Original Article

International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data

Junjun Zhang¹, Joachim Baran¹, A. Cros¹, Jonathan M. Guberman¹, Syed Haider², Jack Hsu¹, Yong Liang¹, Elena Rivkin¹, Jianxin Wang¹, Brett Whitty¹, Marie Wong-Erasmus¹, Long Yao¹ and Arek Kasprzyk^{1,*}

¹Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada and ²Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK

*Corresponding author: Tel: 647 258 4321; Email: arek.kasprzyk@gmail.com

Submitted 14 April 2011; Revised 11 May 2011; Accepted 17 May 2011

The International Cancer Genome Consortium (ICGC) is a collaborative effort to characterize genomic abnormalities in 50 different cancer types. To make this data available, the ICGC has created the ICGC Data Portal. Powered by the BioMart software, the Data Portal allows each ICGC member institution to manage and maintain its own databases locally, while seamlessly presenting all the data in a single access point for users. The Data Portal currently contains data from 24 cancer projects, including ICGC, The Cancer Genome Atlas (TCGA), Johns Hopkins University, and the Tumor Sequencing Project. It consists of 3478 genomes and 13 cancer types and subtypes. Available open access data types include simple somatic mutations, copy number alterations, structural rearrangements, gene expression, microRNAs, DNA methylation and exon junctions. Additionally, simple germline variations are available as controlled access data. The Data Portal uses a web-based graphical user interface (GUI) to offer researchers multiple ways to quickly and easily search and analyze the available data. The web interface can assist in constructing complicated queries across multiple data sets. Several application programming interfaces are also available for programmatic access. Here we describe the organization, functionality, and capabilities of the ICGC Data Portal.

Database URL: http://dcc.icgc.org

Project description

The International Cancer Genome Consortium (ICGC) is a multidisciplinary, multi-institutional collaborative effort aiming to systematically and comprehensively characterize somatic mutations in 50 different cancer types and subtypes (1). Five hundred tumor genomes, as well as matched normal control genomes for each cancer type, will be analyzed using high-throughput next-generation sequencing technologies to detect a wide range of somatic mutations, including single nucleotide mutations, small insertions/ deletions, copy number alterations, translocations and other chromosomal structural rearrangements. Genomewide methylation state analysis and whole-transcriptome sequencing have also been planned to provide additional molecular-level characterizations. To make the effort more scalable each member institution specializes in generating data for a particular tumor type. At the time of writing, 12 countries have joined the effort (Figure 1).

One of the major goals of ICGC is to rapidly bring these data to the cancer research community in order to accelerate studies on the discovery of cancer causes, to enhance the accuracy of diagnoses and to improve treatments. In order to achieve this task, the data generated by the consortium members have to be managed efficiently. Amongst the most important data management challenges faced by the consortium is the high complexity and heterogeneity of the data types involved, the necessity to link

© The Author(s) 2011. Published by Oxford University Press.

This is Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http:// creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. Page 1 of 10

Original Article

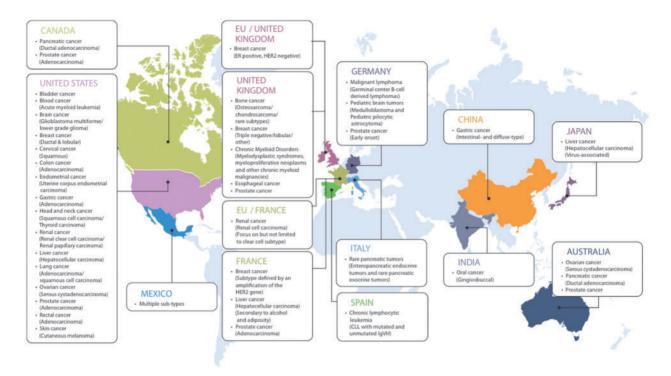


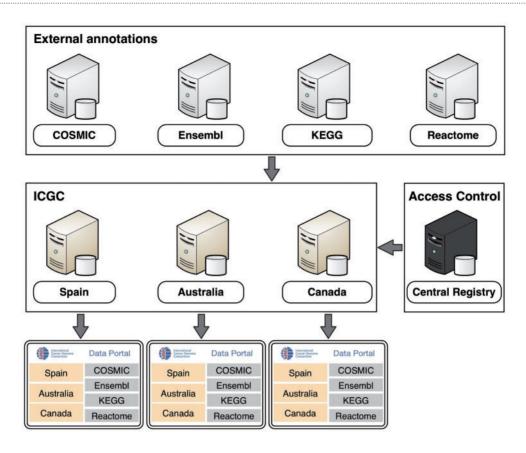
Figure 1. International Cancer Genome Consortium (ICGC) projects (March 2011).

different data types, and the need to protect controlled data. Furthermore, the high volume of data and the distributed nature of the sources make traditional centralized approaches to data management impractical. Consequently, the ICGC has adopted federated data architecture to address their data management needs. The scalability of the system is improved by having each member institution store and process data locally; the data federation software then presents these separate sources as a single access point for remote data access.

BioMart, an open source data federation system (2), has been chosen as the ICGC data management platform. A number of features make BioMart an attractive solution for the ICGC. BioMart's flexible data model makes it generally applicable to a wide range of biological data types and built-in query optimizations make it suitable for large data sets. In addition, BioMart supports an industrystandard security framework that is needed to provide secure access to the controlled data. Finally, a large number of existing expert-maintained public annotation databases can be readily federated, adding value to the interpretation of the ICGC experimental data.

The architecture of the ICGC data management system has been modeled on the BioMart Central Portal (3, 4). For a number of years, the portal has been successfully providing a single point of access to a large number of biological databases distributed across the world. Each BioMart database federated in this portal is maintained independently, released and updated on its own schedule. Similarly, each ICGC member maintains their local BioMart database where clinical annotations and the data produced from genomic analyses are deposited. However, in order to make these data comparable across different cancer projects and create a unified ICGC data set, a mechanism for enforcing uniformity is required. Thus, the same set of data models, controlled vocabularies, ontologies and reference data sets are used in all of the ICGC member databases. Furthermore, a gene-centric data model adopted from Ensembl Mart (5) is used to reference experimental data to the same set of annotated genes.

In order to link individual BioMart databases into a unified system, each of the ICGC member institutions maintains an instance of the BioMart server. These servers can each access data in three ways: they can access their own local databases directly; they can communicate with the other ICGC BioMart servers, in order to retrieve data from remote databases; and they can communicate with non-ICGC BioMart servers, such as COSMIC (6) and Reactome (7), to retrieve publicly-available annotation data. In this manner, the BioMart server maintained by each ICGC member acts as a fully featured ICGC data portal, providing unified access to all the consortium data. The users select a combination of different data sets and specify query criteria using a variety of graphical user interfaces and analysis tools that are available from the portal. Application programming interfaces (APIs) for



Java, REST, SOAP and SPARQL are also available for programmatic access. Behind the scenes, BioMart software breaks down this query into smaller parts that are distributed to the remote data sources. The results are collected and compiled into a single unified results set that is presented to the user.

Some of the data in the portal has access restrictions in order to preserve patient privacy. To protect this sensitive data, communication between BioMart servers is secured by HTTPS and authentication is handled using the oAuth protocol. Access to controlled data is managed by the Central Registry server. When a user logs into an ICGC data portal server using his/her OpenID, the server will consult the central registry to ensure that the user is authorized to access the data (Figure 2).

Data content

At present the portal contains data from 24 cancer projects, consisting of 3478 genomes and 13 cancer types and subtypes. This includes the data generated from seven studies performed by five ICGC participating institutions located in four countries. The Data Portal also hosts data from other large-scale cancer genome projects including The Cancer

Genome Atlas (TCGA) (8, 9), Tumor Sequencing Project (TSP)(10) and Johns Hopkins University (11-14). Open access data sets include (i) simple somatic mutations, (ii) copy number alterations, (iii) structural rearrangements, (iv) gene expression, (v) miRNA, (vi) DNA methylation and (vii) exon junctions. Secure access to controlled data sets, such as germline variations, is available to authorized users. A summary of data currently available on the Data Portal is shown in Table 1. In addition to cancer genomic data, the Data Portal also federates several public databases (Table 2). Currently this includes Ensembl genome database (5), Kyoto Encyclopedia of Genes and Genomes (KEGG) (15-17), Reactome (7), COSMIC (6), Pancreatic Expression Database (18) and the Breast Cancer Campaign Tissue Bank (19), and the number of resources is continuously growing.

Query examples

The Data Portal provides three major interactive entry points: Identifier Search, Analysis and Database Search (Figure 3). Identifier search (Figure 3A) lets users input identifiers that are commonly used in public annotation databases (e.g. HGNC gene symbol, Ensembl ID, RefSeq ID,

Source	Cancer project	Data set							
		Simple mutations	Copy number alterations	Structural rearrangements	Gene expression	miRNA expression	Exon junctions	DNA methylation	Germline variations
ICGC	Breast carcinoma (WTSI, UK) Liver cancer (NCC, JP) Liver cancer (RIKEN, JP) Malignant melanoma (WTSI, UK)	•••		••••					
	Pancreatic cancer (OICR, CA) Pancreatic cancer (QCMG, AU) Small cell lung carcinoma (WTSI, UK)	••	•	••	••	••	•		• •
TCGA	Acute myeloid leukemia Breast invasive carcinoma Colon adenocarcinoma GlioblastomaMultiforme Kidney renal clear cell carcinoma Kidney renal papillary cell carcinoma Lung adenocarcinoma	•			••••				
	Lung squamous cell carcinoma Ovarian serous cystadenocarcinoma Rectum adenocarcinoma Stomach adenocarcinoma Uterine corpus endometrioid carcinoma	•			•••			••••	
Other	Breast cancer (JHU, US) Colorectal cancer (JHU, US) GlioblastomaMultiforme (JHU, US) Lung adenocarcinoma (TSP, US) Pancreatic cancer (JHU, US)	••••							

Page 4 of 10

Table 2. Public databases federated with the ICGC Data Portal

Source	URL	Description of contents
Source	ORE	Description of contents
Ensembl	www.ensembl.org	Genome annotation
Reactome	www.reactome.org	Pathway annotation
KEGG	www.genome.jp/kegg	Pathway annotation
COSMIC	www.sanger.ac.uk/genetics/CGP/cosmic/	Somatic mutations in cancer
Pancreatic Expression Database	www.pancreasexpression.org	Pancreatic cancer expression data
Breast Cancer Campaign Tissue Bank	www.breast cancert is sue bank.org/bio-informatics.php	Breast cancer expression data

cer Genome Data Portal

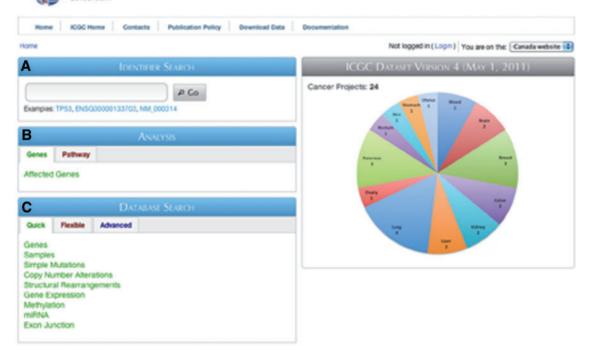


Figure 3. Screenshot of the ICGC Data Portal home page. Three main entry points are available: (A) Identifier search, (B) Analysis and (C) Database search.

UniProt ID and other accessions) and returns links to the corresponding Gene Report page, which displays basic gene description, pathway annotation, mutations found in the COSMIC database, as well as pancreas and breast cancer expression data (Figure 4). In addition, the Gene Report displays a summary of mutation frequencies in the selected gene across all cancer projects (Figure 4D). Records in the gene report (i.e. genomic coordinates, pathway name, publication) are linked to appropriate resources when available, allowing users to easily retrieve additional information on their data of interest.

To help with the interpretation of cancer data, Gene and Pathway analysis tools are available in the Analysis section of the Data Portal (Figure 3B). These tools enable users to view the most commonly affected genes or pathways in one or more cancer projects. Results are presented in an easy-to-follow chart that can be exported as an image file, and the numerical data can also be downloaded for further processing (Figure 5).

Using the Database Search entry point (Figure 3C) users can interactively query the database by several data types: genes, samples, simple mutations, copy number alterations, structural rearrangements, gene expression, miRNA, DNA methylation and exon junctions. Three interfaces are available: *Quick*, *Flexible* and *Advanced*. The *Quick* interface contains a pre-selected set of the most commonly used

					E			- 009	VIII.				
Income had Press and Physics	ENSIGNOOD 12703 DOK	ca .			-					-	WARD MARK THINK		
Construct Cold M. Har.		-			10.0	24.54	Subalization	1.044	SAURIO B	_	241		1.5904
management of the	Artist webst of technological												-
	arming Decord CAL Invite/Aux/HUT				8000	-	Interaction of	CANNER.	THE R. LEWIS CO., NO.		Page 1	e	CO-MORE DelaCol
	a cost but can contribute an out i	From Not Spr.	100441		-	2011	later.		Davision B	_	100		
anti-the A	PROPERTY AND A DESCRIPTION OF A DESCRIPR	mont.	-			-	1000	COLUMN TAX	Same of the				
									Carriero, M				riand
let a	48.1	Cost British	p.0-0.1109		-	-	Substrum.	A MERICA	Section 8	-			-
Andrew Prop. Statute & and	Analys private livel specify advecta		Anna Provincial	Contraction in the second second									-
P	Opening Bargarow, Subscript rul ru	Contract allowed a	and the second	construction of the second									
D. shand be ad 15	official problems' generation of the	and the second s	success where it is	and a number should	-	-	NAME AND	1.000.00		unes.	0.01		-
DENELS'S RECORD AND	ing periods, card-period based on the region	NUM HANKS FROM	is Long-term periodicilities	C regulation of all an									1.000
rosante, teach signal a	pasters, Gritt spaingasters, Pepe	cross mediated benyt	Published, Milanoper	unit, findemer sons may be had	80.00	10001.0	waters.	44		and the second			16.0
other instantion, for so	and it cannot find and in the Read of the	other states	soor, waterstation in a	an, Barren Printer and at	10.0	98.4	Origini-	<9,900.48	Database Auto- 40	in a los	PD1	0	-
							NOCK40						200
Water and a state of the state of the	Raber Levie Groece space where a	Co spece a serva	Residence and Residences	-	10.0	-	Dollars.	434,3858477	Databa hou at		144		-
-	Bernarder Specifiques and Barning in	And Address of the Owner, Name	other states	And American State	-			1.04.0000001					0.04
					47.4	2004	Compter-	18.805.40	Database Re		+55		CUERROR
The rest is not at an	an use more compared strand	or and a local design of the local design of t	of an include	all and see an an an an an an			distant and						10,0047
Denited to WOR, Danishing	S-FAL Development with a select Provide	CONTRACTOR NO.	HE, NO PARADARKI	and a state of the second s			10.812/2004						
states which have no	Speaking og 1924 A. Bendt somernal i Klaffe og	and the same in the	Albertown have been	And the state of t						-	Deserve Cake of	Density 21.0	
					C								
A Desired States of the second second	rang values has har be patent memories.	Adulte & Junited and	office. That have not see the	Control Cost Strictly restrictions		11121	a o m sec-						
or he wanted and find the	rating NOM standingthe neurile surgrouph	Ane athres. Bare	singles interdestors, inte	mular 2 a protong									
_			ringing interdentions, imm	restar 2 a geolog	-								
r te seu ar vel hel te D		And patients, Rev A School as	deglas interdeutura, trae	rmile 3 e peders	F			Proma law	esce Das				
D	Mg/ series Despinant			In the fee upper	F			Prenalise					5.00 V
D	Michael Broger an Michael	Source Co	or and	This fact segment	Full Comp	100.0010		Courses and	Technology	0-191		10478-0	e churke -
	Multime Berger av Multime	Sources Marc House Co	and allow the state	The a last fragmatical	Full Comp	on the second	0.20210	Courses and	Technology	Page 1	tacing the	125755	1221
D Constant Total M particular and arrive protection Total Constant (PTR), 148	Microso Angle as Angle as	states co	april Johan Philippia Arriteri Arriteri Arriteri	This Lee face signal	Full Comp	ons[[[0]]]	10.20210	Courses and	Technology	125237		122723	1221
D Control Control System Control of Control Control Control of Control of Control of Control Control Control of Con	Million Millio	Sciences art contribut Co		The allow New Segment	Fut Coup	ornalis ornalis	no de la composition	Concrete Sensitives and another St. Br	includings includings of	Papijini i	and the second second	ering had	States of the
D constanting of the second s	Microsoft Bogin Bo CLALUNI B. LID	Scherers art destroy Co		Backerine agrief	Fut Coup	Outside contract of	no de la composition	Courses Sensitives and activities	includings includings of	Papijini i	NUMBER OF T	ering had	
Dente International (PA) (4) International (PA) (4) International (PA) (4) International (PA) (4) International (PA) (4) International (PA) (4) International (PA) (4)	Mariana Mariana Katalan Katalan Katalan	Street	94.450 Fabric	The object data straphyridd - Second d Frankish Frankish Frankish Frankish Frankish	Fut Coup	Outside contract of	no de la composition	Concrete Sensitives and another St. Br	includings includings of	Papijini i	NUMBER OF T	ering had	States of the
Designation and area pro- readed for some pro- trace for some pro- trace for some formations of the information of the pro- section of the some pro- pro- pro- pro- pro- pro- pro- pro-	Microsoft Margin to An ang An Ang An Ang Ang Ang Ang Ang Ang Ang Ang Ang Ang	N Schwarzy NATE Republica CO Intern I	April Johnson Production Strategiese Strat		Fut Coup	Persentis same or	ne and a second second	Concrete Sensitives and another St. Br	ternengi motjones o	Paggini I Jun	el federajantijskoj la 19. kompanijskoj la O	Present	
Designation and the particular a	Microsoft Microsoft Clark Mig Microsoft Micros	N Scherery NYC Reporter CO Inter I			Fut Coup	Persentis same or	ne and a second second	Concerner Search and and an and and and an and and and an and an and and an and an and and an a	ternengi motjones o	Paggini I Jun	el federajantijskoj la 19. kompanijskoj la O	Frank Salar	
Designation of the second seco	Microsoft Annu Micros	N SCHWARTS CO HART CARLING CO HART CARL HART CARL H	101.000 Putota 10.000 10.000 10.000 10.000 10.000 10.000 10.000 10.000 10.000 10.000	801.0.1.000 /000 vilgeneel 	Fut Coup	Persentis same or	ne and a second second	Concerner Search and and an and and and an and and and an and an and and an and an and and an a	ternengi motjones o	Paggini I Jun	el federajantijskoj la 19. kompanijskoj la O	Present	Contra or
Control or all system control (ref) from Granes (ref) (4) from Granes (ref) (4) from Granes (ref) (4) (4) (4) (4) (4) (4) (4) (4	Maj sera Maj sera Culta July N. US N. US N	A Streamers International Con- International Int		Stora Lance - Andre uniquement - Seconda Territoria Territoria Territoria Territoria Territoria Territoria Territoria Territoria Territoria Territoria	Fut Coup	Persentis same or	ne and a second second	Concerner Search and and an and and and an and and and an and an and and an and an and and an a	ternengi motjones o	Paggini I Jun	ni indogenijkog k ni indogenijkog in O nij i smeljskog ili Pasa	ren farmen Fransen Fransen	
Design of the second se	Microsoft Microsoft	Covers and events and and and and and and and and and and	001 4000 F000 40 00000 00000 00000 00000 00000 0000 0000	2012 ALC Face council 1019 20 1019 20 100 100 100 100 100 100 100 1	Fut Coup	Persentis same or	ne and a second second	Concerner Search and and an and and and an and and and an and an and and an and an and and an a	ternengi motjones o	Paggini I Jun	el federajantijskoj la 19. kompanijskoj la O	ren farmen Fransen Fransen	
Next oral significant and are pro- ment Garron (2014) All ment Garron (2014) All ment Garron (2014) All den Internet Antonio (2014) All Internet (2014) All Internet (2014) All Internet (2014) All Internet (2014) All Internet (2014) All and Internet (2014) All	Maj sera Maj sera Culta July N. UB N. UB N	Albanett and Australia Co and and and and and and and and		Sea Las Age concerno mais a mais a ma	Fut Coup	Persentis same or	ne and a second second	Concerner Search and and an and and and an and and and an and an and and an and an and and an a	ternenge enclosence o	Paggini I Jun	ni indogenijkog k ni indogenijkog in O nij i smeljskog ili Pasa	ren farmen Fransen Fransen	
Construction of the second sec	Microsoft Microsoft	Sources of an and a second sec	Qe 1 - Andree Funderale - Andree -	300 440 446 040 400 100 8 100 8 10	Fut Coup	Persentis same or	ne and a second second	Concerner Search and and an and and and an and and and an and an and and an and an and and an a	ternenge enclosence o	Paggini I Jun	ni indogenijkog k ni indogenijkog in O nij i smeljskog ili Pasa	ren farmen Fransen Fransen	
Print on system assessments and former AM Mill ones Environme (AM Mill ones Environme (AM Mill ones Environme (AM Mill ones) (Conter (and US) assessite Environme (AM Exclusions ENATIONES) And Environmentation and Environmentation and Environmentation and Environmentation (AM Million (AM Million and Environmentation) And Environmentation (AM Million (AM Million (AM Million (AM Million (AM Million (AM Million) (AM Million (AM Million) (AM Million (AM Million) (AM Million) (AM Million) (AM Million) (AM Milli	Maj sera kaj seg Clah Lilij ki Libi Ki	Katawan Mariana Mariana Mariana Mariana Mariana Mariana Mariana Mariana Mariana Mariana Mariana Mariana	Gr L Ander Fullman Incom Inc	Serie Late Actor according Inclusion Inclu	Fut Coup	Persentis same or	nir valadnik vi	Language of the second	antiportes o	Ner Aut	ni indogenijkog k ni indogenijkog in O nij i smeljskog ili Pasa	ren farmen Fransen Fransen	
An Agents assess prior and farms about a prior and and farms about a prior and and farms about a prior and and instance farmers ab an interaction to the term for the second to the term of the second to the second to the term of	Microsoft Microsoft	A Schwarts Co and Annual Co and Annual Co and Annual Co and Co an	Qr.4. And as Fully and Inclusion	300 400 424 000000 100000 100000 100000 10000 10000 10000 10000	Fut Coup	Persentis same or	nir valadnik vi	Concerner Search and and an and and and an and and and an and an and and an and an and and an a	antiportes o	Ner Aut	ni indogenijkog k ni indogenijkog in O nij i smeljskog ili Pasa	ren farmen Fransen Fransen	
A Contract of the second secon	Maj sere se, sep Color Ling S. Lib S. Lib	Somers Markets Ma	Angel Publick 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000 01000	Serie Las Agencies Inclus Incl	Fut Coup	Persentis same or	no oneren or no primerica primerica p nicorentitatio Base	Language of the second	antiportes o	Ner Aut	ni indogenijkog k ni indogenijkog in O nij i smeljskog ili Pasa	ren farmen Fransen Fransen	
An Agents against (A) and (A)	No, see av, see Color Ally is, LB Color Ally Solution Physics Physic	Concess or Concessor		300 400 424 000000 100000 100000 100000 10000 10000 10000 10000	G	Annual Control of Control	no conserve or no programme processive no constitution finant constitution finant constitution finant constitution finant constitution finant	Concerning	Trans Long (1)	Add	el federaj medij ka di në federaj medij ka di në j komunijano j k Ram Calendarij d	Prod No.	
Construction of the second sec	Maj sere kaj sere Culta Lalij Kaj La	A Sources of Automation of Automation of Automation of Automation of Automation of Automation of Automatical Automation of Automatical Aut		Serie Las Agencies Transa Tran	G	Annual Control of Control	no ostante di no procession ni constituin finante finante Constituine	Concerning	Sanaray Sanaray Sanaray Sanaray Sanaray Sanaray Sanaray Sanaray Sanaray	Add	el federaj medij ka di në federaj medij ka di në j komunijano j k Ram Calendarij d	ering het Pessen enselige le malifie Pare	
Control (1997) Control (1997)	No. 1992 No. 1992 No. 1992 No. 1993 No. 19	Concess or Conces	21 Joint Fallent 0.0000 0.00000 0.00000 0.0000 0.0000	2010 400 441 0 000 000 1010 00 1010 00 1000 00 1000 00 1000 00 1000 00 1000 00 1000 00 1000 00 1000 0	G	Annual Control of Control	no conserve or no programme processive no constitution finant constitution finant constitution finant constitution finant constitution finant	Concerning	Trans Long (1)	Add	el federaj medij ka di në federaj medij ka di në j komunijano j k Ram Calendarij d	Prod No.	
Description of the second seco	Majoran Majoran Colo Ling K. US K. US K. US Majoran Majoran K. US K. US	а (ранита) ала силина — со на казак марака		Serie Las Agencies Transa Tran	G	Annual Control of Control	no conserve or no programme processive no constitution finant constitution finant constitution finant constitution finant constitution finant	Concerning	Trans Long (1)	Add COTR BOX OF	el federaj medij ka di në federaj medij ka di në j komunijano j k Ram Calendarij d	ren fa men Person Person enale Person ment factor	
Control (1997) Control (1997)	Мулики мулиц Минули М Сальму К. 100 Сольму К. 100 Сольму Мулор Соле Мулор Соле Мулор Соле Марана (Сале К. 100 Соле Марана (Сале Марана) К. 100 Соле Марана (Сале) К. 100 Соле Мара	Sources Alexandre Colored Sources Sou	2 1000 10000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.000000	300 40 445 000000 100000 100000 100000 100000 100000 1000000 1000	G	Annual Control of Control	no conserve or no programme processive no constitution finant constitution finant constitution finant constitution finant constitution finant	Concerning	Trans Long (1)	Add COTR BOX OF	an internet of the second seco	ren fa men Person Person enale Person ment factor	
Control of the point of the control	Majorana Majorana Kutok				G	Annual Control of Control	no conserve or no programme processive no constitution finant constitution finant constitution finant constitution finant constitution finant	Concerning	Trans Long (1)	Add COTR BOX OF	an internet of the second seco	ren fa men Person Person enale Person ment factor	
_	Мулики мулиц Минули М Сальма К. 201 Сольма К. 201 Сольма Мала Сольма Мала Сольма К. 201 Сольма Сальма	Sources Alexandre Colored Sources Sou	2 1000 10000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.000000	300 40 445 000000 100000 100000 100000 100000 100000 1000000 1000	G	Annual Control of Control	no conserve or no programme processive no constitution finant constitution finant constitution finant constitution finant constitution finant	Concerning	Trans Long (1)	Add COTR BOX OF	an internet of the second seco	ren fa men Person Person enale Person ment factor	

Figure 4. Gene Report for KRAS includes: (A) Gene annotation data from Ensembl, Pathway annotation from (B) KEGG and (C) Reactome, (D) Summary of mutation frequencies in each cancer project, (E) Mutation data from COSMIC, Expression data from (F) Pancreatic Expression Database and (G) Breast Cancer Campaign Tissue Bank (BCCTB). Different sections of this page come from federated BioMart sources.

filters such as gene type, mutation type and chromosome, and outputs a fixed set of attributes. The *Flexible* interface contains additional filters and a selection of attributes, allowing the user to choose which data are displayed in the output. The *Advanced* interface contains the complete set of filters and attributes, including technological platforms used for sequencing, and clinical parameters such as patient gender and tumor histopathology.

To demonstrate the utility of the ICGC Data Portal we present several queries that can be performed using different query interfaces.

Query #1 (Quick Search): 'Search for genes affected by copy number loss and also detected as deletion from structural rearrangement analysis.' (Figure 6)

Data sets	Cancer projects	Filters	Attributes
Genes	Pancreatic Cancer (QCMG, AU)	Rearrangement type: deletion	Ensembl Gene ID
		Copy number alteration type: loss	Gene symbol
			Gene description

Query #2 (Flexible Search): 'Retrieve clinical staging data for colorectal cancer patients with non-synonymous simple mutations in genes that are involved in WNT signaling pathway.' (Figure 7)

Data sets	Cancer projects	Filters	Attributes
Genes	Colorectal cancer (JHU, US)	Pathway: Signaling by Wnt	Tumor sample ID
		Consequence type: non_synonymous_ coding	Mutation ID
			Mutation type
			Donor ID
			Diagnosis ID
			Clinical staging (WHO)
			Ensembl Gene ID
			Gene Symbol

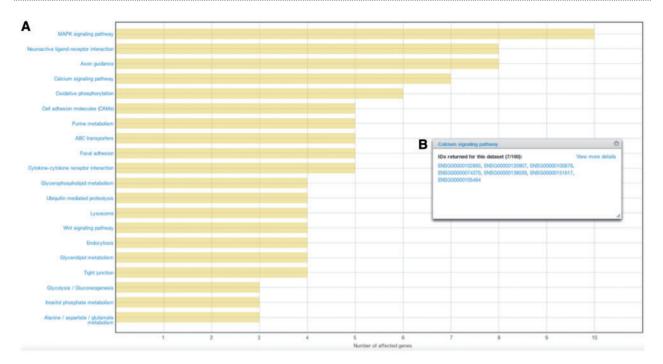


Figure 5. Results from affected pathways analysis for breast cancer (JHU, US). (**A**) Affected pathways are shown in a chart, with bars representing the number of affected genes in each pathway. (**B**) By clicking on the bar, users are able to view and download the genes that were mutated in each pathway.

1. SELECT CANCER TYPES			2. RESTRICT SEARCH		
elect multiple databases by holding down the csrs1 key, # on a Mac.		Pathway name:	Select		
Kidney Renal Papillary Cell Caroinome (FCGA, US)		Gene type:	Select		
Liver Cancer (NCC, JP)	⇒.	Circula en antica encara una a trav	Select		
Liver Cancer (PIKEN, JP)		Simple mutation consequence type:	select		
Lung Adenocaroinoma (TOGA, US)		Copy number alteration type:	loss	•	
Lung Adenocarcinoma (TSP, US)		Structural reamangement type:	deletion		
Lung Squamous Cell Cardinoma (TCGA, US)					
Malignant Melanoma (WTSI, UK)		Novel splice form:	Select		
Ovarian Serous Cystadenocarcinoma (TCGA, US)		Fusion gene:	Select		
Pancreatic Cancer (JHU, US)		Methylation beta value:	No deta	1.01	
Pancreatic Cancer (OICIR, CA)			INF BILLS		
Panovatic Canoir (DCMG, AU)		Clinical staging (WHO):	Select		
Rectum Adenocarcinoma (TOGA, US)		Entries with following IDs:	Select		
Small Cell Lung Cardinoma (WTSI, UK)					
Stomach Adenocarcinoma (TCGA, US)					
Uterine Corpus Endometricid Carcinoma (TCGA,					

Figure 6. A screenshot of the Quickquery interface.

	CANER	111D	
low:	Genes		
ancer types:	Acute Myeloid Leukemia (Breast Carcier (HU, US) Breast Carcinema (MYS), U Breast Invasive Carcinoma Colon Adenocarcinoma (T Celosectal Cancer (HU, U) Ciloblastoma Multiforme (Ciloblastoma Multiforme (Kidney Renal Clear Cell Ca Kidney Renal Clear Cell Ca	N) (TCCA, US) CCA, US) (HU, US) TCGA, US) TCGA, US) (TCGA, US)	
	Filt	RS	
PATHWAYS			
Pathway name:			
Signaling by Wint	. 0		
	TIONE		
SIMPLE SOMATIC MUTA			
Mutation type:	141	Consequence type: non_synonymous_coding	. 0
Annotated: Select	1.0	Validation status: Select	
	Аттка	UTES	
GENERAL	Artike	utts	
GENERAL		UTIS	
GENERAL Cancer Type			
GENERAL Cancer Type GENE	Assembly Version		
GENERAL Cancer Type GENE Ensembl Gene ID	Assembly Version		
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype	Assembly Version		
GENERAL Gancer Type GENE Gene Biotype EXPERIMENT DATA	Assembly Version		
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS	Assembly Version		
GENERAL Gancer Type GENE Ensembl Gene ID Gene Biotype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION	Assembly Version Ensembl Transcript ID	Gene Symbol	
GENERAL Cancer Type GENE Ensembl Gene ID Gene Biotype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION MUTATION TUMOUT sample ID	Assembly Version Ensembl Transcript ID Mutation ID	Gene Symbol	
GENERAL Gancer Type GENE GENE Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION Turnour sample ID Mutation	Assembly Version Ensembl Transcript ID Mutation ID Consequence type	Gene Symbol Mutation type Chromosome	
GENE Gene Biotype EXPERIMENT DATA SIMPLE SOMATIC MUTATION MUTATION MUTATION C Turnour sample ID C Mutation C Chromosome start	Assembly Version Ensembl Transcript ID Mutation ID Consequence type	Gene Symbol Mutation type Chromosome	
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION Tumour sample ID Mutation Chromosome start Is annotated	Assembly Version Ensembl Transcript ID Mutation ID Consequence type	Gene Symbol Mutation type Chromosome	
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION Tumour sample ID Mutation Chromosome start Is annotated DONOR	Assembly Version Ensembl Transcript ID Mutation ID Consequence type	Gene Symbol Mutation type Chromosome	

Figure 7. A screenshot of the Flexiblequery interface.

EXPERIMENT DATA			1	Sciences	o view XM
SIMPLE SOMATIC MUTATIONS CO	PY NUMBER SOMATIC M	UTATIONS		View Genes	
STRUCTURAL SOMATIC MUTATIONS		JUNCTION mIRNA		Cancer Types Pancreatic Cancer (GCMO, AU)	
EXPRESSION				Filters Mutation type: gain	
Sample ID	Normalized re	ad count		Platform: SOLD sequencing	
Raw read count	Validation stat	us		Attributes Ensembl Gene ID	
Platform				Sample ID	
				Raw read count	
ANALYSIS				Normalized read count	
Base calling algorithm	Alignment alg	prithm			
Raw data repository	Raw data acc	ession	_		

Figure 8. A screenshot of the Advanced query interface.

Query #3 (Advanced Search): 'In pancreatic cancer data set, retrieve all RNA-seq expression data for genes that are affected by copy number gains.' (Figure 8)

Data sets	Cancer projects	Filters	Attributes
Genes	Pancreatic Cancer (QCMG, AU)	Copy number alteration type: gain	Ensembl Gene ID
		Platform: SOLiD sequencing	Gene Expression: Sample ID
			Normalized read count
			Raw read count

Conclusion and future directions

The ICGC Data Portal is the first project to successfully federate large amounts of cancer genomics data and rich annotation data in a single access point. It presents a scalable approach, not only in the traditional sense of parallelizing data processing and storage, but also in a more general sense of outsourcing the external annotation expertise, by federating annotations from independently maintained databases. This approach has proved to be successful in addressing ICGC data management needs and can be useful for similar, large-scale collaborative projects.

The ICGC Data Portal will continue to expand by adding more data, both from within the project in the form of new cancer genomics data from ICGC members, and by integrating other public annotation databases to further the depth of analysis possible through the data portal. Additionally, new tools will be developed to increase the flexibility and utility of the system. This includes tools for ICGC deployers, to further streamline the data processing, transformation and loading processes, and tools for users, to add new methods of data visualization and analysis to the portal.

Acknowledgements

The authors would like to thank members of the International Cancer Genome Consortium who have provided data, use cases, and/or input into the ICGC Data Portal. The authors are also grateful to the programmers and bioinformaticians who have contributed to the BioMart project over the course of its development.

Funding

Ontario Institute for Cancer Research and the Ontario Ministry for Research and Innovation. Funding for open access charge: XXX.

Conflict of interest. None declared.

References

- 1. Hudson,T.J., Anderson,W., Artez,A. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**(7291), 993–998.
- 2. Zhang, J., Haider, S., Guberman, J.M. *et al.* (2011) BioMart: a data federation framework for large collaborative projects. *Database* (This issue).
- 3. Guberman, J.M., Arnaiz, O., Baran, J. *et al.* (2011) BioMart central portal: an open database network for the biological community. *Database* (This issue).
- Haider, S., Ballester, B., Smedley, D. et al. (2009) BioMart Central Portal–unified access to biological data. Nucleic Acids Res., 37, W23–W27.
- 5. Kinsella,R.J., Kahari,A., Haider,S. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across the taxonomic space. *Database* (This issue).

- 6. Shepherd,R., Forbes,S.A., Beare,D. *et al.* (2011) Data mining using the catalogue of somatic mutations in cancer BioMart (COSMICMart). *Database* (This issue).
- 7. Haw,R., Croft,D., Yung,C. *et al.* (2011) The reactome BioMart. *Database* (This issue).
- 8. The Cancer Genome Atlas. http://cancergenome.nih.gov/.
- 9. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
- 10. Ding,L., Getz,G., Wheeler,D.A. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.
- Jones, S., Zhang, X., Parsons, D.W. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.
- Parsons, D.W., Jones, S., Zhang, X. et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321, 1807–1812.
- Sjoblom, T., Jones, S., Wood, L.D. et al. (2006) The consensus coding sequences of human breast and colorectal cancers. Science, 314, 268–274.

- Wood,L.D., Parsons,D.W., Jones,S. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
- 15. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28, 27–30.
- Kanehisa, M., Goto, S., Furumichi, M. et al. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res., 38, D355–D360.
- Kanehisa, M., Goto, S., Hattori, M. et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res., 34, D354–D357.
- 18. Cutts, R.J., Emanuela, G., Lemoine, N.R. *et al.* (2011) Using BioMart as a framework to manage and query pancreatic cancer data. *Database* (This issue).
- 19. Breast Cancer Campaign Tissue Bank. http://breastcancertissuebank .org/bio-informatics.php.