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

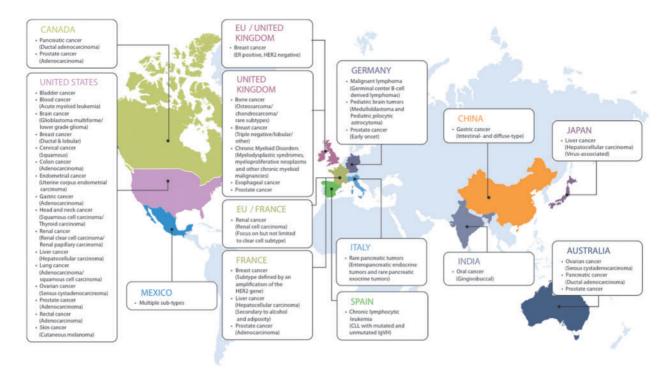


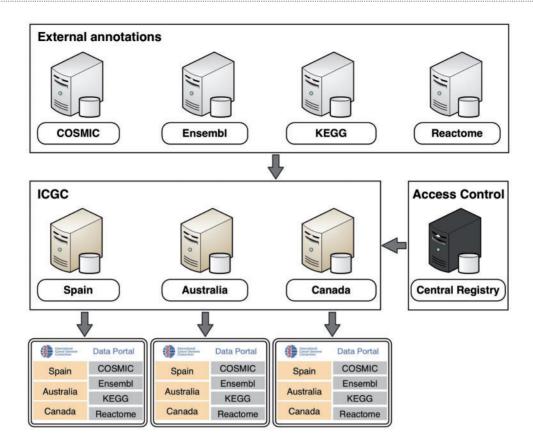
Figure 1. International Cancer Genome Consortium (ICGC) projects (March 2011).

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,

Table 1.	Table 1. The summary of data available on the ICGC Data Portal divided by cancer project	GC Data Portal	divided by can	cer project					
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 (RKEN, 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 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)								

Table 2. Public databases federated with the ICGC Data Portal

Source	URL	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

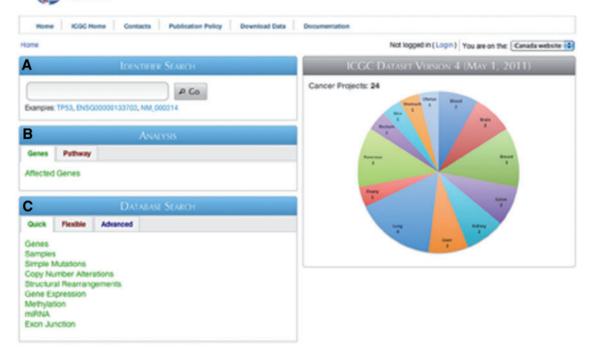


Figure 3. Screenshot of the ICGC Data Portal home page. Three main entry points are available: (A) Identifier search, (B) Analysis and (C) Database search.

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

A Gest lero					E COBVEC						
Inwented Gene II	D: ENSGROOM 11723 P.R.A.	ទា			_		ABOUT THE ! I	No. Andreas Property of the	INVALUE AND POST OF	CAMPAGE/AND 1 10	8181-100-10 de
					***	1996	Mallum	0.047	SAULUS READ	2489	1.5904
tion is private	+#122 wilds it because a conduct	a hanna and a			-		helitan	CANNA.	Intelligibility Bellands	patters	COMPANY
	service Device (CAL Synthetics (MCT)	Form 2nd Rev.	100491								240
					1018	294	Endneren.	8.7	Edult.ton Bisem	+0.01	-
Care ford those	3-108-1	and a	4		anus.		10.64	SPI, MICH	SHORAD REAL PROPERTY.	1.85/5	ALCOHOL: N
Let.	ALC .	Cons Branger	2-2-0, 1-10y		-	-	NAMES OF T	AND	Sauthan Room	para	V-MARCE
And an International Name	Water of the Property of the P	Danielles dendry a	elana, Dana tasimita	is break as they add to.							
	Anderson Barry with Relationship and the										
-					-	1997	NAMES AND	1.000.00	NAME AND BROADS	0.0100	1175,000
NUMBER OF STREET	bend bewelt children bowarder harves	Also when it fight	AT THE AREA RECIPCION	 require many 							1000
pusha any basic sign	die sales, Gritt speingschus, Pepu	cross mediated benyt	ruturation, tableringer	usis, there we repared	-	100010	termine.	44	1000	p.e.	16.0
of all index place, they	ways is care to. Colorected control, Rend Cally,	on home, Par roads	property Restoration (are	on, Brown, Personancercar,	***	101	COMPRO- MODERUP	<9,9004F	Datable July- Reinardas	PO 90	110.000
NAMES ADDRESS OF TAXABLE	A Market Lances, Concest repaired instantia, in		Internet and American	-			CARACTURE.				
					10.0		Collans.	434,3856/7	Datalik helo- distrator	1404	111000
C	he if you also signify an out, figuring is	Inchestry and	openingen, RD-au	inclusion, Bernstein, 1818							244
	AND REAL PROPERTY AND ADDRESS OF THE OWNER,				4516	2000	Complex-	10.1014	LANKA, ANY REPORT	*508	0,090
							distant.				70,04
period to VOL Band	No 1977E, Development within, Provi	CONTRACTOR NO.	IR, NO PRODUCE	any, enderstand			10.0			Internet Canada and David	
arkadar, philippinases	na, Tagonaling, ng USAPA, Banah asaminu'n BSAPA a y	paing the state in t	Officiarsky low-w	Namindanik Namaka Kati	-						
STREET, STATE AND INC.	grang is tea for he pairs hereign.	Sprang Sounds, Sul		CARLS AND INCOME AND INCOME.	_	1114347	A D IN MARY				
				the Walter of the local data							
ar ber nepalate stall, frag (Brains KOM dandrate scale or presi	Ane patteron. By-	elegia interduction, irm	e malak 2 a geodera							
n be see de rei hel		, Anne galanton, Bayr In Schneidte	elegia Hinduina, Ira	enslar 2 x peaks	F			Percent law	mon Das		
-	McCarlo	a Sciences		Instanting	•				and a second second		6 .00
D	Muji sata Bargin ku	a Sciences			Fell Oringe	971036-789 (M)		Courses and	Interior	an Altrophysics	NE-ALE CRIME
	Microso Degeter Society	a Sciences			Fell Oringe	-01121/01212	10.22021-07	Courses and	Interior		NE-ALE CRIME
	Microso Degeter Society	a Sources	operation includes	Texas fee agrie 1	Fell Oringe		10.1203210	Courses and	Interior		NE-ALE CRIME
Denne Green (PA) (A)	McCarel Bright Ad Ad	A School C	igel. Alder Kr. Miles Sec. 19	The a large fragmanical	Fell Oringe	-Out-Mircold #51	increases to	Courses Sensitives and activities	Including Including Including Including Including	essection de la recentra	arts and
D constant insum Garware (244) (48) insum Garware (244) (48) insum Garware (244)	Miran Kala ol Kala ol Gala da	a Sharen	April Andre Multicale Arrienta April Andrea Arrienta	The a loss face angement	Fell Oringe	economication (M) consumerous and (M) Facarmine same con	increases to	Courses	Including Including Including Including Including	internation in Courts	and state
D transform and the pro- transform (PA) off transformer (PA) of	Microsoft Notes, and Al- policies, and Constant Al-	Notes Contest	ogel Johone Roberte Science Science Science Science	Ber allere Rate singereent Hermal Hermal Hermal Hermal	Fell Oringe	-Out-Mircold #51		Courses	Including Including Including Including Including	internation in Courts	and state
Constitution (PA) (A) Install Garware (PA) (A) Install Garware (PA) (A) Install Install of Constantion Constitution (Constantions) Constalling Constantions)	Miriana Naki Mirika Naki Mirika Naki Mirika Naki Mirika Naki Mirika	Sources		De chat fear argeneri Innia Innia Innia Innia Innia Innia	Fell Oringe	Paramite tone or	ni galerika	Concerner Search and and an and and and an and and and and an and an and and an and an and and an an and an and and an a	analatina deservas nacional deservas nacional deservas nacional deservas	internet and the first first state of the second state of the seco	
Deserver Insue Gartere (PALLA) Insue Gartere (PALLA) Insue Gartere (PALLA) Insue Gartere (PALLA) Insue (Insuence (PALLA) Insue (Insuence (PALLA) Insue (Insuence (PALLA)) Insue (Insuence (PALLA))	We need	A School C		The advancement of the second se	Fell Oringe	Paramite tone or	ni galerika	Courses	analatina deservas nacional deservas nacional deservas nacional deservas	interpreta la Cros fa	
Dente on Martin and any p form forward (244) Mil form forward (244) Mil form formation (244) Million (244) Million (244) Million (244) Million (244) Million (244) Million (245) Million (244) Million (245) Million (245) Million (245) Million (2	Victoria New, With Enry R M (Control of Control of Cont	N SUMOUTS	April Johan Paulitak Merikan Merikan Merikan Merikan Merikan Merikan Merikan	An Les les agents innes rec rec rec reces rec reces reces reces reces reces reces reces reces reces rec rec rec rec rec rec rec rec rec rec	Fell Oringe	Paramite tone or	ni galerika	Concerner Search and and an and and and an and and and and an and an and and an and an and and an an and an and and an a	analatina deservas nacional deservas nacional deservas nacional deservas	interpreta la Cros fa	
Design of the second se	We need	N SUMOUT		Stor a Local-Andre sequences	Fell Oringe	Paramite tone or	ni galerika	Concerner Search and and an and and and an and and and and an and an and and an and an and and an an and an and and an a	analatina deservas nacional deservas nacional deservas nacional deservas	interpreta to from fo the system of the form for the system of the system for the system of the syst	nasigi maristi nasista
Designation of the second seco	Without With Comparison Particular Autor Particular Autor Parti	N SOMORY VICE OF ANY ANY CONTRACT CONTRACTOR CONTRA	12 (A) (A) (A) (A) (A) 12 - 12 (A) 12 -		Fell Oringe	Paramite tone or	ni galerika	Concerner Search and and an and and and an and and and and an and an and and an and an and and an an and an and and an a	analatina deservas nacional deservas nacional deservas nacional deservas	interpreta la Cros fa	nasigi maristi nasista
Contract Contract (Contract Contract Co	Microsoft Allo Allo Callos Allo Callos Allos Callos Allos Ca	A Scherenti and Construction Con- state state state state Scherenti Sc		En o Los Ano expenses encos reces rec reces rec reces reces reces reces reces reces reces reces reces	Fell Oringe	Paramite tone or	ni galerika	Concerner Search and and an and and and an and and and and an and an and and an and an and and an an and an and and an a	analatina deservas nacional deservas nacional deservas nacional deservas	interpreta to from fo the system of the form for the system of the system for the system of the syst	nasigi maristi nationa
Design of the Application and an energy of the Applications and an energy of the applications of the Applications in the the A	Without Without Strength Million Strengt	Statutes ACC Available Constant	101 Anna Antonio Marina	301.040 fag cignred 100.08 1	Fell Oringe	Paramite tone or	ni galerika	Concerner Search and and an and and and an and and and and an and an and and an and an and and an an and an and and an a	analatina deservas nacional deservas nacional deservas nacional deservas	interpreta to from fo the system of the form for the system of the system for the system of the syst	nasigi maristi nasista
Design of the second se	We need	 Streets Antibiotics 		Billy Las Age age and minute minute resulta re	Fell Oringe	Paramite tone or	ne) protection (net constitution) net constitution)	Language of the second	la tangg moporti sen tan tan Bagin ti mulatnan Jawa mulatnan Jaka	interpreta to from fo the system of the form for the system of the system for the system of the syst	nasigi maristi nasista
Protect one hypothesistemes pro- ment forware 1944 Mil- ment forware 1944 Mil- ment forware 1944 Mil- ment forware 1944 Mil- ment forware 1944 Mil- Million 1944 Million 1944 Mil- Million 1944 Million 1944 Million Million 1944 Million 1944 Million 1944 Million Million 1944 Million 1944 Million 1944 Million 1944 Million Million 1944 Million 1944 M	Microsoft Nine, Wij Lingel B (A) (A) (A) (A) (B) (A)	Sources S		301.440 449 449 449 449 101.00 100 1	Fell Oringe	Paramite tone or	ne) protection (net constitution) net constitution)	Language of the second	analatina deservas nacional deservas nacional deservas nacional deservas	interpreta to from fo the system of the form for the system of the system for the system of the syst	nasigi maristi nasista
Construction of the second sec	Vectors teau, seg A0 Concept to A0 Concept to A0 Conce	k Schweite Heit Austrick Color Heit Australie Heit Australi	Gel Ander Fullende Incinité Incini	201 40 40 500 400 101 50 101 50 100 101 50 101 50 100 1	G	Paramite tone or	ne) protection (net constitution) net constitution)	Exercises and property of subscription 2, to subscription 2, to subscription 2, to subscription 2, to subscription 2, to subscription 2, to	Anterio operante montres operante montres Just montres Ante	interpreta to from fo the system of the form for the system of the system for the system of the syst	nasigi maristi nasista
Period one hypothypothypothypothypothypothypothypot	Without With Comparison The Comparison	Solaron Alexandre	Contact Futures Reference Refer	300.440 449 copress 100.00 100000000	Fell Oringe	Pesendi sere ori Pesendi sere ori Pesen	normanie w normanie powerkani nie oceanie anie nie oceanie anie oce	Concerning and an and a second	Anna Kan (KLTR	interpretation in the first of	
Construction of the second sec	Victoria AD English	A Sources A Constant Constant Sources		2012/2014/2019 00/00000 1010/20 100/20 100/20 1010/20 1010/20 1010/20	G	Persenti tenerori Persenti tenerori Persenti 1	normanie w normanie powerkani nie oceanie anie nie oceanie anie oce	Concerning and a second and a s	Anterio operante montres operante montres Just montres Ante	integrandjang karter Intergrandjang og Ennerginger Stanson Stanson Franklik Stanson Frank Stanson Franklik S	
Construction of the second sec	Microsoft Nick, Mill Except in (2004) Life Except in (2004) Life <td< td=""><td>Scheelers Market Scheelers Market Scheelers Schelers Scheelers Schelers Schel</td><td>Ander Publication 00000</td><td>201 40 40 500 400 101 50 101 50 100 101 50 101 50 100 100 100 100 100 100 100 1</td><td>G</td><td>Persenti tenerori Persenti tenerori Persenti 1</td><td>no odanjete se no je na odvenička je koratika ko Bane Bane Oda</td><td>Concerning and a second and a s</td><td>Annego Hospitela desegues Hospitela Jusi Hospitela Main Tracellas (KCTR Hospitela Bosco</td><td>integrandjang karter Intergrandjang og Ennerginger Stanson Stanson Franklik Stanson Frank Stanson Franklik S</td><td></td></td<>	Scheelers Market Scheelers Market Scheelers Schelers Scheelers Schelers Schel	Ander Publication 00000	201 40 40 500 400 101 50 101 50 100 101 50 101 50 100 100 100 100 100 100 100 1	G	Persenti tenerori Persenti tenerori Persenti 1	no odanjete se no je na odvenička je koratika ko Bane Bane Oda	Concerning and a second and a s	Annego Hospitela desegues Hospitela Jusi Hospitela Main Tracellas (KCTR Hospitela Bosco	integrandjang karter Intergrandjang og Ennerginger Stanson Stanson Franklik Stanson Frank Stanson Franklik S	
Construction of the second sec	Коллон М. Сону К. Коллон М. Коллон М. К	Solutions Markets		201 43 43 40 40 40 101 43 101 43 10	G	Persenti tenerori Persenti tenerori Persenti 1	no odanjete se no je na odvenička je koratika ko Bane Bane Oda	Concerning and a second and a s	Anne Ras (RCT)	intercontrigens, include intercontrigens, include the second second second second second second second distances from second distanc	
Design of the second se	Microsoft Nick, Wij Europie III (2014) Lillij 201 (2014) Lilliji 201 <	A Sources A Constant Constant Sources		201 La 42 Consta 1000	G	Persenti tenerori Persenti tenerori Persenti 1	no odanjete se no je na odvenička je koratika ko Bane Bane Oda	Concerning and a second and a s	Anne Ras (RCT)	integrandjang karter Intergrandjang og Ennerginger Stanson Stanson Franklik Stanson Frank Stanson Franklik S	
Control System control on the Control of Control on the Control of Control on the Control of Control on the Con	Коллон Маллени Коллон Маллении Коллон Маллении Коллон Маллении Коллон Маллении Коллон Маллении Ко	A Sources	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	20140 448 449 449 449 10140 100100 100000000	G	Persenti tenerori Persenti tenerori Persenti 1	no odanjete se no je na odvenička je koratika ko Bane Bane Oda	Concerning and a second and a s	Anne Ras (RCT)	intercontrigens, include intercontrigens, include the second second second second second second second distances from second distanc	
-	Microsoft Nick, Wij Europie III All IIII Spicial All IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	A Solutions A Constant of a second		20142 448 449 449 449 10142 1014 1014 1014 1014 1014 1014 1014 1014 1	G	Persenti tenerori Persenti tenerori Persenti 1	no odanjete se no je na odvenička je koratika ko Bane Bane Oda	Concerning and a second and a s	Anne Ras (RCT)	intercontrigens, include intercontrigens, include the second second second second second second second distances from second distanc	 Interpreter program in the program in the last program in the last program in the last program in the last program is the last program in the last program is the last program in the last program is the last progra

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

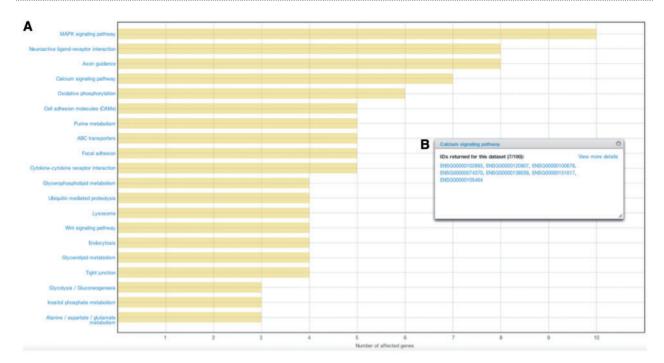


Figure 5. Results from affected pathways analysis for breast cancer (JHU, US). (A) Affected pathways are shown in a chart, with bars representing the number of affected genes in each pathway. (B) By clicking on the bar, users are able to view and download the genes that were mutated in each pathway.

1. SELECT CANCER TYPES		2. RESTRICT SEARCH		
elect multiple databases by holding down the curst key, r # on a Mac.	Pathway name:	Select		
Kidney Renal Papillary Cell Caroinoma (TCGA, US)	Gene type:	Select		
Liver Cancer (NCC, JP)	Single mutation consequence type:	Select	140	
Liver Cancer (FIRCEN, JP)	ompre metalion consequence rype.	select	•	
Lung Adenocarcinoma (TCGA, US)	Copy number alteration type:	loss	•	
Lung Adenocarcinoma (TSP, US)	Structural rearrangement 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 data	100	
Panorestic Canoer (CICR, CA)		INV GULA		
Pancreatic Cancer (OCMG, AU)	Clinical staging (WHO):	Select		
Rectum Adenocarcinome (TCGA, US)	Entries with following IDs:	Select		
Small Cell Lung Cardinoma (WTSI, UK)				
Stomach Adenocarcinoma (TCGA, US)				
Uterine Corpus Endometrioid Carcinoma (TCGA, US)				

Figure 6. A screenshot of the Quickquery interface.

lew:	(Canada			
ancer types:	Cenes Acute Myeloid Leuke Breast Cancer (HU, U Breast Carcinoma (W Breast Invasive Carcin Colon Adenocarcinee	mia (TCGA, US) S) ISI, UKO Ioma (TCCA, US)	Ī	
	Colorectal Cancer (IH) Clioblastoma Multifo	U, US) me (HU, US)		
		me (TCGA, US) ell Carcinoma (TCGA, US) y Cell Carcinoma (TCGA, US)		
		Filters		
PATHWAYS				
Pathway name:				
Signaling by Wint				
SIMPLE SOMATIC MUTA	TIONS			
Mutation type:		Consequence type:		
Select		non_synonymous_cod	ling	0
Annotated:		Validation status:		
Select		Select		
	A	TRIBUTES		
GENERAL	A Assembly Version			
ATTRIBUTES GENERAL Geneer Type GENE				
GENERAL Cancer Type		· · · · · ·	s Symbol	
GENERAL Cancer Type GENE Ensembl Gene ID	C Assembly Version	· · · · · ·	e Symbol	
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype	C Assembly Version	· · · · · ·	s Symbol	
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype	C Assembly Version		• Symbol	
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION	C Assembly Version	pt ID 🦉 Gene		
GENERAL Cancer Type GENE Ensembl Gene ID Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS	C Assembly Version	pt ID 🦉 Gene	 Symbol tation type 	
GENERAL Gencer Type GENE Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION	C Assembly Version	pt ID Migen		
GENERAL Gancer Type GENE Gene Biotype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION Turnour sample ID	C Assembly Version	pt ID ⊠ Germ Ø Mu pe ⊡ Ch	tation type	
GENERAL Gancer Type GENE GENE Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION Grunour sample ID Guidation	Assembly Version Ensembl Transcrip Mutation ID Consequence to	pt ID ⊠ Germ Ø Mu pe ⊡ Ch	tation type romosome	
GENERAL Gencer Type GENE Gene Bictype EXPERIMENT DATA EMPLE SOMATIC MUTATIONS MUTATION Grunour sample ID Grunour sample ID Grunosome start Is annotated	Assembly Version Ensembl Transcrip Mutation ID Consequence to	pt ID ⊠ Germ Ø Mu pe ⊡ Ch	tation type romosome	
GENERAL Gancer Type GENE GENE Gene Bictype EXPERIMENT DATA SIMPLE SOMATIC MUTATIONS MUTATION Turnour sample ID Mutation Chromosome start	Assembly Version Ensembl Transcrip Mutation ID Consequence to	pt ID ⊠ Germ Ø Mu pe ⊡ Ch	tation type romosome	
GENERAL Gencer Type GENE Gene Bictype EXPERIMENT DATA BMPLE SOMATIC MUTATIONS MUTATION Gromosome start Is annotated DONOR	Assembly Version Ensembl Transcrip Mutation ID Consequence to	pt ID Gen pt ID On pe On pd On	tation type romosome	

Figure 7. A screenshot of the Flexiblequery interface.

EXPERIMENT DATA			10	Schuler	co 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			ווו ר	Filters Mutation type: gain	
Sample ID	Normalized re	ad count		Platform: SOLD sequencing	Ċ
Raw read count	Validation stat	us		Attributes Ensembl Gene ID	
C 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 Advancedquery 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

- 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).

- 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.
- 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.