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?

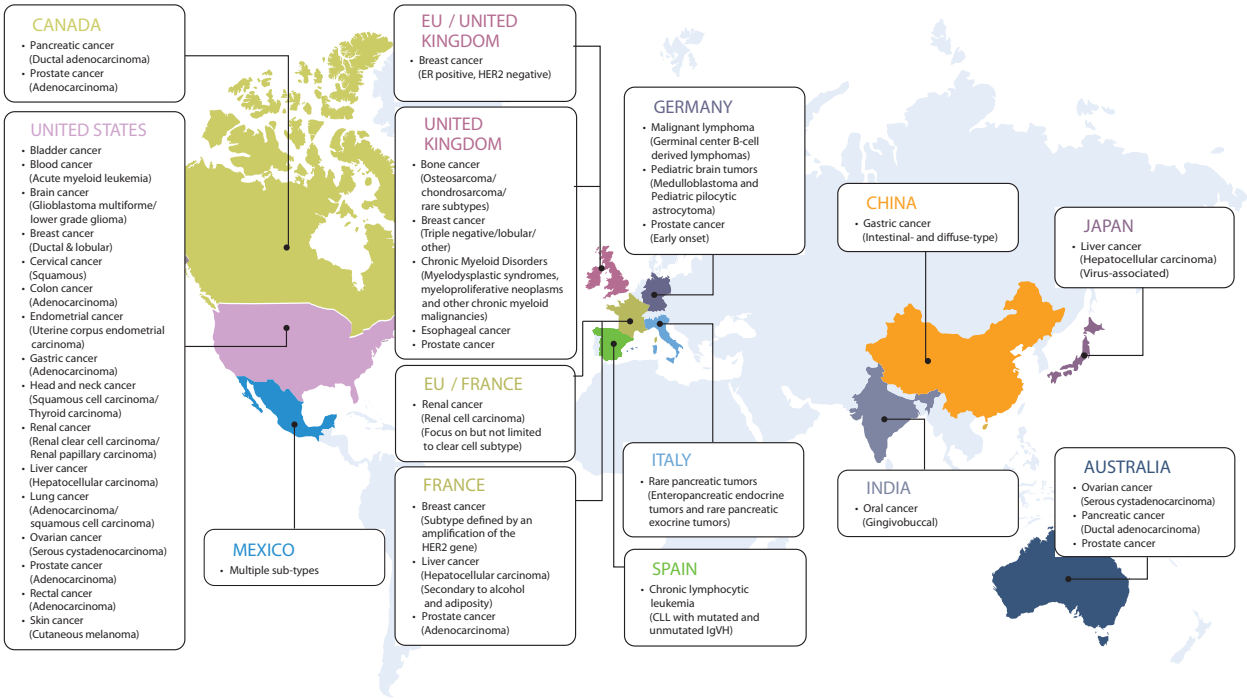
Free, open-source federated data management system that makes it possible to make distributed biological data accessible to the research community through a unified user interface

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

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.

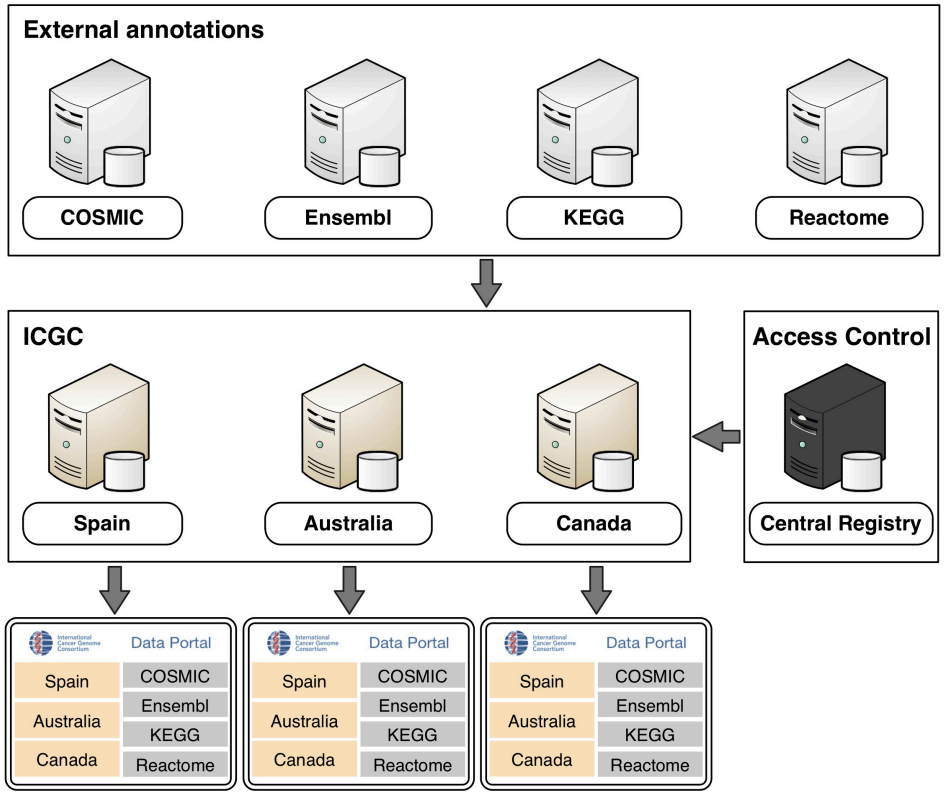
ICGC Data Categories

ICGC Open Access Datasets	ICGC Controlled Access Datasets
➢ Cancer Pathology Histologic type or subtype Histologic nuclear grade ➢ Donor Gender Age range ➢ RNA expression (normalized) ➢ DNA methylation ➢ Genotype frequencies ➢ Somatic mutations (SNV, CNV and Structural Rearrangement)	➢Detailed Phenotype and Outcome Data Patient demography Risk factors Examination Surgery/Drugs/Radiation Sample/Slide Specific histological features Protocol Analyte/Aliquot ➢Gene Expression (probe-level data) ➢Raw genotype calls (germline) ➢Gene-sample identifier links ➢Genome sequence files

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.

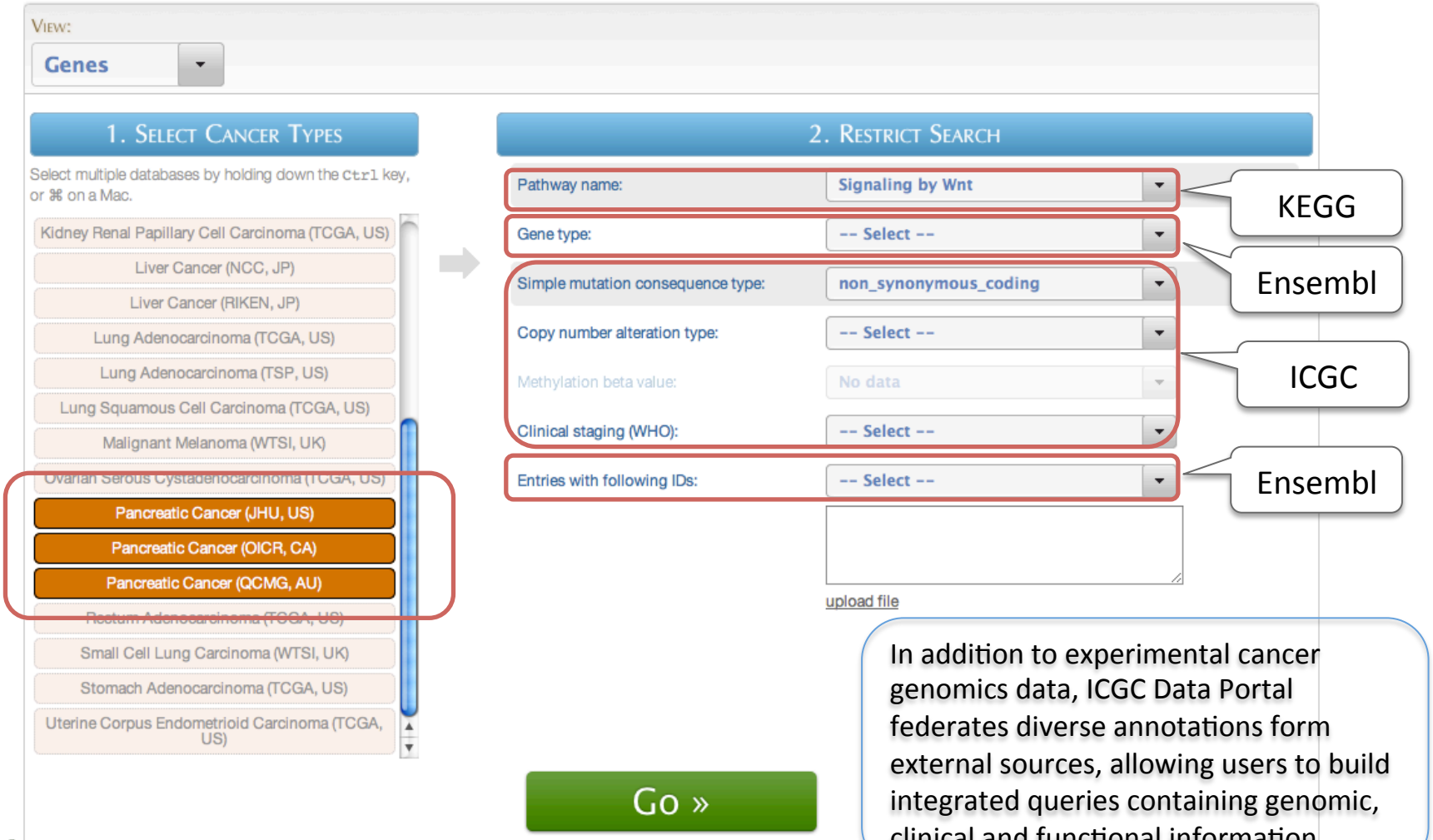


Each ICGC member deposits its site-specific data into its local BioMart database. These data are integrated with diverse annotations from external sources (Ensembl, Reactome, EKGG, COSMIC, Pancreatic Expression database, Breast Cancer Campaign Tissue Bank). Finally, all data are exposed through the ICGC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



In addition to experimental cancer genomics data, ICGC Data Portal federates diverse annotations form external sources, allowing users to build integrated queries containing genomic, clinical and functional information.

Home Page (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

Examples: TP53, ENSG00000133703, NM_000314

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

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

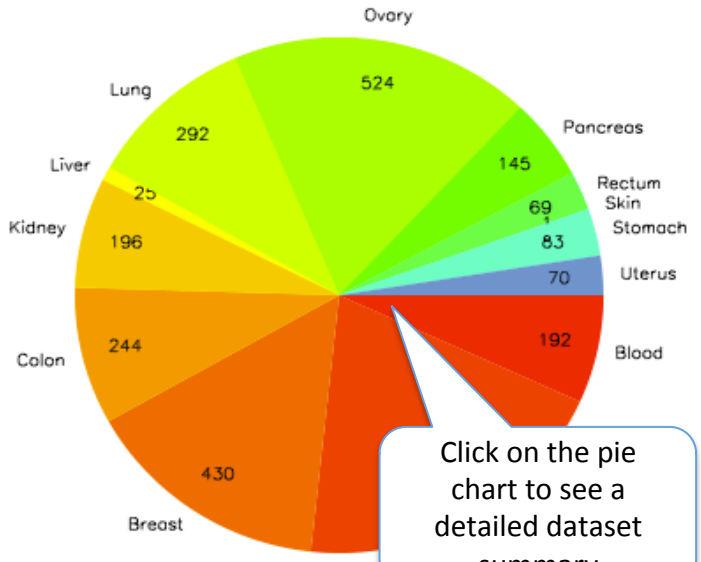
Query the data using one of several data entry points through Quick, Flexible, or Advanced search interface

Login to view controlled datasets

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

Cance

Donors by Tissue



Click on the pie chart to see a detailed dataset summary

Total Donors: **2,837**

Powered by biomart

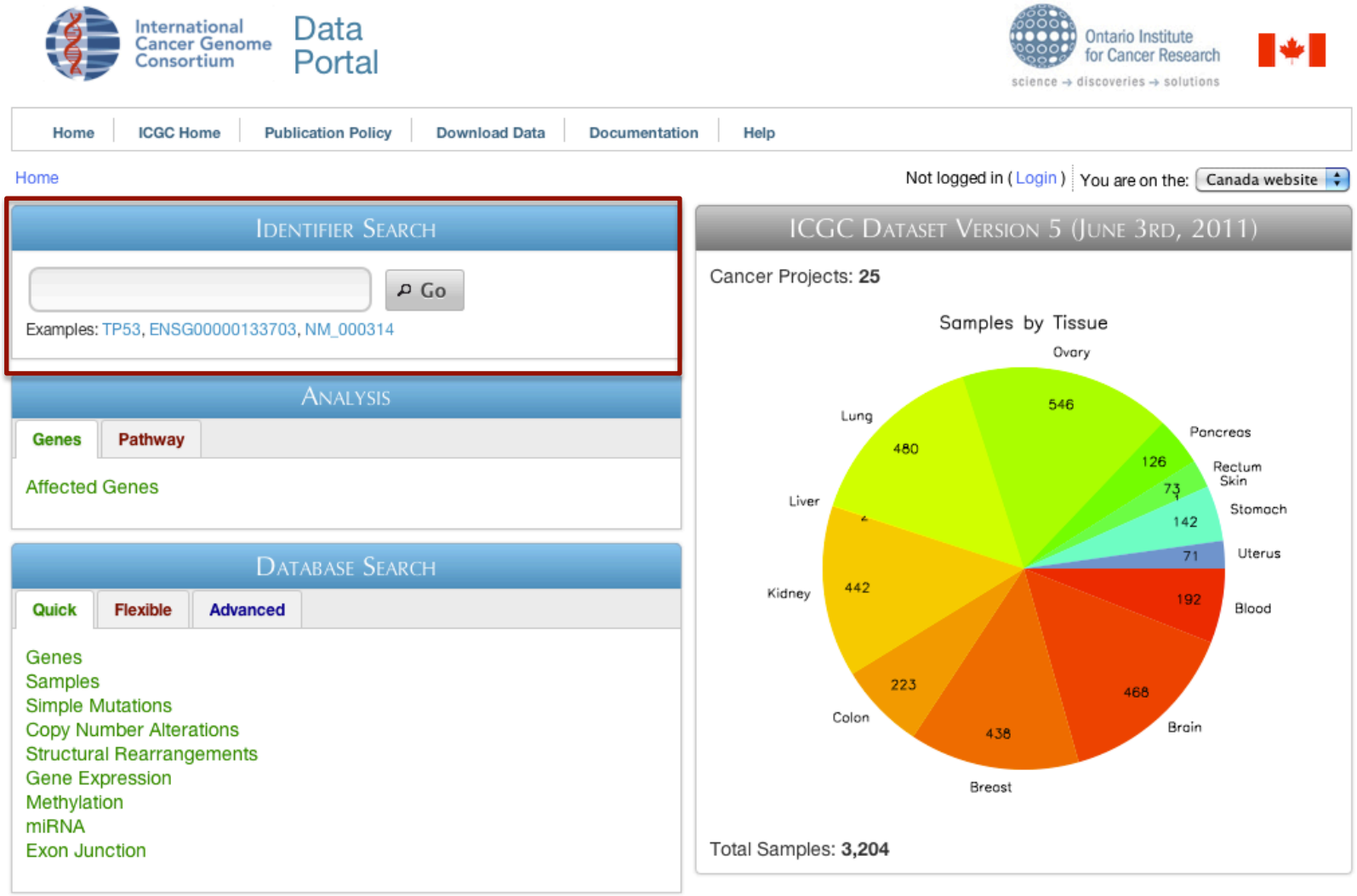
OICR Ontario Institute for Cancer Research

ICGC Data Portal Dataset Summary Table

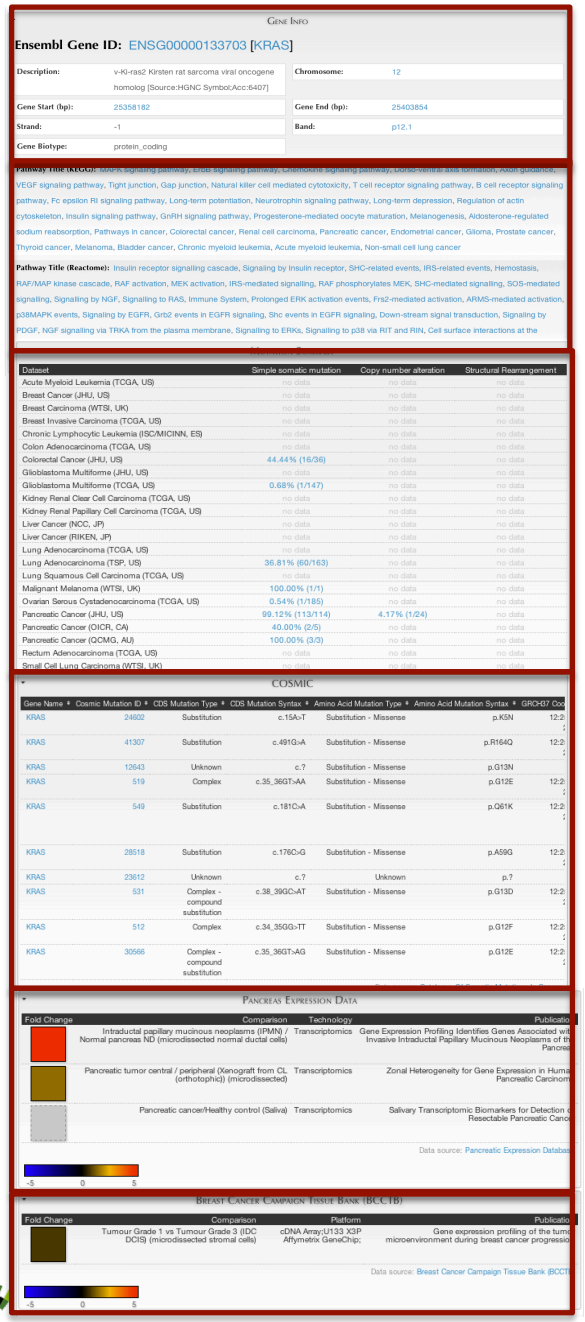
Source	Cancer Project	Tissue	Dataset									
			Donors	Samples	Simple Mutations	CNV	Structural Rearrangments	Gene Expression	miRNA Expression	Exon Junction	DNA Methylation	Germline Variations*
ICGC	Acute Myeloid Leukemia (TCGA, US)	Blood	188	188	-	-	-	-	-	-	188	-
	Chronic Lymphocytic Leukemia (ISC/MICINN, ES)[6]	Blood	4	4	4	4	-	-	-	-	-	-
	Glioblastoma Multiforme (TCGA, US)[1]	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)[5]	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)			142	-	-	-	-	-	-	142	-
	Uterine Corpus Endometrioid Carci				-	-	-	-	-	-	70	-
Other	Glioblastoma Multiforme (JHU, US)[4]				89	22	-	-	-	-	-	-
	Breast Cancer (JHU, US)[7,8]				42	-	-	-	-	-	-	-
	Colorectal Cancer (JHU, US)[7,8]				36	-	-	-	-	-	-	-
	Lung Adenocarcinoma (TSP, US)[2]				163	-	-	-	-	-	-	-
	Pancreatic Cancer (JHU, US)[3]	Pancreas	114	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



Gene Report



Ensembl

KEGG
Reactome

Mutation frequencies from cancer projects with data distributed around the globe

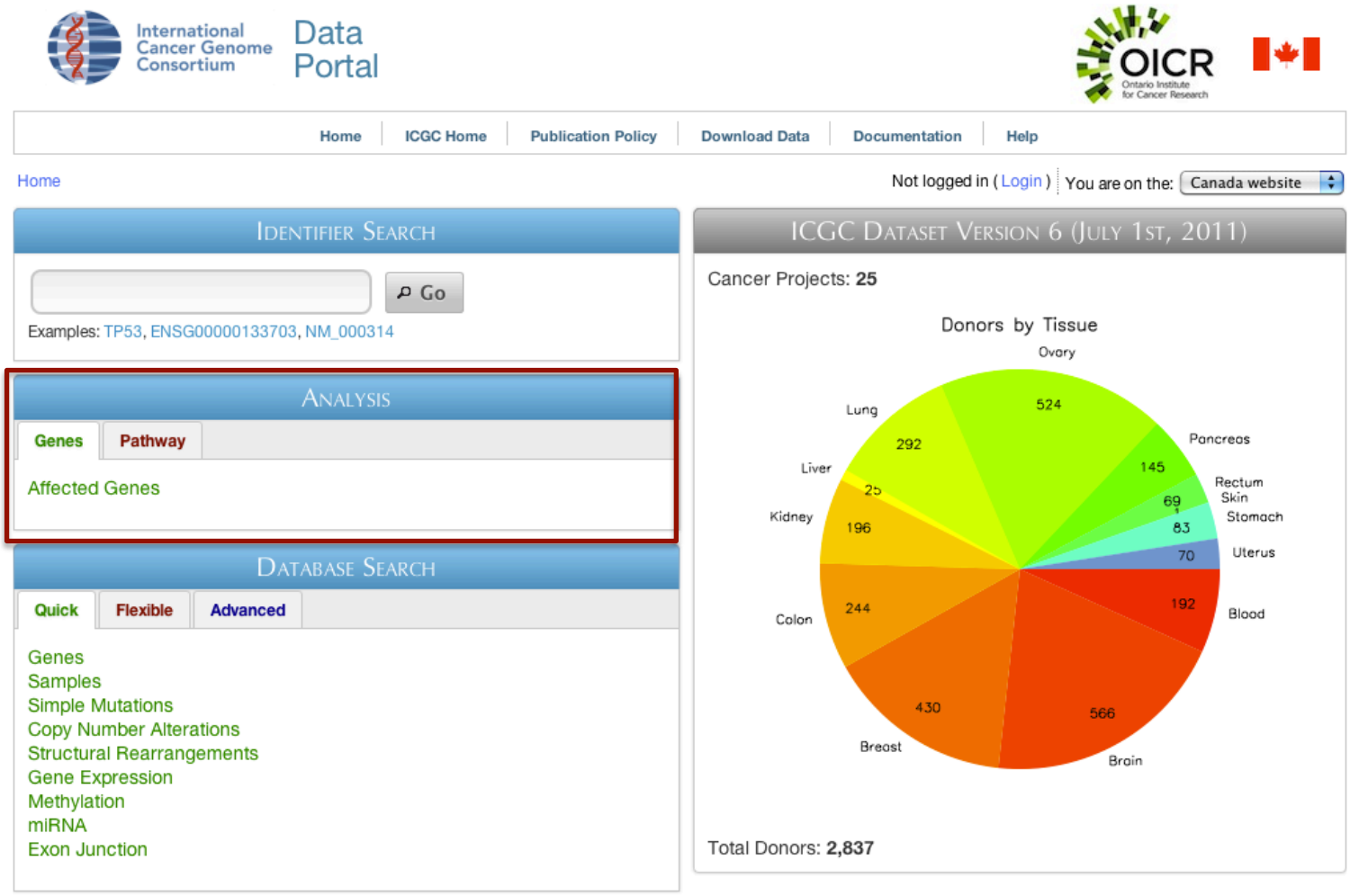
COSMIC

Pancreatic Expression Database (PED)

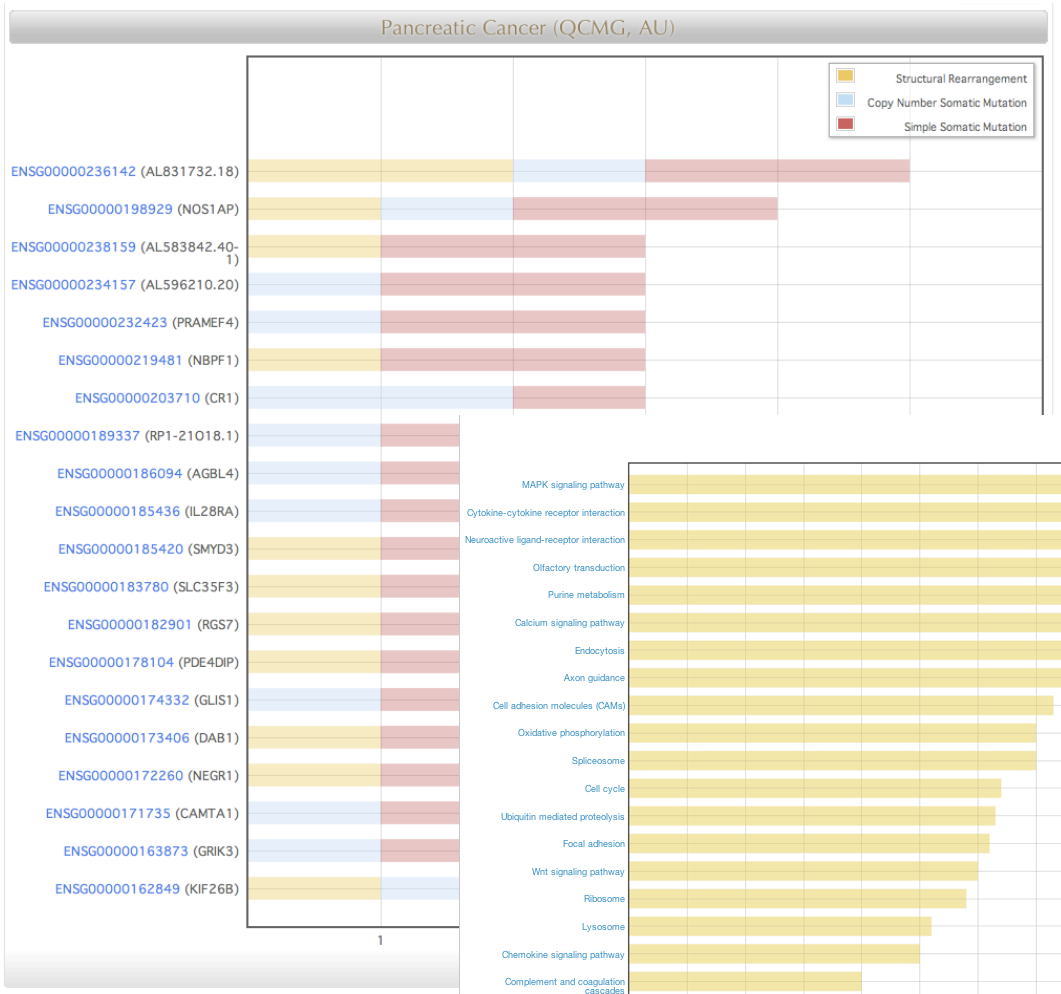
Breast Cancer Campaign Tissue Bank (BCCTB)

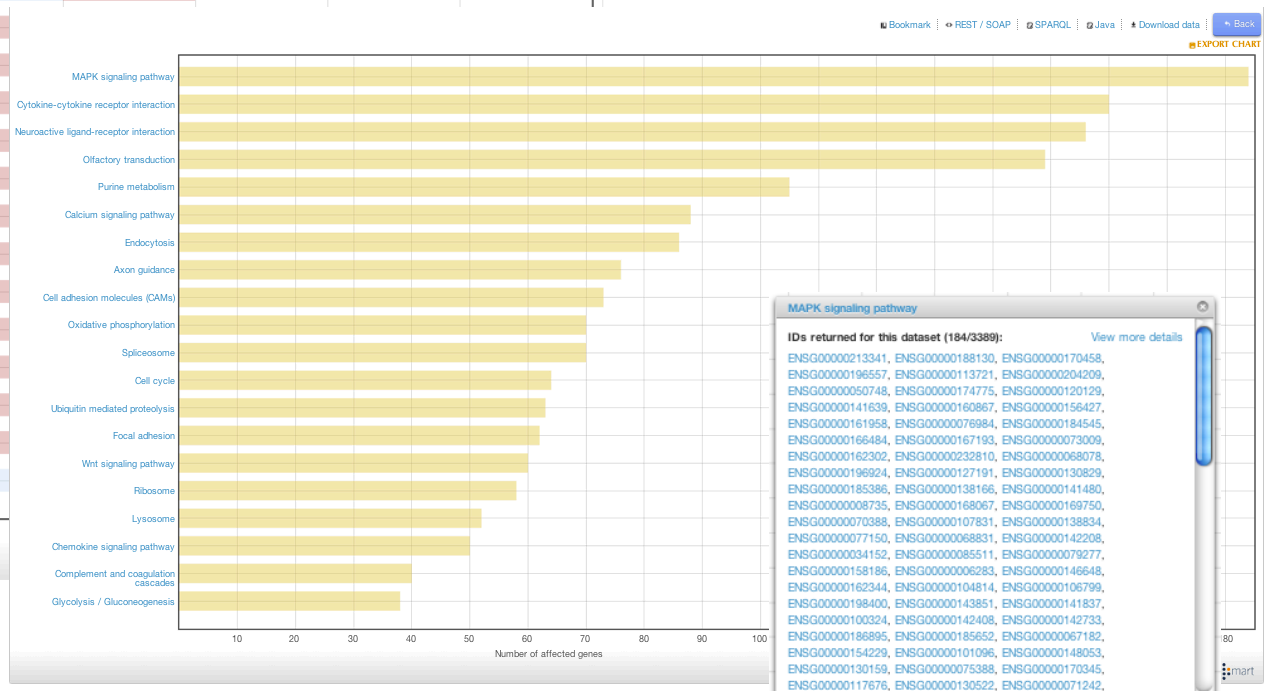
Displays data related to a single identifier, federated from multiple data sources

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



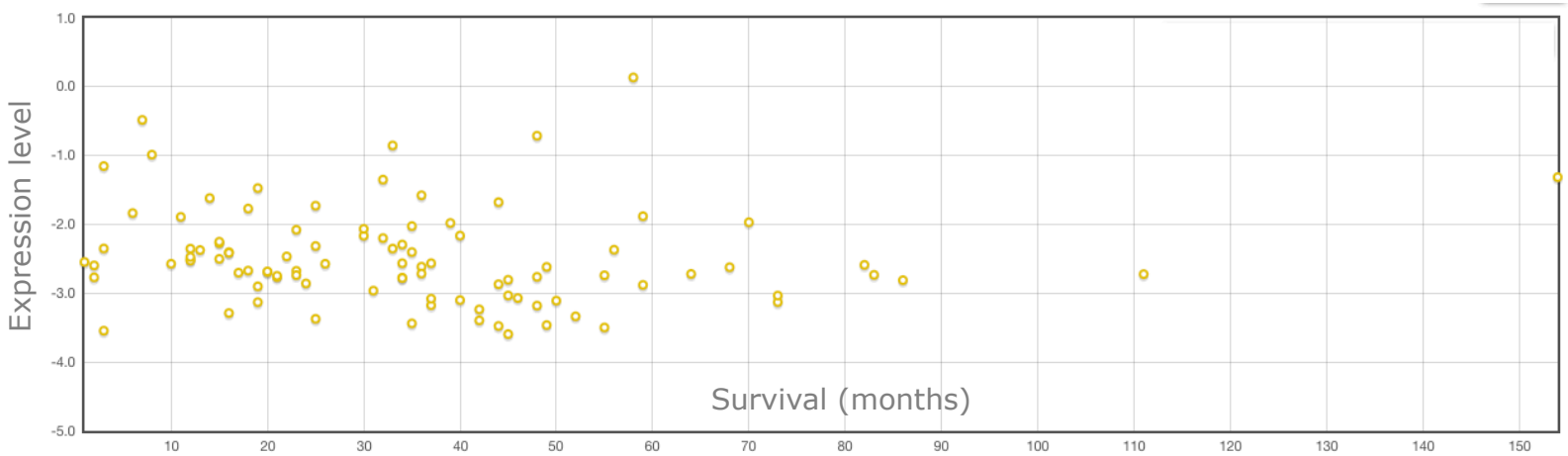
Most frequently affected genes or pathways

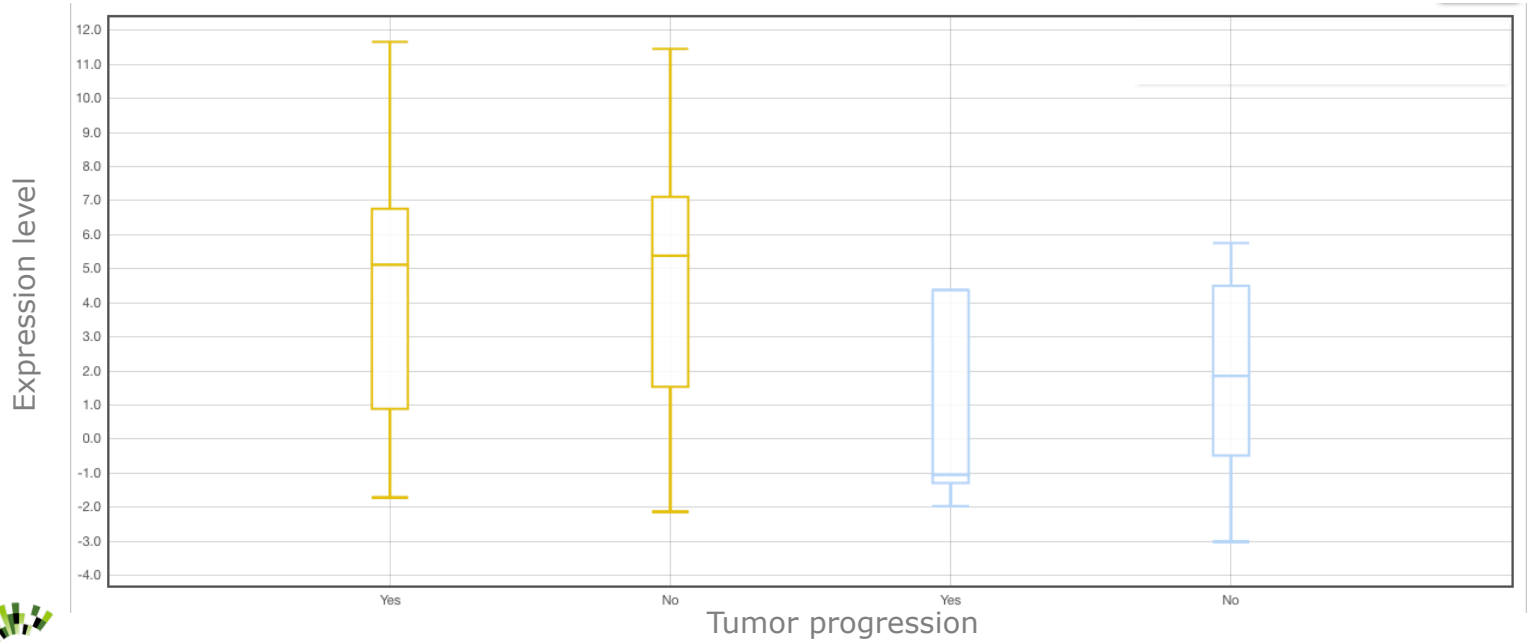




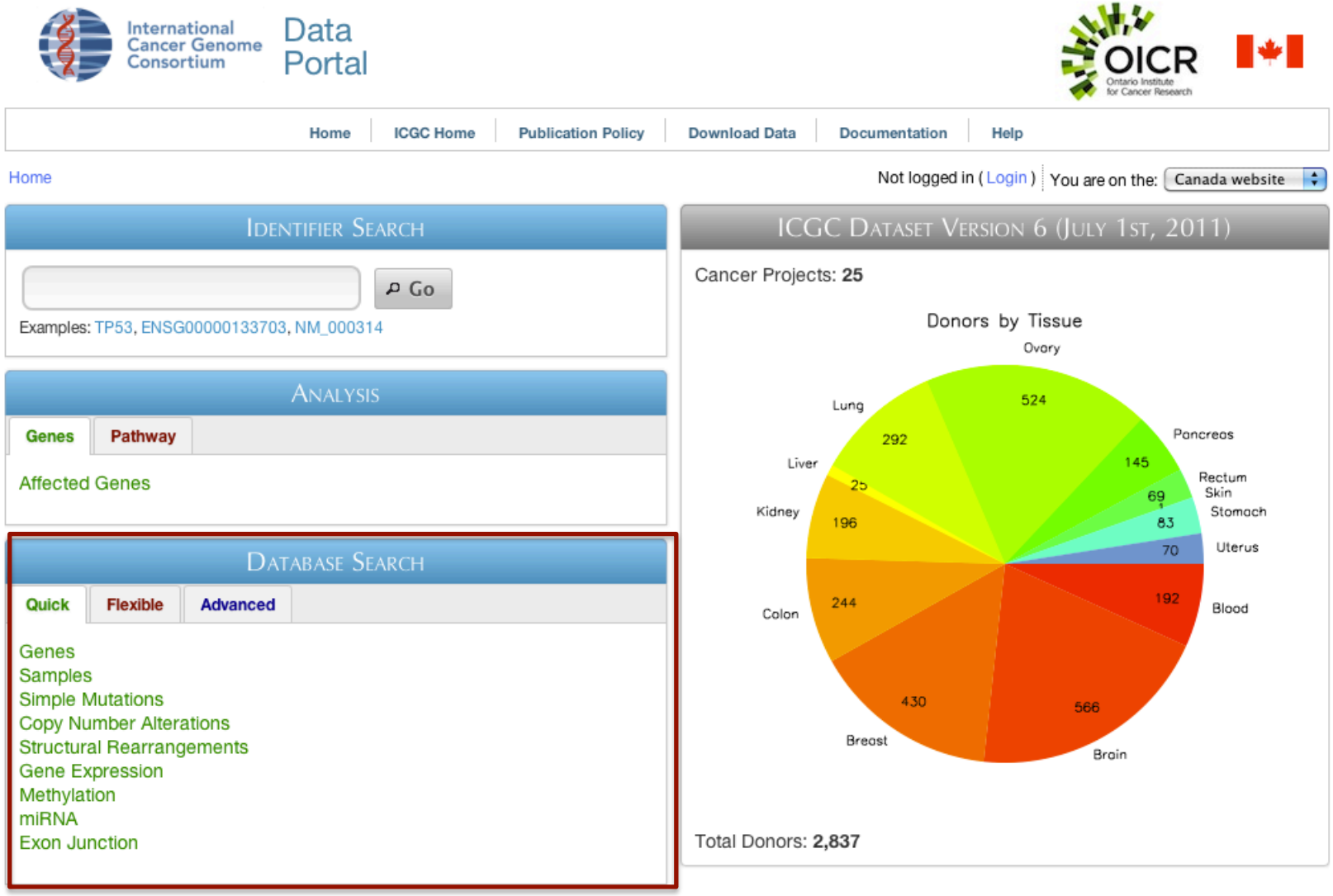
Data is processed by BioMart Plugins

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





Database Search

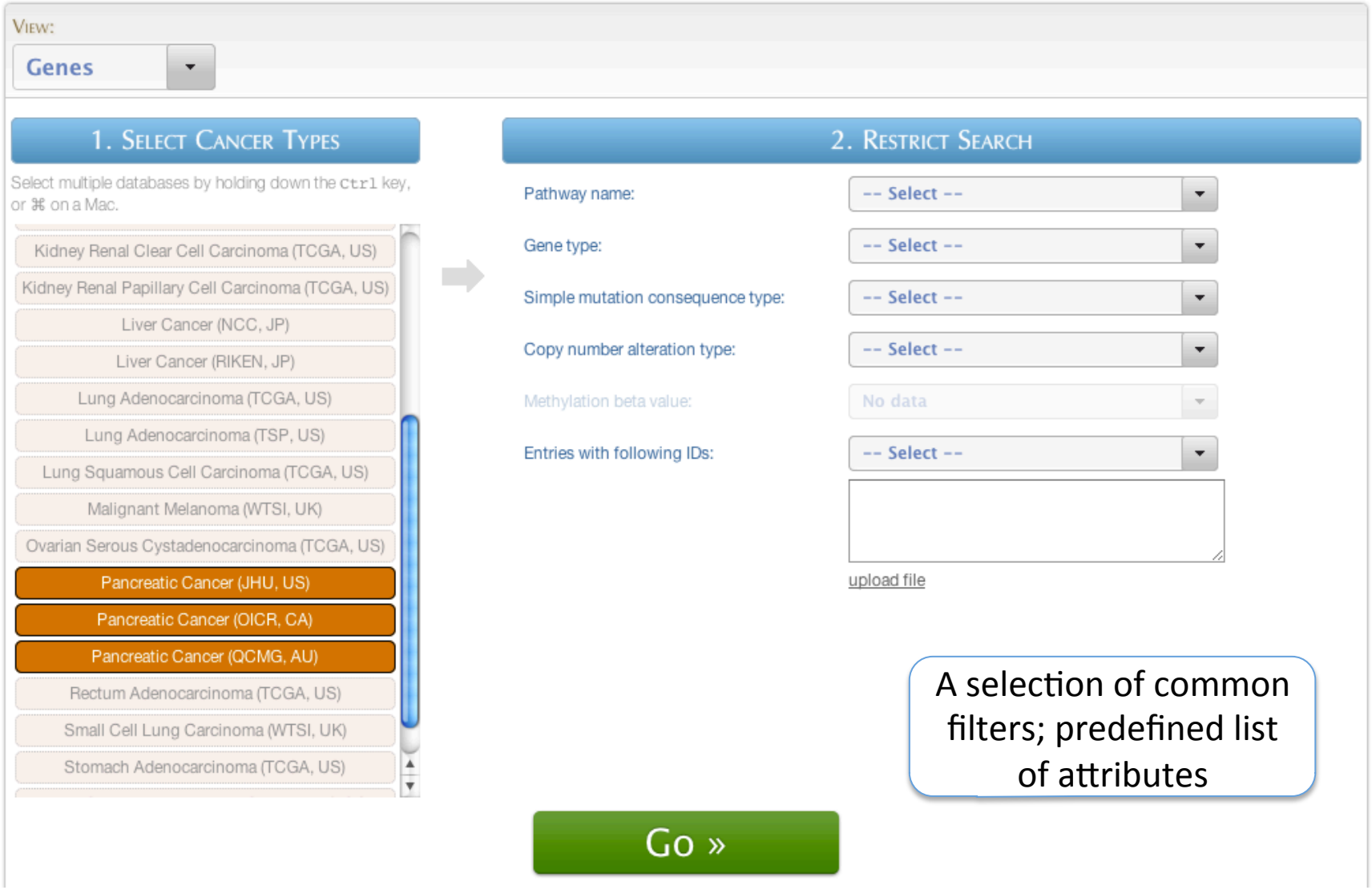


Database Search

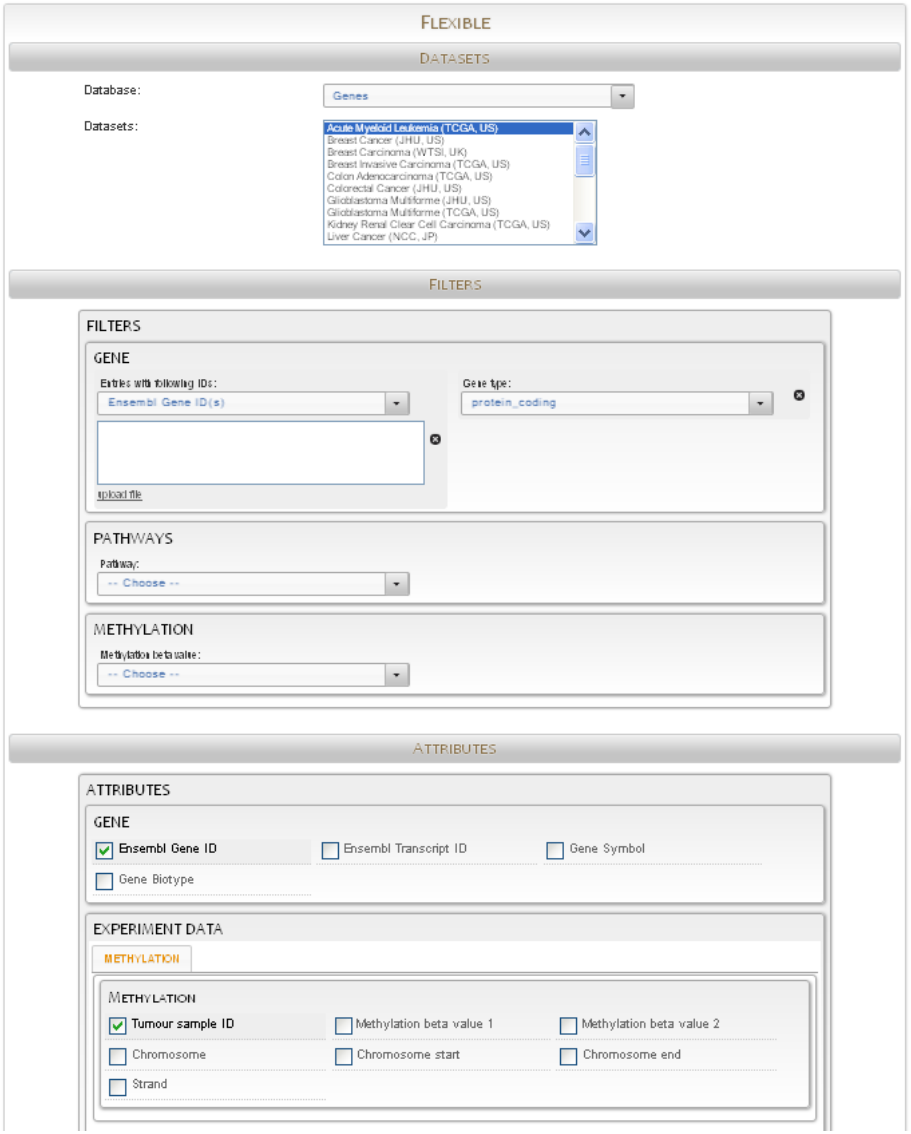
All queries in the Database Search consists of these few basic steps:

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

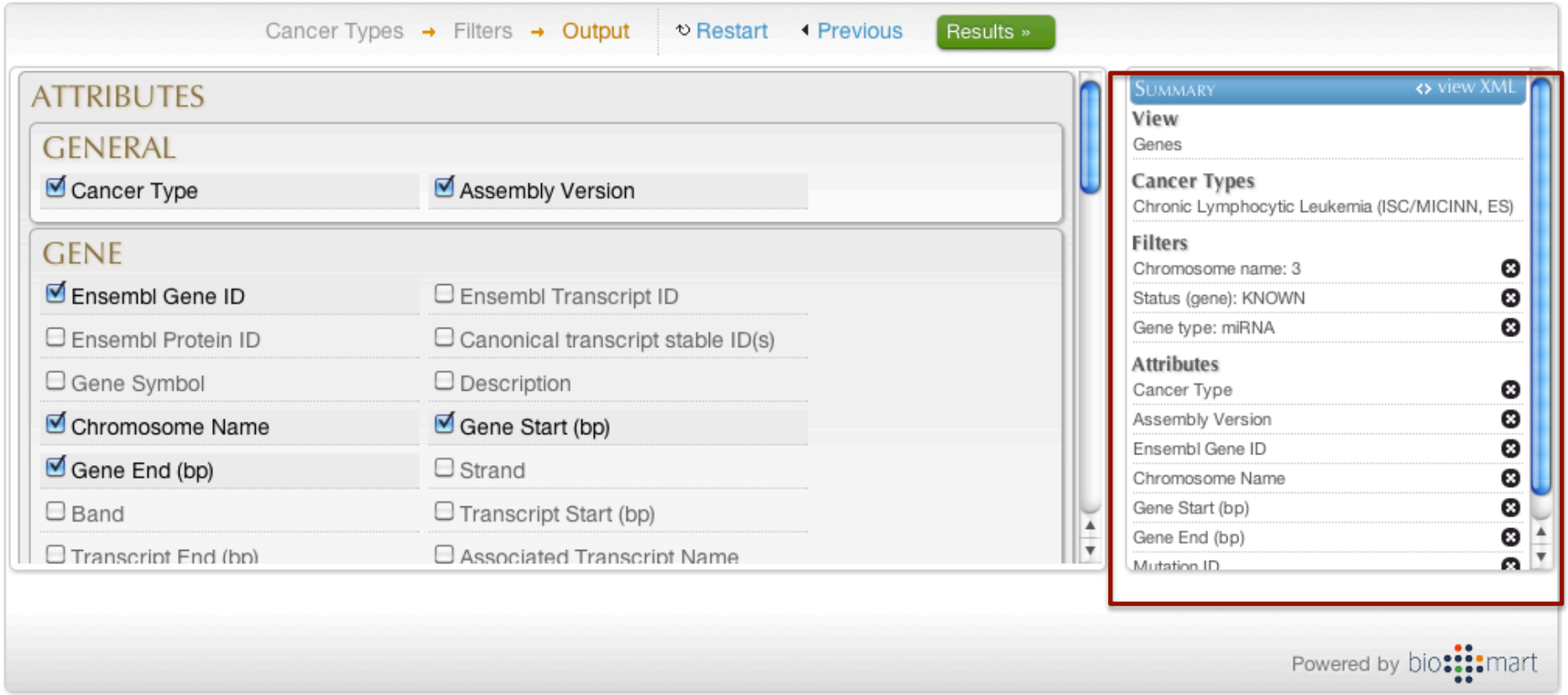


Flexible Search



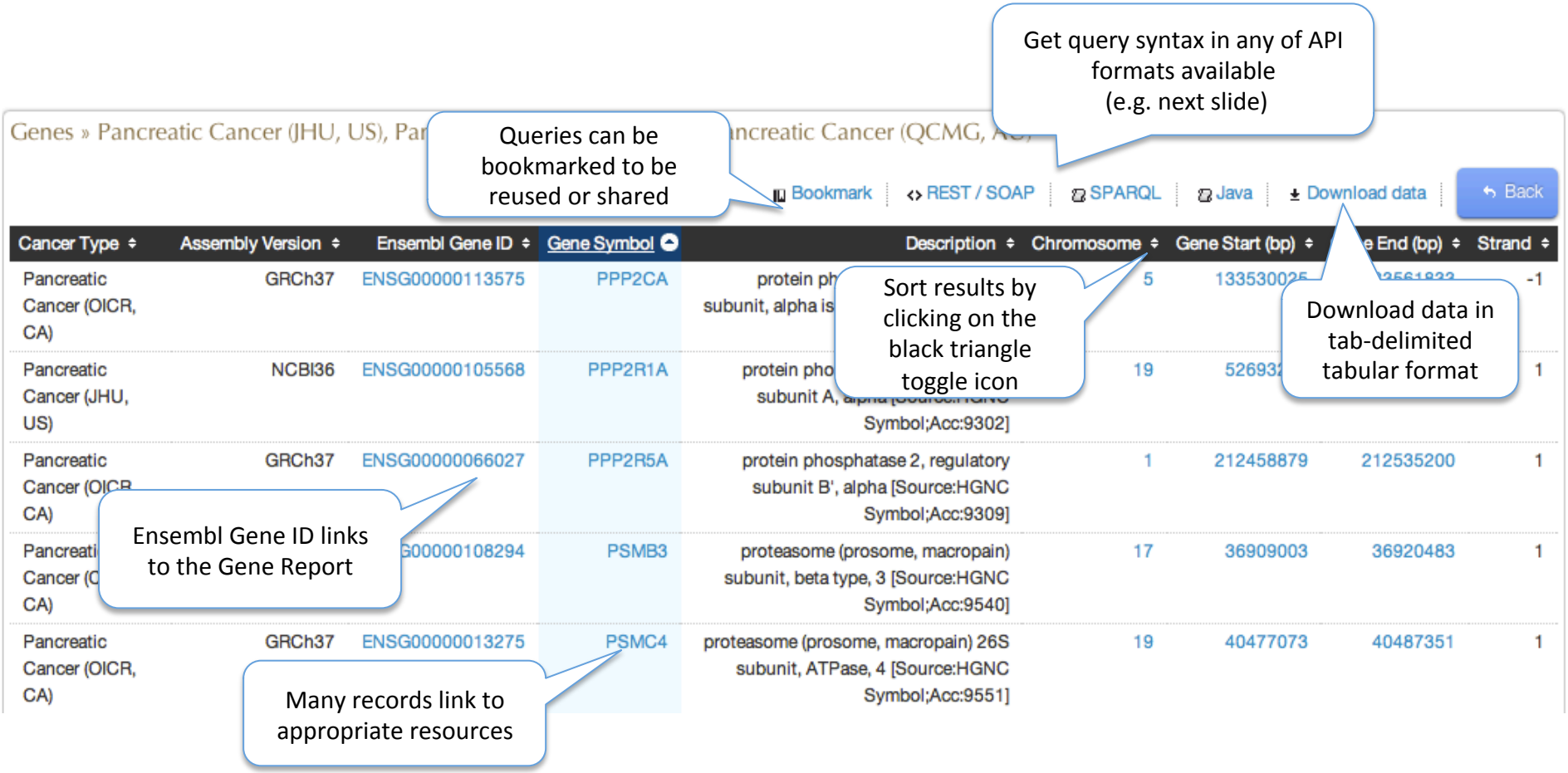
A selection of filters and attributes

Advanced Search



- 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.

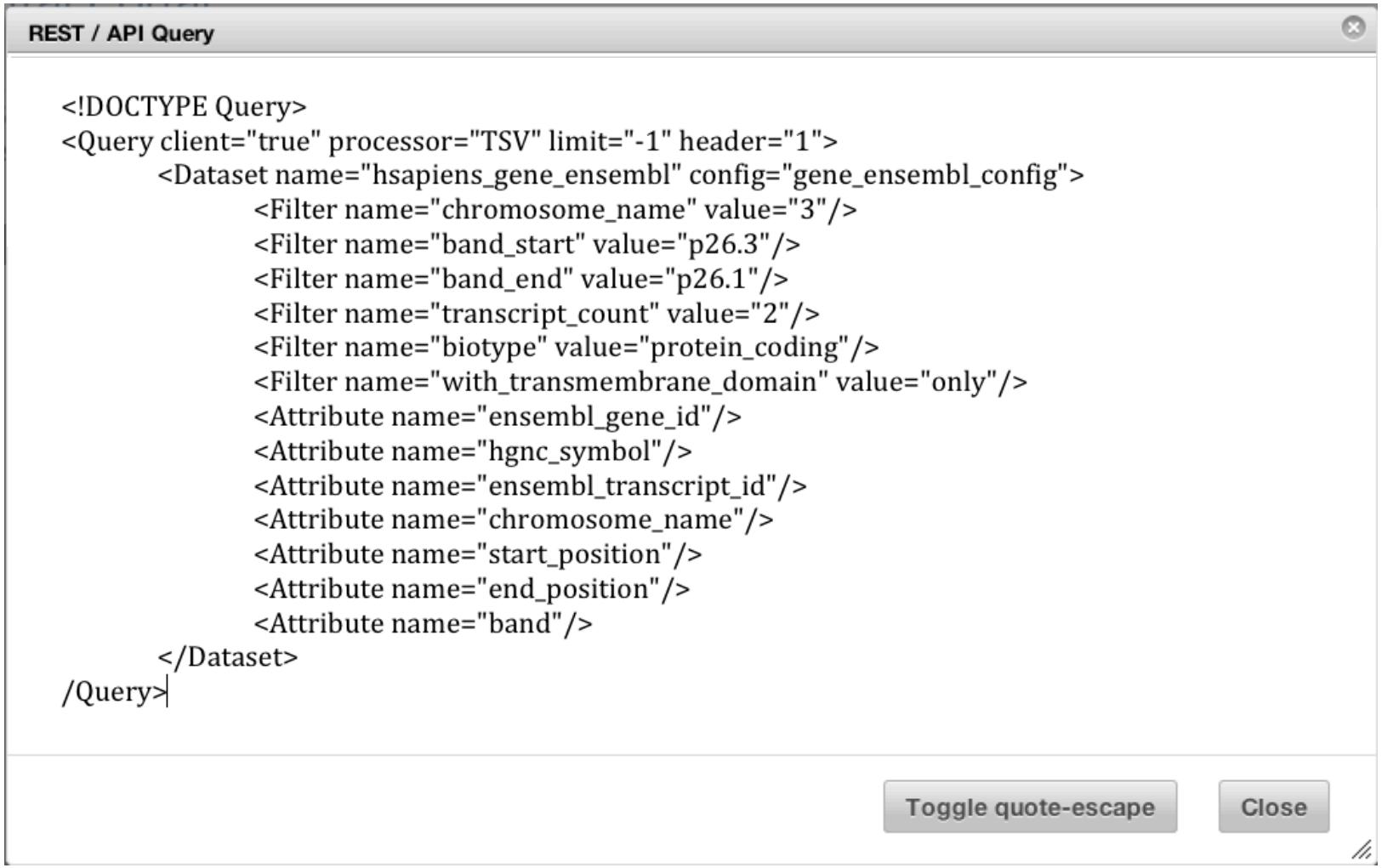
Results Page



Cancer Type	Assembly Version	Ensembl Gene ID	Gene Symbol	Description	Chromosome	Gene Start (bp)	Gene End (bp)	Strand
Pancreatic Cancer (OICR, CA)	GRCh37	ENSG00000113575	PPP2CA	protein ph subunit, alpha is	5	133530025		-1
Pancreatic Cancer (JHU, US)	NCBI36	ENSG00000105568	PPP2R1A	protein pho subunit A, Symbol;Acc:9302]	19	526932		1
Pancreatic Cancer (OICR CA)	GRCh37	ENSG00000066027	PPP2R5A	protein phosphatase 2, regulatory subunit B', alpha [Source:HGNC Symbol;Acc:9309]	1	212458879	212535200	1
Pancreati Cancer (C CA)		G00000108294	PSMB3	proteasome (prosome, macropain) subunit, beta type, 3 [Source:HGNC Symbol;Acc:9540]	17	36909003	36920483	1
Pancreatic Cancer (OICR, CA)	GRCh37	ENSG00000013275	PSMC4	proteasome (prosome, macropain) 26S subunit, ATPase, 4 [Source:HGNC Symbol;Acc:9551]	19	40477073	40487351	1

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.



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
Mailing list: users@biomart.org
ICGC Data Portal: dcc-support@lists.oicr.on.ca