ICGC International Cancer Genome Consortium in the cloud

Integrated cloud computing infrastructures for ICGC data

Christina K Yung[1], Guillaume Bourque[2], Paul C Boutros[1,3], Khaled El Emam[4], Bartha M Knoppers[5], Brian O'Connor[1], BF Francis Ouellette[1,3], Cenk Sahinalp[6], Sohrab P Shah[7], Vincent Ferretti[1], Lincoln D Stein[1,3], Cancer Genome Collaboratory Consortium

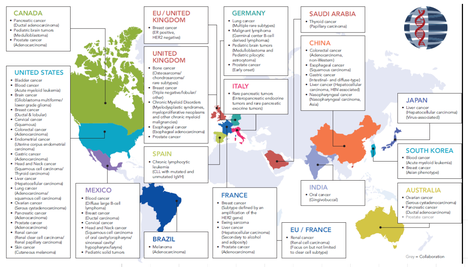
[1]Ontario Institute for Cancer Research, Toronto, ON, Canada; [2]McGill University, Montreal, QC, Canada; [3]University of Toronto, Toronto, ON, Canada; [4]University of Ottawa, Ottawa, ON, Canada; [5]Centre of Genomics and Policy, Montreal, QC, Canada; [6]Simon Fraser University, Vancouver, BC, Canada; [7]BC Cancer Agency Research Centre, Vancouver, BC, Canada

ABSTRACT

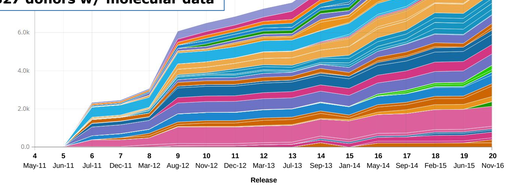
In November 2015, members of this consortium and the International Cancer Genome Consortium (ICGC) jointly announced the availability of whole cancer genomes in the Amazon Web Services (AWS) and the Cancer Genome Collaboratory. To date, more than 1400 whole cancer genomes are available in AWS' Elastic Compute Cloud (EC2), and more than 1900 in the Cancer Genome Collaboratory, an academic cloud being built by this consortium. By making the data available in cloud compute form, researchers benefit from the high availability, scalability and economy offered by cloud services, and to avoid the large investment in compute resources and the time needed to download the data. Over the next year, we will increase the number of ICGC genomes available in the cloud, with the goal of placing the entire ICGC data set of ~25,000 donors in academic and commercial clouds when the project is completed in 2018. For information and a getting-started guide, see https://dcc.icgc.org/icgc-in-the-cloud.

What is ICGC?

ICGC Goal: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumour types and/or subtypes which are of clinical and societal importance across the globe. As of April 2016, ICGC has 88 projects in 17 jurisdictions.



Release 20 (December 2016):
- **66 projects**
- **17,927 donors total**
- **14,827 donors w/ molecular data**



What is PCAWG?

A practical demonstration of the power of working in compute clouds comes from our ongoing collaboration with the PanCancer Analysis of Whole Genomes Project (PCAWG; https://dcc.icgc.org/pcawg), which seeks to interpret patterns of variation in both coding and non-coding portions of cancer genomes. Upwards of **2800 ICGC whole cancer genomes** were subjected to a uniform data processing pipeline that included whole genome alignment, uniform quality control, and standardized germline and somatic variant calling using a large number of software packages that were adapted to run efficiently in the cloud.

Using a series of **14 academic and commercial compute clouds**, we were able to process this **800 terabyte data set** in just over a year's time. Given the improvements in the software that occurred over this period, the whole project would take less than 4 months on just a single commercial cloud if we were to start over. When the project is completed later in 2016, we will again use academic and compute clouds to publish the PCAWG data, its major results, and all the software used during the analysis, thereby allowing the research community to integrate PCAWG with their own data sets, and apply the same analytic procedures.

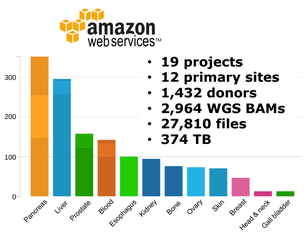
Why Cloud Computing?

Cloud computing represents a fundamental shift in the way that cancer genomics is performed. Because of the large size of the ICGC data set, it can take many months to download the data across a typical university broadband connection, and it requires a substantial investment in hardware in order to analyze it. In practice, this has meant that only large computational groups could perform whole-genome analysis at scale. Using the cloud, research groups of any size can launch large analytic processes, pay only for the compute that they use, and avoid charges for data transfer and long-term data storage.

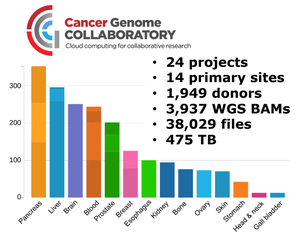
	Size	Transfer on dedicated 10G link	Transfer on shared University link
PCAWG data set today	0.8 PB	9 days	8 months
ICGC data set in 2018	5.0 PB	2 months	4 years

Two Clouds Hosting PCAWG Data: Amazon Web Services – commercial Cancer Genome Collaboratory – academic

amazon web services™
- **19 projects**
- **12 primary sites**
- **1,432 donors**
- **2,964 WGS BAMs**
- **27,810 files**
- **374 TB**

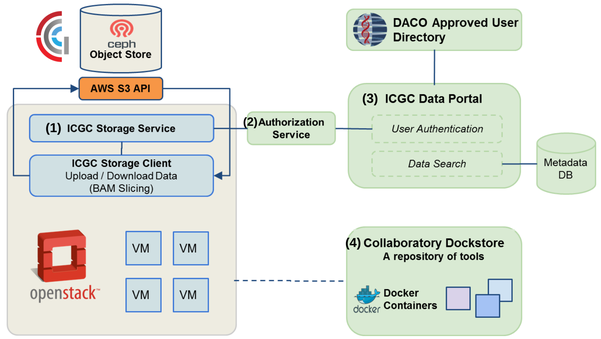


Cancer Genome COLLABORATORY
Cloud computing for collaborative research
- **24 projects**
- **14 primary sites**
- **1,949 donors**
- **3,937 WGS BAMs**
- **38,029 files**
- **475 TB**



Current data holdings include whole-genome sequencing (WGS) BAMs and variants from 4 variant calling pipelines (Sanger, DKFZ/EMBL, Broad and Muse). In the near future, merged variant sets, mini BAMs, and RNA-Seq alignments will also be available.

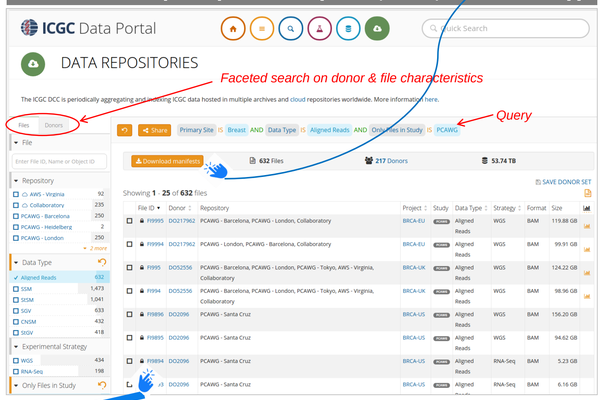
Infrastructure Overview



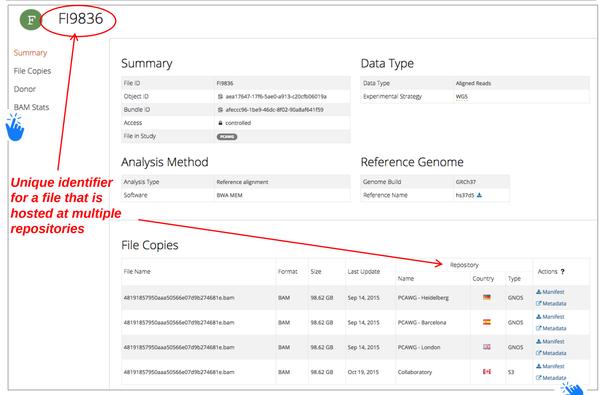
The "ICGC in the cloud" software infrastructure includes four main components: (1) a client-server storage module used to upload/download large data files to/from the Ceph object store; (2) an authentication and authorization module to ensure that only approved researchers from the ICGC Data Access Compliance Office (DACO; https://icgc.org/daco) can access ICGC control-tier data; (3) a user-friendly data search tool in the ICGC Data Portal and finally, (4) an open platform called Dockstore (dockstore.org) to share data analytic pipeline packaged in Docker containers. While the above figure depicts the infrastructure at Collaboratory, a similar system is used at AWS but with S3 in place of Ceph, and EC2 instances in place of OpenStack VMs.

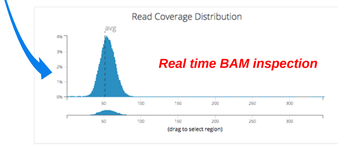
How to Search and Download Data?

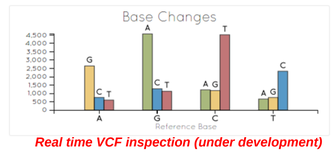
1. Perform faceted search at the ICGC Data Portal for data hosted at multiple repositories (GNOS, AWS, Collaboratory)



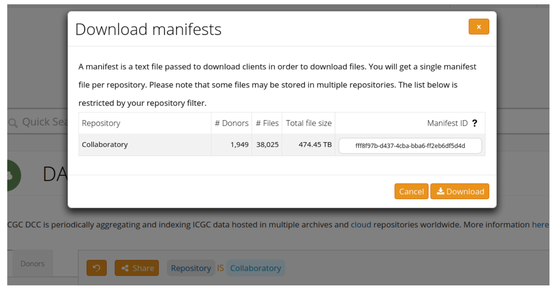
2. View metadata and other details on the file entity page







3. Save the list of files as a manifest



4. Download the files using the ICGC Storage Client

The ICGC Storage Client can be run in an EC2 instance or OpenStack VM. It supports:
- high-throughput multi-part download
- resumability after transfer interruption
- fault tolerant against data corruptions in transit
- MD5 checksum validation
- BAM slicing
- Filesystem in Userspace (FUSE)

Download manifest data

```
%: icgc-storage-client download --manifest 4jdyyqs099ew22
--output-dir data --output-layout bundle
```

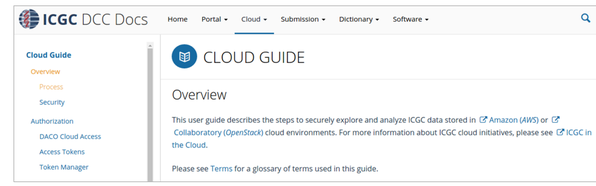
Download BAM slices

```
%: icgc-storage-client view --object-id ea17647-17f6-5ae0
--query 12:25357723-25403870
```

Mounting a manifest (FUSE)

```
%: icgc-storage-client mount --manifest 4jdyyqs099ew22 --
mount-point /tmp/
%: ls /tmp
```

Comprehensive User Guide on docs.icgc.org



Please contact help@cancercollaboratory.org for more information

Acknowledgements

INNOVATION.CA CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
NSERC CRSNG
Ontario
OICR Ontario Institute for Cancer Research
PCAWG PanCancer Analysis of WHOLE GENOMES
GenomeCanada
CIHR IRSC Canadian Institutes of Health Research | Instituts de recherche en santé du Canada