



# ICGC Data Portal: One-stop shop for cancer genomics data

Elena Rivkin  
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

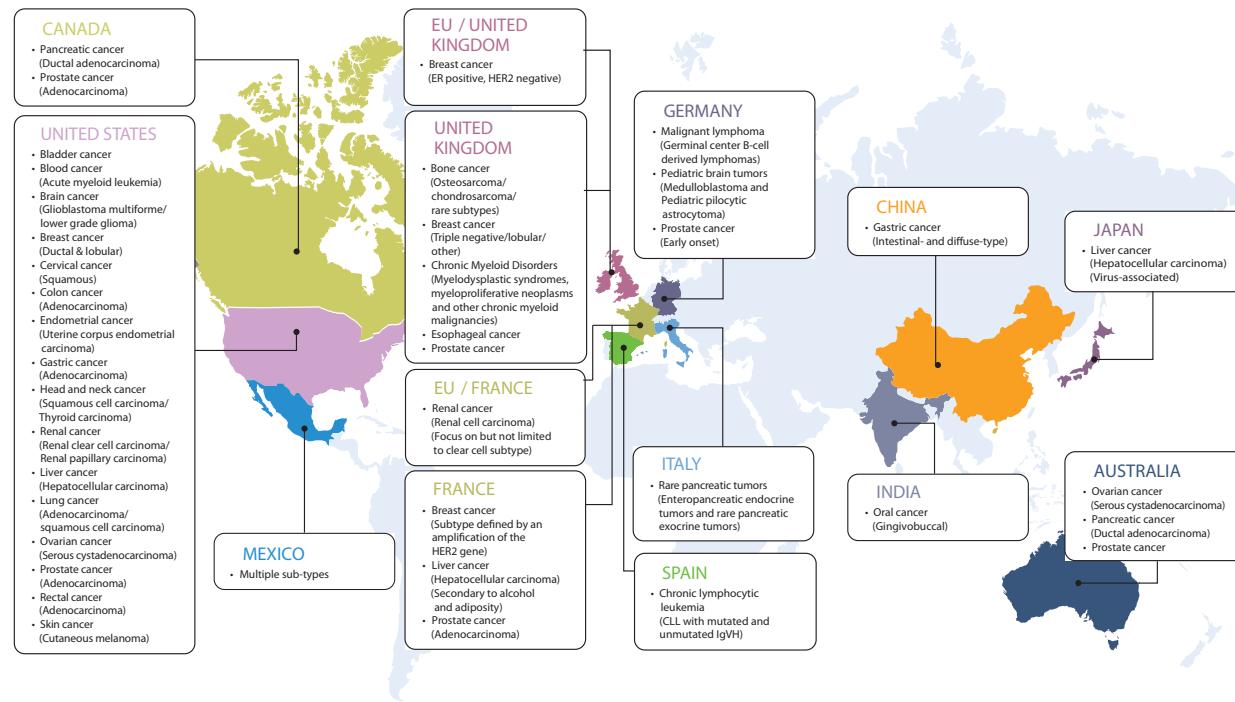
## **Key features:**

- Easy installation using a single integrated tool
- Supports data federation ideal for collaborative projects
- Optimized for large-scale data retrieval important for high-throughput cancer genomic experiments
- Variety of graphical and programmatic query interfaces
- Enterprise level security features to manage sensitive data
- Robust plug-in framework for data analysis and visualization

[www.biomart.org](http://www.biomart.org)

3

# International Cancer Genome Consortium (ICGC)



## Goals

- To obtain comprehensive description of genomic, transcriptomic, and epigenomic changes in 50 different cancer types and/or subtypes. 500 cancer and matched control samples will be analyzed per cancer type. At present, 12 countries and two European consortia have joined the effort.
- To make the data available rapidly and with minimal restrictions, to accelerate research of the causes and control of cancer.

## ICGC Data Categories

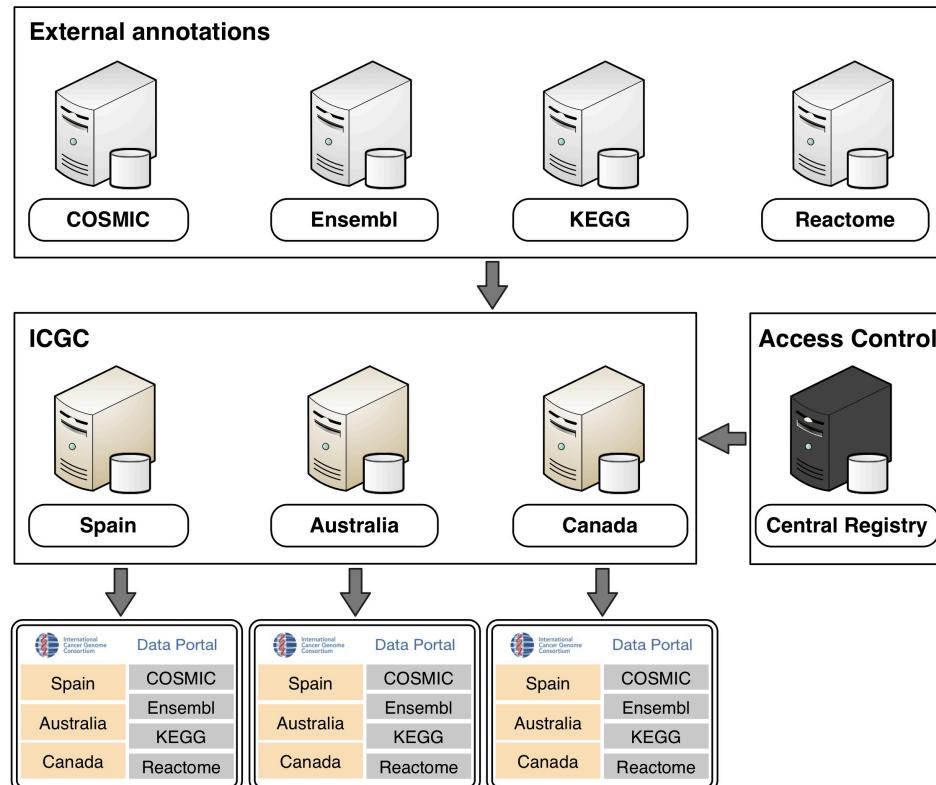


ICGC Open Access Datasets	ICGC Controlled Access Datasets
<ul style="list-style-type: none"><li>➤ Cancer Pathology<ul style="list-style-type: none"><li>Histologic type or subtype</li><li>Histologic nuclear grade</li></ul></li><li>➤ Donor<ul style="list-style-type: none"><li>Gender</li><li>Age range</li></ul></li><li>➤ RNA expression (normalized)</li><li>➤ DNA methylation</li><li>➤ Genotype frequencies</li><li>➤ Somatic mutations (SNV, CNV and Structural Rearrangement)</li></ul>	<ul style="list-style-type: none"><li>➤ Detailed Phenotype and Outcome Data<ul style="list-style-type: none"><li>Patient demography</li><li>Risk factors</li><li>Examination</li><li>Surgery/Drugs/Radiation</li><li>Sample/Slide</li><li>Specific histological features</li><li>Protocol</li><li>Analyte/Aliquot</li></ul></li><li>➤ Gene Expression (probe-level data)</li><li>➤ Raw genotype calls (germline)</li><li>➤ Gene-sample identifier links</li><li>➤ Genome sequence files</li></ul>

Most of the data in the portal is publically available without restriction. However, access to some data, like the germline mutations, requires authorization by the Data Access Compliance Office (DACO)

## 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.



Each ICGC member deposits its site-specific data into its local BioMart database. These data are integrated with diverse annotations from external sources (Ensembl, Reactome, EKGG, COSMIC, Pancreatic Expression database, Breast Cancer Campaign Tissue Bank). Finally, all data are exposed through the ICGC Data Portal. To the user, the multiple databases appear as a single integrated database.

- Multi-portal deployment, no transferring of large data, flexible update
- Shared data models across all ICGC member sites
- Seamlessly federate ICGC data sources as well as external annotations
- HTTPS and OAuth secured server-to-server communication; OpenID based user authentication

# The power of data federation: cross-dataset query against distributed data sources



Query: Give me all the genes that are mutated with non-synonymous coding consequence and are involved in Wnt signaling pathway from all pancreatic cancer datasets

The screenshot illustrates the biomart interface for performing a cross-dataset query. The process is divided into two main steps:

- 1. SELECT CANCER TYPES**: This step involves selecting specific cancer types from a list. A red box highlights the "Pancreatic Cancer" entries, which are further subdivided into three specific datasets: JHU, CA, and QCMG.
- 2. RESTRICT SEARCH**: This step involves specifying search parameters. The "Pathway name" is set to "Signaling by Wnt". The "Gene type" dropdown is set to "non\_synonymous\_coding". The "Copy number alteration type" dropdown is set to "No data". The "Methylation beta value" dropdown is set to "No data". The "Clinical staging (WHO)" dropdown is set to "Select". The "Entries with following IDs" dropdown is set to "Select".

Annotations on the right side of the interface map specific datasets to their corresponding search parameters:

- KEGG**: Associated with the "Pathway name" field.
- Ensembl**: Associated with the "Gene type", "Copy number alteration type", and "Entries with following IDs" fields.
- ICGC**: Associated with the "Methylation beta value" and "Clinical staging (WHO)" fields.
- upload file**: Associated with the "Entries with following IDs" field.

A large green button at the bottom center is labeled "Go »". A callout box on the right provides additional context: "In addition to experimental cancer genomics data, ICGC Data Portal federates diverse annotations from external sources, allowing users to build integrated queries containing genomic, clinical and functional information."



Data  
Portal

# Home Page (dcc.icgc.org)



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

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

**IDENTIFIER SEARCH**

Lets users input gene identifiers and return links to the corresponding Gene Report

Examples: TP53, ENSG00000133703, NM\_000314

**ANALYSIS**

**Genes** **Pathway**

Affected Genes

Analysis tools allow users to view commonly affected genes or pathways

**DATABASE SEARCH**

**Quick** **Flexible** **Advanced**

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

Query the data using one of several data entry points through Quick, Flexible, or Advanced search interface

**Donors by Tissue**

Total Donors: 2,837

Ovary: 524  
Lung: 292  
Liver: 25  
Kidney: 196  
Colon: 244  
Breast: 430  
Blood: 192  
Pancreas: 145  
Rectum: 69  
Skin: 1  
Stomach: 83  
Uterus: 70

Login to view controlled datasets

Query data through the portal hosted in any of the hosting countries

Click on the pie chart to see a detailed dataset summary

Powered by biomart



# ICGC Data Portal Dataset Summary Table



Source	Cancer Project	Tissue	Dataset									
			Donors	Samples	Simple Mutations	CNV	Structural Rearrangements	Gene Expression	miRNA Expression	Exon Junction	DNA Methylation	Germline Variations*
ICGC	Acute Myeloid Leukemia (TCGA, US)	Blood	188	188	-	-	-	-	-	-	188	-
	Chronic Lymphocytic Leukemia (ISC/MICINN, ES) <sup>6</sup>	Blood	4	4	4	4	-	-	-	-	-	-
	Glioblastoma Multiforme (TCGA, US) <sup>1</sup>	Brain	460	465	-	-	-	461	405	-	268	-
	Breast Carcinoma (WTSI, UK)	Breast	24	24	-	-	24	-	-	-	-	-
	Breast Invasive Carcinoma (TCGA, US)	Breast	358	350	-	-	-	350	-	-	183	-
	Colon Adenocarcinoma (TCGA, US)	Colon	207	167	-	-	-	154	-	-	166	-
	Kidney Renal Clear Cell Carcinoma (TCGA, US)	Kidney	179	90	-	-	-	57	-	-	89	-
	Kidney Renal Papillary Cell Carcinoma (TCGA, US)	Kidney	17	22	-	-	-	-	-	-	22	-
	Liver Cancer (NCC, JP)	Liver	11	11	11	-	1	-	-	-	-	-
	Liver Cancer (RIKEN, JP)	Liver	14	16	16	-	1	-	-	-	-	-
	Lung Adenocarcinoma (TCGA, US)	Lung	44	44	-	-	-	25	-	-	44	-
	Lung Squamous Cell Carcinoma (TCGA, US)	Lung	59	59	-	-	-	59	-	-	58	-
	Small Cell Lung Carcinoma (WTSI, UK) <sup>5</sup>	Lung	1	1	1	-	1	-	-	-	-	-
	Ovarian Serous Cystadenocarcinoma (TCGA, US)	Ovary	524	509	-	-	-	507	487	-	418	-
	Pancreatic Cancer (OICR, CA)	Pancreas	26	43	43	10	-	-	-	-	-	5
	Pancreatic Cancer (QCMG, AU)	Pancreas	5	9	3	3	3	3	6	3	-	3
	Rectum Adenocarcinoma (TCGA, US)	Rectum	69	69	-	-	-	69	-	-	69	-
	Malignant Melanoma (WTSI, UK)	Skin	1	1	-	-	1	-	-	-	-	-
	Stomach Adenocarcinoma (TCGA, US)		42	-	-	-	-	-	-	-	142	-
	Uterine Corpus Endometrioid Carci		-	-	-	-	-	-	-	-	70	-
Other	Glioblastoma Multiforme (JHU, US) <sup>4</sup>		89	22	-	-	-	-	-	-	-	-
	Breast Cancer (JHU, US) <sup>7,8</sup>		42	-	-	-	-	-	-	-	-	-
	Colorectal Cancer (JHU, US) <sup>7,8</sup>		36	-	-	-	-	-	-	-	-	-
	Lung Adenocarcinoma (TSP, US) <sup>2</sup>		163	-	-	-	-	-	-	-	-	-
	Pancreatic Cancer (JHU, US) <sup>3</sup>	Pancreas	114	114	24	-	-	-	-	-	-	-

Clicking on the number links  
to the relevant query  
interface (e.g. Samples)

# Identifier Search – Type in a gene identifier and link to the corresponding Gene Report



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

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

### 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 5 (JUNE 3RD, 2011)

Cancer Projects: 25

Samples by Tissue

Tissue	Count
Ovary	546
Lung	480
Kidney	442
Colon	223
Breast	438
Brain	468
Blood	192
Uterus	71
Stomach	142
Rectum	73
Skin	1
Pancreas	126
Liver	2

Total Samples: 3,204

# Gene Report



**GENE INFO**

**Ensembl Gene ID: ENSG00000133703 [KRAS]**

Description:	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog [Source:HGNC Symbol;Acc:6407]	Chromosome:	12
Gene Start (bp):	25358182	Gene End (bp):	25403854
Strand:	-1	Band:	p12.1
Gene Biotype:	protein_coding		

**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 phosphotyrosine 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, cSMAPK 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

**Dataset**

	Simple somatic mutation	Copy number alteration	Structural Rearrangement
Acute Myeloid Leukemia (TCGA, US)	no data	no data	no data
Breast Cancer (UHL, US)	no data	no data	no data
Breast Carcinoma (WT3, UK)	no data	no data	no data
Breast Invasive Carcinoma (TCGA, US)	no data	no data	no data
Chronic Lymphocytic Leukemia (ISCM/CINHN, ES)	no data	no data	no data
Codon-Activating Cell (TCGA, US)	44.44% (16/36)	no data	no data
Colon Cancer (UHL, US)	0.68% (1/147)	no data	no data
Glioblastoma Multiforme (UHL, US)	0.68% (1/147)	no data	no data
Glioblastoma Multiforme (TCGA, US)	0.68% (1/147)	no data	no data
Kidney Renal Clear Cell Carcinoma (TCGA, US)	no data	no data	no data
Kidney Renal Papillary Cell Carcinoma (TCGA, US)	no data	no data	no data
Liver Cancer (NCC, JP)	no data	no data	no data
Liver Cancer (RIKEN, JP)	no data	no data	no data
Lung Adenocarcinoma (TCGA, US)	36.81% (60/163)	no data	no data
Lung Adenocarcinoma (TSP, US)	no data	no data	no data
Lung Squamous Cell Carcinoma (TCGA, US)	no data	no data	no data
Malignant Melanoma (WT3, UK)	100.00% (1/1)	no data	no data
Ovarian Serous Cystadenocarcinoma (TCGA, US)	0.54% (1/185)	no data	no data
Pancreatic Cancer (UHL, US)	55.12% (13/14)	4.17% (1/24)	no data
Pancreatic Cancer (OICR, CA)	40.00% (2/5)	no data	no data
Pancreatic Cancer (COMG, AU)	100.00% (3/3)	no data	no data
Rectum Adenocarcinoma (TCGA, US)	no data	no data	no data
Small Cell Lung Carcinoma (WTSI, UK)	no data	no data	no data

**COSMIC**

Gene Name	Cosmic Mutation ID	CDS Mutation Type	CDS Mutation Syntax	Amino Acid Mutation Type	Amino Acid Mutation Syntax	GRCh37 Cox
KRAS	24802	Substitution	c.15a>T	Substitution - Missense	p.K5N	12.2
KRAS	41307	Substitution	c.491G>A	Substitution - Missense	p.R164Q	12.2
KRAS	12843	Unknown	c.7	Substitution - Missense	p.G13N	12.2
KRAS	819	Complex	c.35_36GT>AA	Substitution - Missense	p.Q12E	12.2
KRAS	549	Substitution	c.181C>A	Substitution - Missense	p.Q61K	12.2
KRAS	28518	Substitution	c.176G>Q	Substitution - Missense	p.A59Q	12.2
KRAS	23812	Unknown	c.7	Unknown	p.?	12.2
KRAS	531	Complex - compound substitution	c.38_39G>AT	Substitution - Missense	p.G13D	12.2
KRAS	512	Complex	c.34_35G>TT	Substitution - Missense	p.G12F	12.2
KRAS	30566	Complex - compound substitution	c.35_36GT>AG	Substitution - Missense	p.Q12E	12.2

**PANCREAS EXPRESSION DATA**

Fold Change: Intraductal papillary mucinous neoplasms (IPMN) / Transcriptomics Normal pancreas ND (microdissected normal ductal cells) Comparison: Gene Expression Profiling Identifies Genes Associated with Invasive Intraductal Papillary Mucinous Neoplasms of the Pancreas Technology: Transcriptomics Publication: Zonal Heterogeneity for Gene Expression in Human Pancreatic Carcinoma

Fold Change: Pancreatic tumor central / peripheral (Xenograft from CL (orthotopic)) (microdissected) Comparison: Transcriptomics Pancreatic cancer/Healthy control (Saliva) Technology: Transcriptomics Publication: Salivary Transcriptomic Biomarkers for Detection of Resectable Pancreatic Cancer

Data source: Pancreatic Expression Database

**BREAST CANCER CAMPAIGN TISSUE BANK (BCCTB)**

Fold Change: Tumour Grade 1 vs Tumour Grade 3 (IDC DCIS) (microdissected stromal cells) Comparison: cDNA Array/U133 XSP Affymetrix GeneChip Technology: microenvironment during breast cancer progression Publication: Gene expression profiling of the

Data source: Breast Cancer Campaign Tissue Bank (BCCTB)

Ensembl

KEGG  
Reactome

*Mutation frequencies from cancer projects with data distributed around the globe*

Displays data related to a single identifier, federated from multiple data sources

COSMIC

Pancreatic Expression Database (PED)

Breast Cancer Campaign Tissue Bank (BCCTB)

## Analysis – Identify genes or pathways commonly affected by somatic mutations.



**International Cancer Genome Consortium Data Portal**

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

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 1ST, 2011)**

Cancer Projects: 25

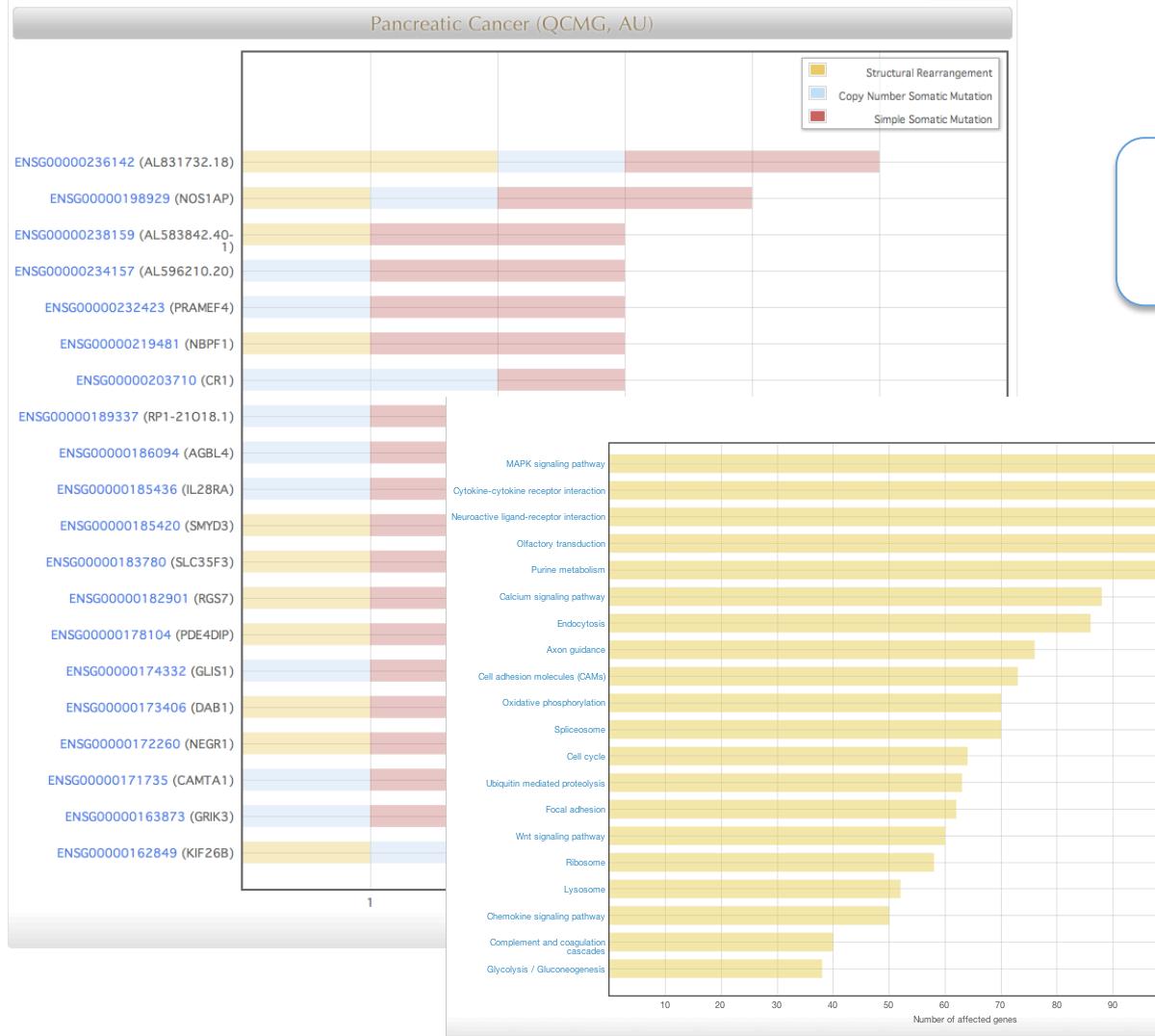
Donors by Tissue

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

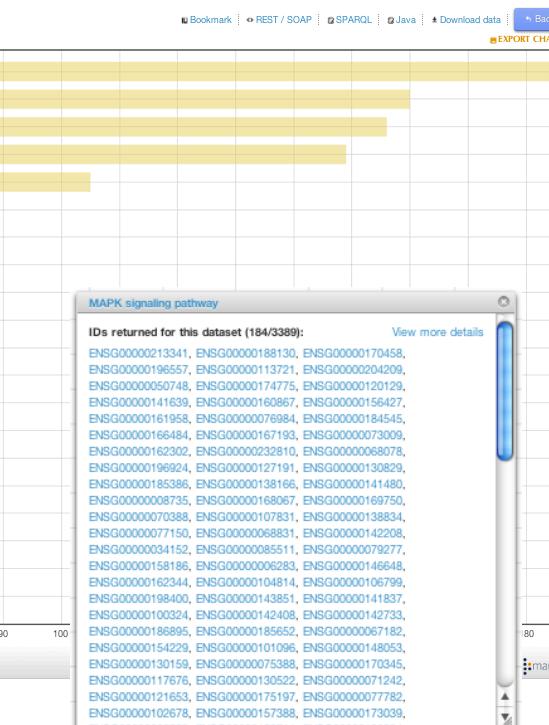
Total Donors: 2,837



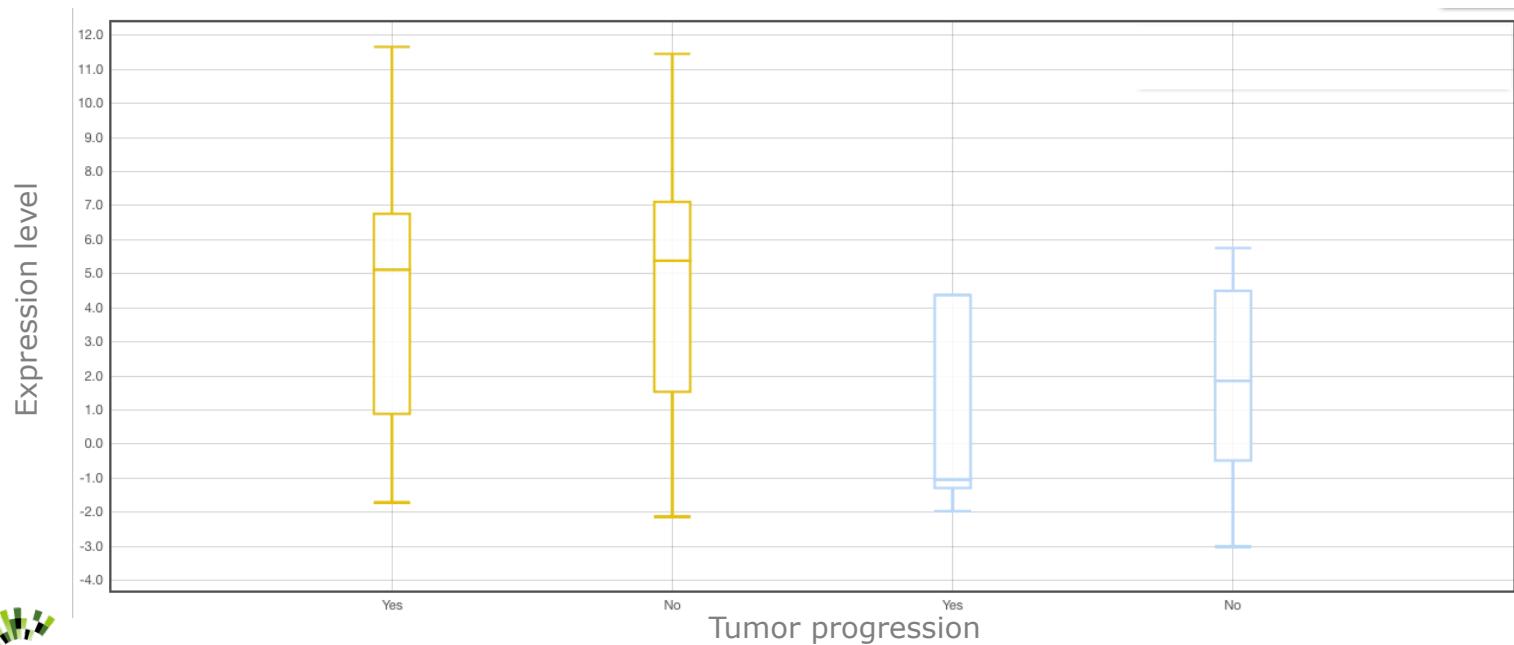
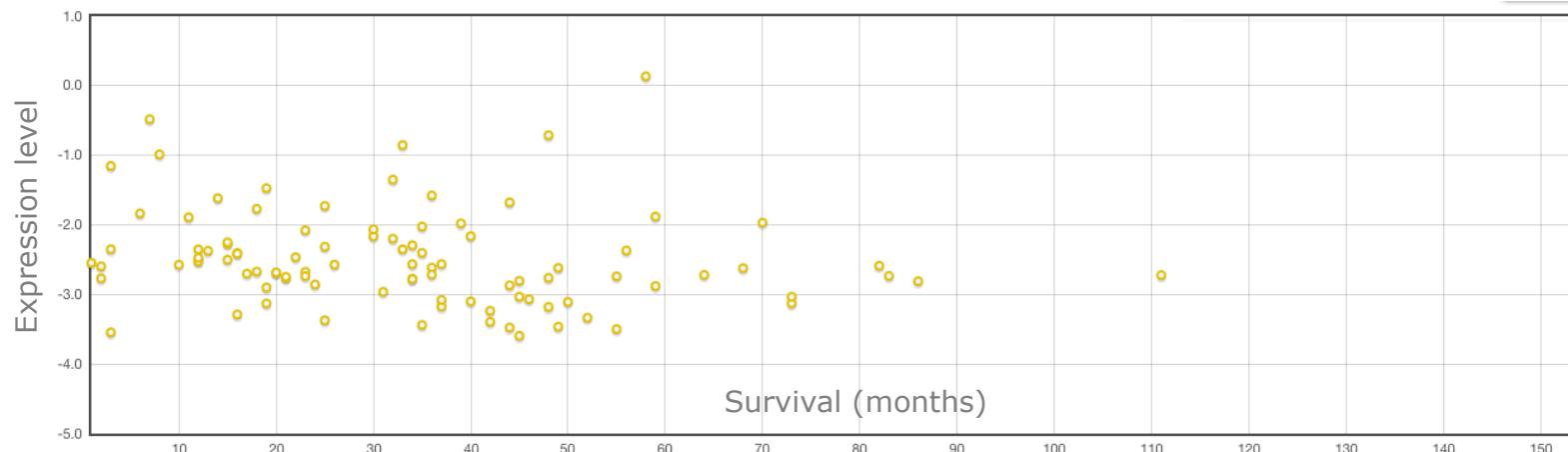
## Most frequently affected genes or pathways



Data is processed by  
BioMart Plugins



Other visualization and analysis plugins are under development.   
E.g. Association analysis – scatter plot, box plot



# Database Search



International  
Cancer Genome  
Consortium

Data  
Portal



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

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

**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 1ST, 2011)**

Cancer Projects: 25

**Donors by Tissue**

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

Total Donors: 2,837



## Database Search



All queries in the Database Search consists of these few basic steps:

1. Select entry point (based on the central item you are querying for) (e.g. Genes)
2. Select Query interface: Quick, Flexible, or Advanced.
3. Select cancer type(s)/project (e.g. Pancreatic cancer, OICR, Canada)
4. Select *Filters* (Optional) to restrict your query (Chromosome 1, or enter Gene ID of interest)
5. Select data elements (*Attributes*) that you would like to receive (ex. Gene ID, gene Type, normalized expression)

## Quick Search

VIEW: **Genes**

**1. SELECT CANCER TYPES**

Select multiple databases by holding down the **ctrl** key, or **⌘** on a Mac.

- Kidney Renal Clear Cell Carcinoma (TCGA, US)
- Kidney Renal Papillary Cell Carcinoma (TCGA, US)
- Liver Cancer (NCC, JP)
- Liver Cancer (RIKEN, JP)
- Lung Adenocarcinoma (TCGA, US)
- Lung Adenocarcinoma (TSP, US)
- Lung Squamous Cell Carcinoma (TCGA, US)
- Malignant Melanoma (WTSI, UK)
- Ovarian Serous Cystadenocarcinoma (TCGA, US)
- Pancreatic Cancer (JHU, US)
- Pancreatic Cancer (OICR, CA)
- Pancreatic Cancer (QCMG, AU)
- Rectum Adenocarcinoma (TCGA, US)
- Small Cell Lung Carcinoma (WTSI, UK)
- Stomach Adenocarcinoma (TCGA, US)

**2. RESTRICT SEARCH**

Pathway name: **-- Select --**

Gene type: **-- Select --**

Simple mutation consequence type: **-- Select --**

Copy number alteration type: **-- Select --**

Methylation beta value: **No data**

Entries with following IDs:

**Go »**

A selection of common filters; predefined list of attributes

## Flexible Search

**FLEXIBLE**

**DATASETS**

Database: Genes

Datasets:

- Acute Myeloid Leukemia (TCGA, US)
- Breast Cancer (JHU, US)
- Breast Carcinoma (WTSI, UK)
- Breast Carcinoma (TCGA, US)
- Colon Adenocarcinoma (TCGA, US)
- Colorectal Cancer (JHU, US)
- Glioblastoma Multiforme (JHU, US)
- Glioblastoma Multiforme (TCGA, US)
- Kidney Renal Clear Cell Carcinoma (TCGA, US)
- Liver Cancer (NCC, JP)

**FILTERS**

**GENE**

Entries with following IDs:

Gene type:

**PATHWAYS**

Pathway:

**METHYLATION**

Methylation beta value:

**ATTRIBUTES**

**GENE**

Ensembl Gene ID    Ensembl Transcript ID    Gene Symbol  
 Gene Biotype

**EXPERIMENT DATA**

**METHYLATION**

Tumour sample ID    Methylation beta value 1    Methylation beta value 2  
 Chromosome    Chromosome start    Chromosome end  
 Strand

A selection of filters and attributes

## Advanced Search

Cancer Types → Filters → Output ⏪ Restart ⏴ Previous Results

**ATTRIBUTES**

**GENERAL**

Cancer Type     Assembly Version

**GENE**

<input checked="" type="checkbox"/> Ensembl Gene ID	<input type="checkbox"/> Ensembl Transcript ID
<input type="checkbox"/> Ensembl Protein ID	<input type="checkbox"/> Canonical transcript stable ID(s)
<input type="checkbox"/> Gene Symbol	<input type="checkbox"/> Description
<input checked="" type="checkbox"/> Chromosome Name	<input checked="" type="checkbox"/> Gene Start (bp)
<input checked="" type="checkbox"/> Gene End (bp)	<input type="checkbox"/> Strand
<input type="checkbox"/> Band	<input type="checkbox"/> Transcript Start (bp)
<input type="checkbox"/> Transcript End (bp)	<input type="checkbox"/> Associated Transcript Name

**SUMMARY** ⏴ view XML

**View**  
Genes

**Cancer Types**  
Chronic Lymphocytic Leukemia (ISC/MICINN, ES)

**Filters**

- Chromosome name: 3 ✖
- Status (gene): KNOWN ✖
- Gene type: miRNA ✖

**Attributes**

- Cancer Type ✖
- Assembly Version ✖
- Ensembl Gene ID ✖
- Chromosome Name ✖
- Gene Start (bp) ✖
- Gene End (bp) ✖
- Mutation ID ✖

Powered by bioMart

- Complete set of available filters and attributes
- Attributes determine the order of columns in the results. Drag and drop to reorder, click on the x to remove.

## Results Page

Genes » Pancreatic Cancer (JHU, US), Pancreatic Cancer (QCMG, AC)

Queries can be bookmarked to be reused or shared

Bookmark REST / SOAP SPARQL Java Download data Back

Cancer Type	Assembly Version	Ensembl Gene ID	Gene Symbol	Description	Chromosome	Gene Start (bp)	Gene End (bp)	Strand
Pancreatic Cancer (OICR, CA)	GRCh37	<a href="#">ENSG00000113575</a>	PPP2CA	protein phosphatase 2, catalytic subunit, alpha isozyme [Source:HGNC Symbol;Acc:9302]	5	133530025	13561922	-1
Pancreatic Cancer (JHU, US)	NCBI36	<a href="#">ENSG00000105568</a>	PPP2R1A	protein phosphatase 2, regulatory subunit A, alpha [Source:HGNC Symbol;Acc:9302]	19	526932	528000	1
Pancreatic Cancer (OICR, CA)	GRCh37	<a href="#">ENSG00000066027</a>	PPP2R5A	protein phosphatase 2, regulatory subunit B', alpha [Source:HGNC Symbol;Acc:9309]	1	212458879	212535200	1
Pancreatic Cancer (OICR, CA)	GRCh37	<a href="#">ENSG00000108294</a>	PSMB3	proteasome (prosome, macropain) subunit, beta type, 3 [Source:HGNC Symbol;Acc:9540]	17	36909003	36920483	1
Pancreatic Cancer (OICR, CA)	GRCh37	<a href="#">ENSG00000013275</a>	PSMC4	proteasome (prosome, macropain) 26S subunit, ATPase, 4 [Source:HGNC Symbol;Acc:9551]	19	40477073	40487351	1

Sort results by clicking on the black triangle toggle icon

Download data in tab-delimited tabular format

Ensembl Gene ID links to the Gene Report

Many records link to appropriate resources

## 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.

**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](#)

## Conclusions

- ICGC Data Portal demonstrates how BioMart can scale to manage large collaborative projects involving next generation sequencing data
- Data federation increases scalability and flexibility, and fully takes advantage of expert-curated and independently maintained external annotation databases, to enhance interpretability of experimental data
- A variety of interfaces (graphical and programmatic) provide numerous query options, even allowing to build integrated queries containing genomic, clinical and functional information.

## Acknowledgements



Joachim Baran  
Anthony Cros  
Jonathan Guberman  
Jack Hsu  
Yong Liang  
Long Yao  
Elena Rivkin  
Brett Whitty  
Marie Wong-Erasmus  
Christina Yung  
Jianxin Wang  
Gnaneshan Saravanamuttu  
Syed Haider  
Arek Kasprzyk

*Website: [dcc.icgc.org](http://dcc.icgc.org)*  
*Mailing list: [users@biomart.org](mailto:users@biomart.org)*  
*ICGC Data Portal: [dcc-support@lists.oicr.on.ca](mailto:dcc-support@lists.oicr.on.ca)*



MINISTRY OF RESEARCH AND INNOVATION

