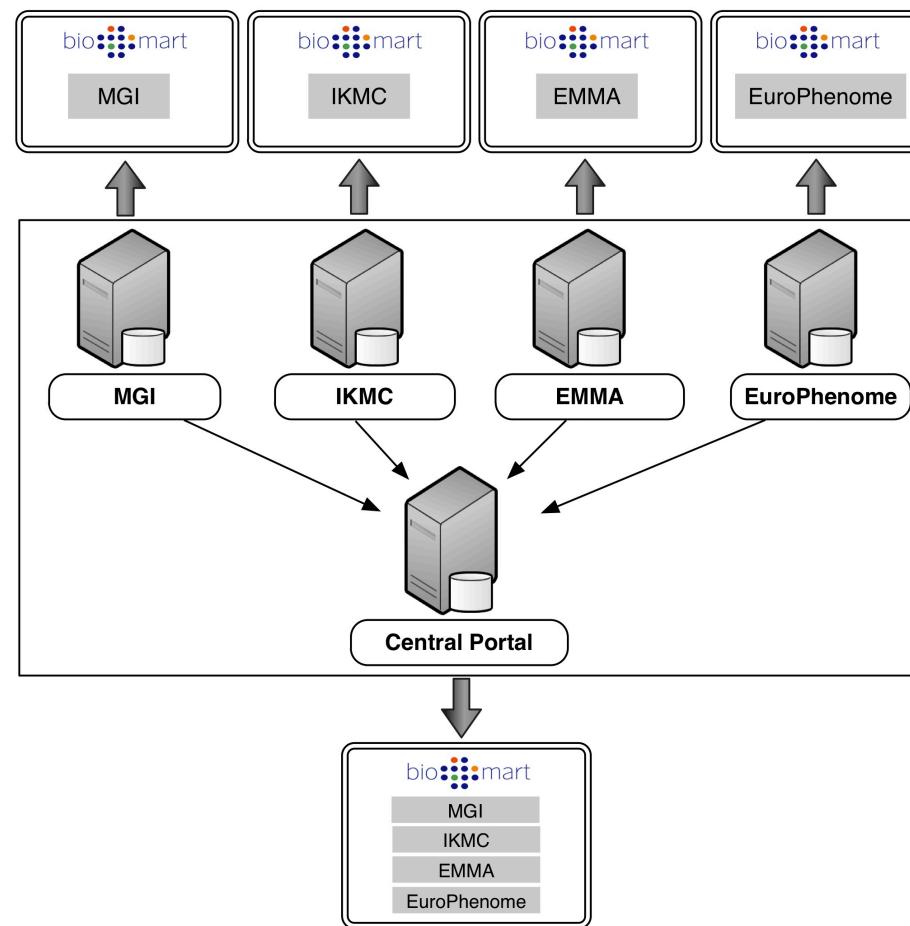
Search by type **Search by organism**

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

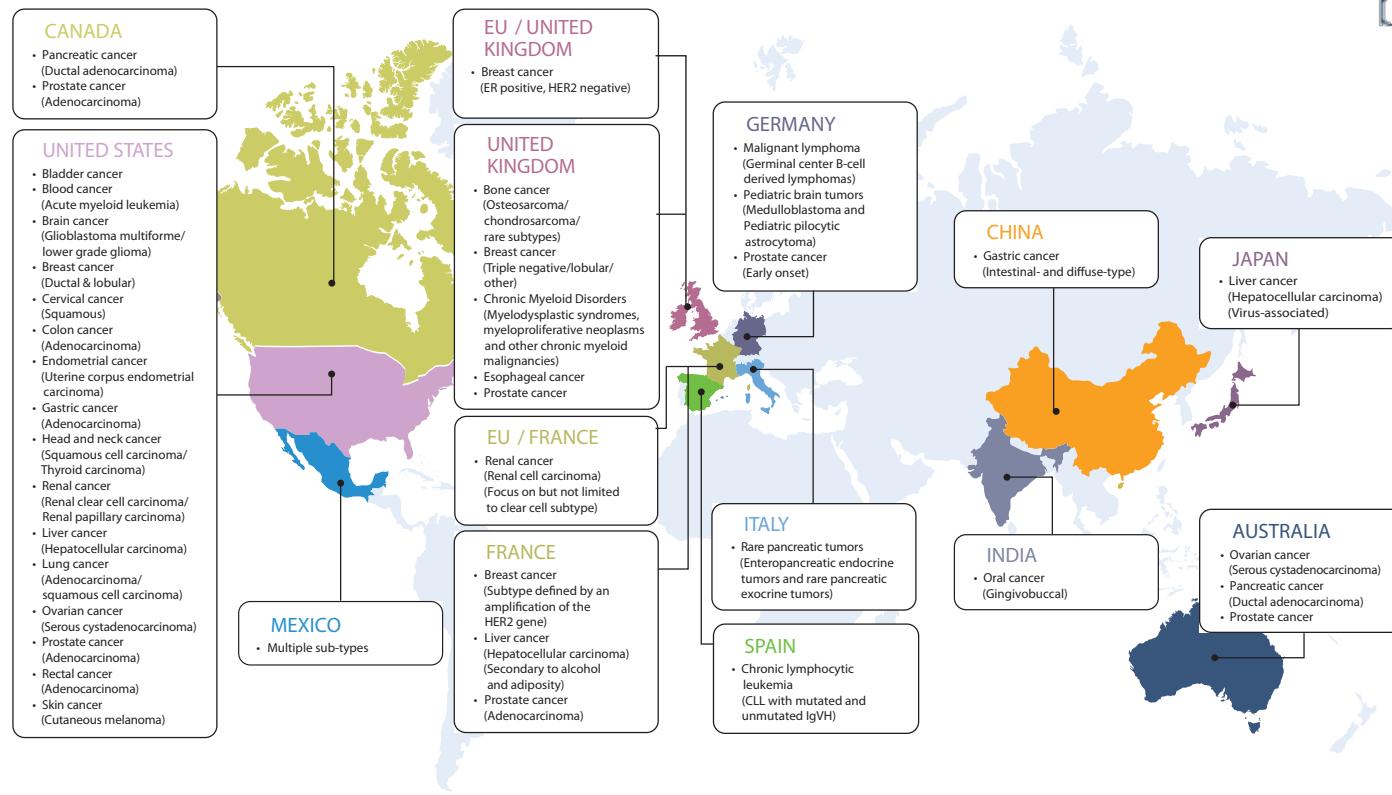


BioMart Central Portal Architecture

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



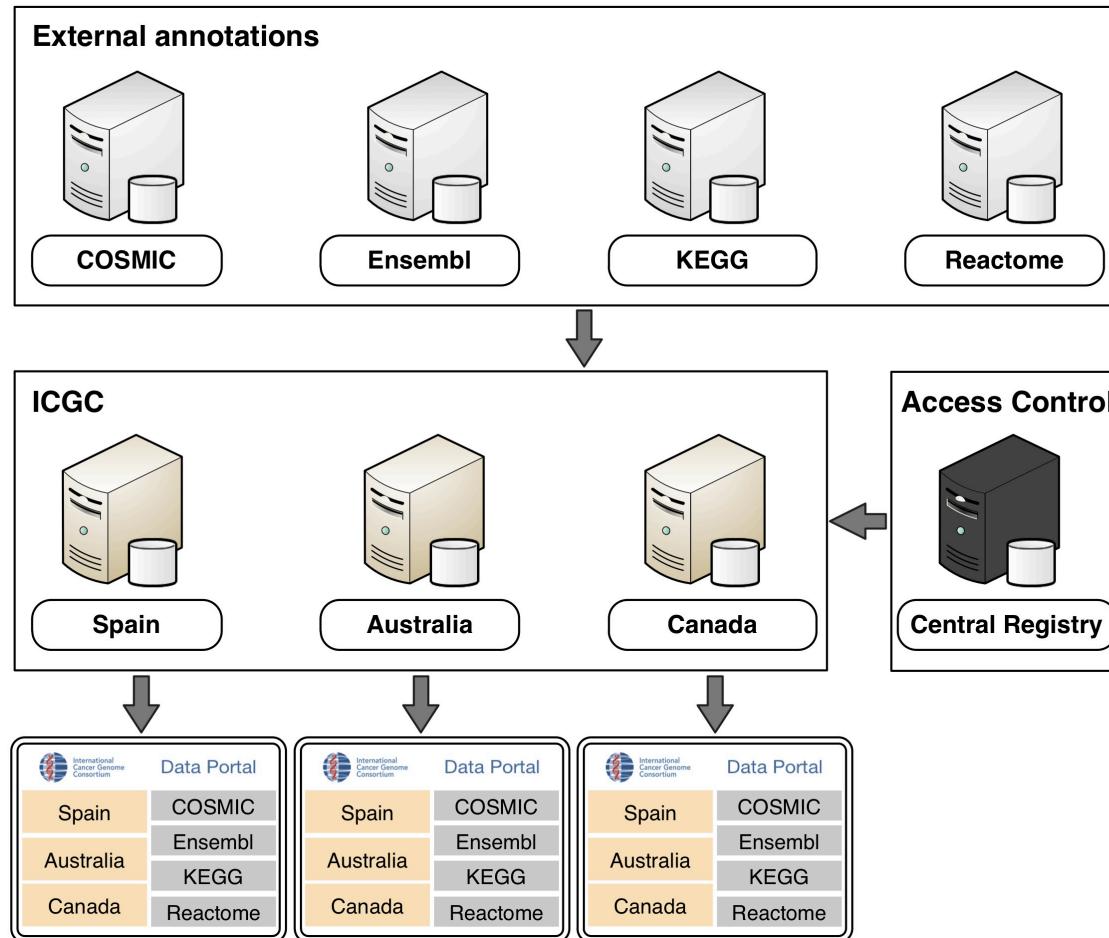
GOALS: To obtain a comprehensive description of genomic, transcriptomic, and epigenomic changes in 50 different tumor types and/or subtypes, which are of clinical and societal importance across the globe. 500 tumor and matched control samples will be analyzed per tumor type. At present, 12 countries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

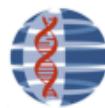
To make the data available rapidly and with minimal restrictions, to accelerate research of the causes and control of cancer.

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.





International
Cancer Genome
Consortium

Data
Portal

(dcc.icgc.org)



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

Not logged in ([Login](#)) | You are on the: [Canada website](#)

Home

IDENTIFIER SEARCH

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

Donors by Tissue

Tissue	Count
Ovary	524
Lung	292
Kidney	196
Colon	244
Breast	430
Brain	566
Blood	192
Pancreas	145
Rectum	69
Skin	1
Stomach	83
Uterus	70

Total Donors: 2,837

Powered by biomart

25



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

