Cancer Genome Analysis for Everyone: The Cancer Genome Collaboratory

Michelle Brazas, OICR

Cancer: A Disease of the Genome

- An accumulation of genomic alterations can lead to unregulated cell growth
- *One-size-fits-all standard treatment* models do not take into account distinct molecular characteristics of each tumour
  - Response to standard therapy is highly variable
- Development of *targeted therapies* needs a comprehensive catalogue of molecular alterations and models of how these alterations give rise to tumour phenotype

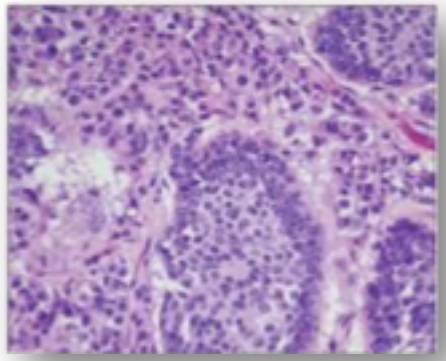
International Cancer Genome Consortium (ICGC)

- Launched in **2008**
- Goal: To sequence **25,000** tumor genomes (with matched normal) across **50** tumor types or subtypes by 2018
- Make this '**big 1.5PB cancer data**' available to community & public

Donor's normal genome

...GATTATTCCAGGTAT...

Donor's tumor genome

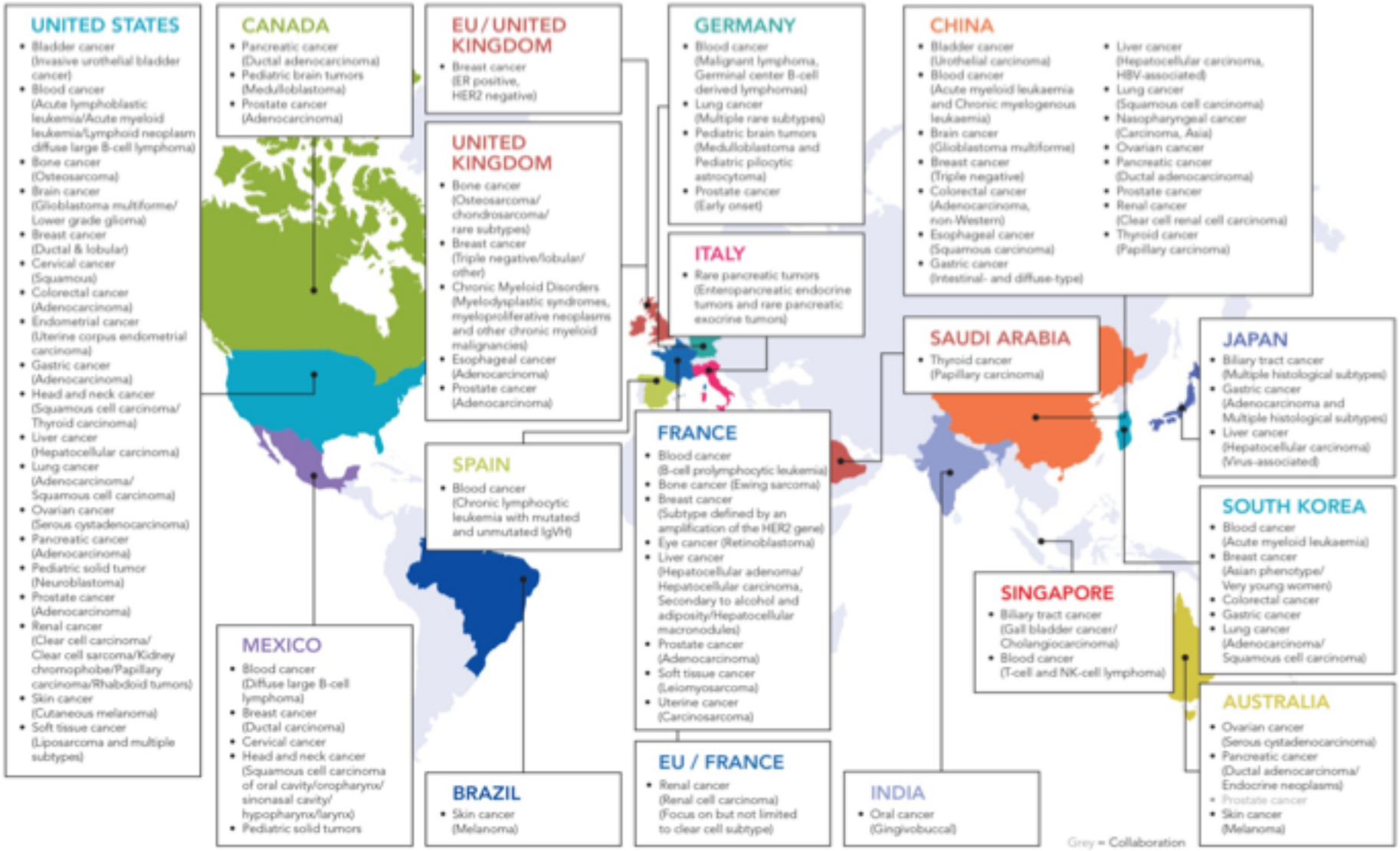


...GATTATTGCAGGTAT...

Catalogue genomic changes

...GATTATTGCAGGTAT...

April **2017** - Commitment from **107 ICGC Projects** for **29,000+ tumor genomes**



PCAWG
PanCancer Analysis OF WHOLE GENOMES

- Goal: Understand the nature and consequences of somatic and germline variations in both coding and non-coding regions of the cancer whole genome and to identify common patterns of mutation
  - 580+ researchers
  - 130+ research projects
  - 16 thematic working groups

- Research Plan:
  - Uniformly analyze **2800+ tumor/normal whole genome pairs** from ICGC
  - Make this **'big 0.8PB cancer data'** available to research community

‘Big Data’ is a relative term

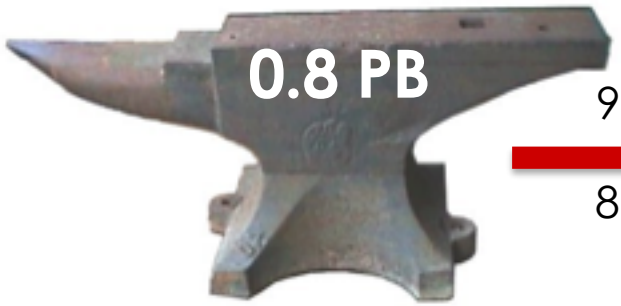
- This is what a **5MB** hard drive looked like in 1956 (note the forklift)

http://goo.gl/f1PkV

- This is what a **5TB** drive looks like in 2017 (1 million times more storage)

With the current pace of sequencing, 1.5PB (ICGC dataset) will soon become a small dataset!

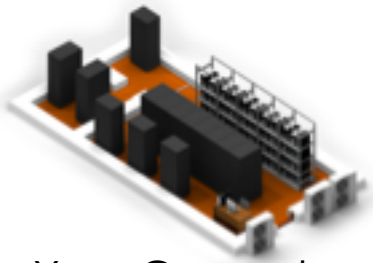
Genomic Data Distribution is a Challenge



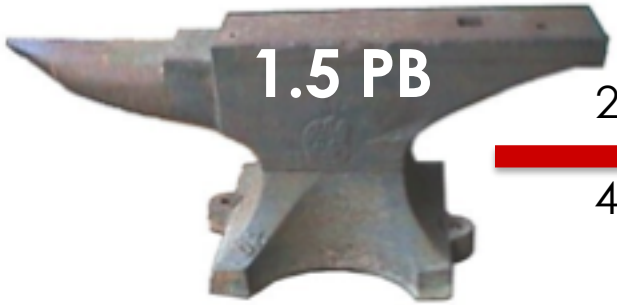
PCAWG Dataset Today
2800+ donors

9 days on a dedicated 10G link

8 mo on a shared University link



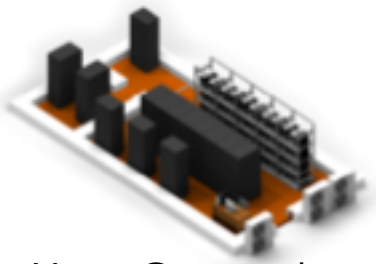
Your Compute Cluster



ICGC Dataset in 2018
29,000+ donors

2 mo on a dedicated 10G link

4 yr on a shared University link



Your Compute Cluster

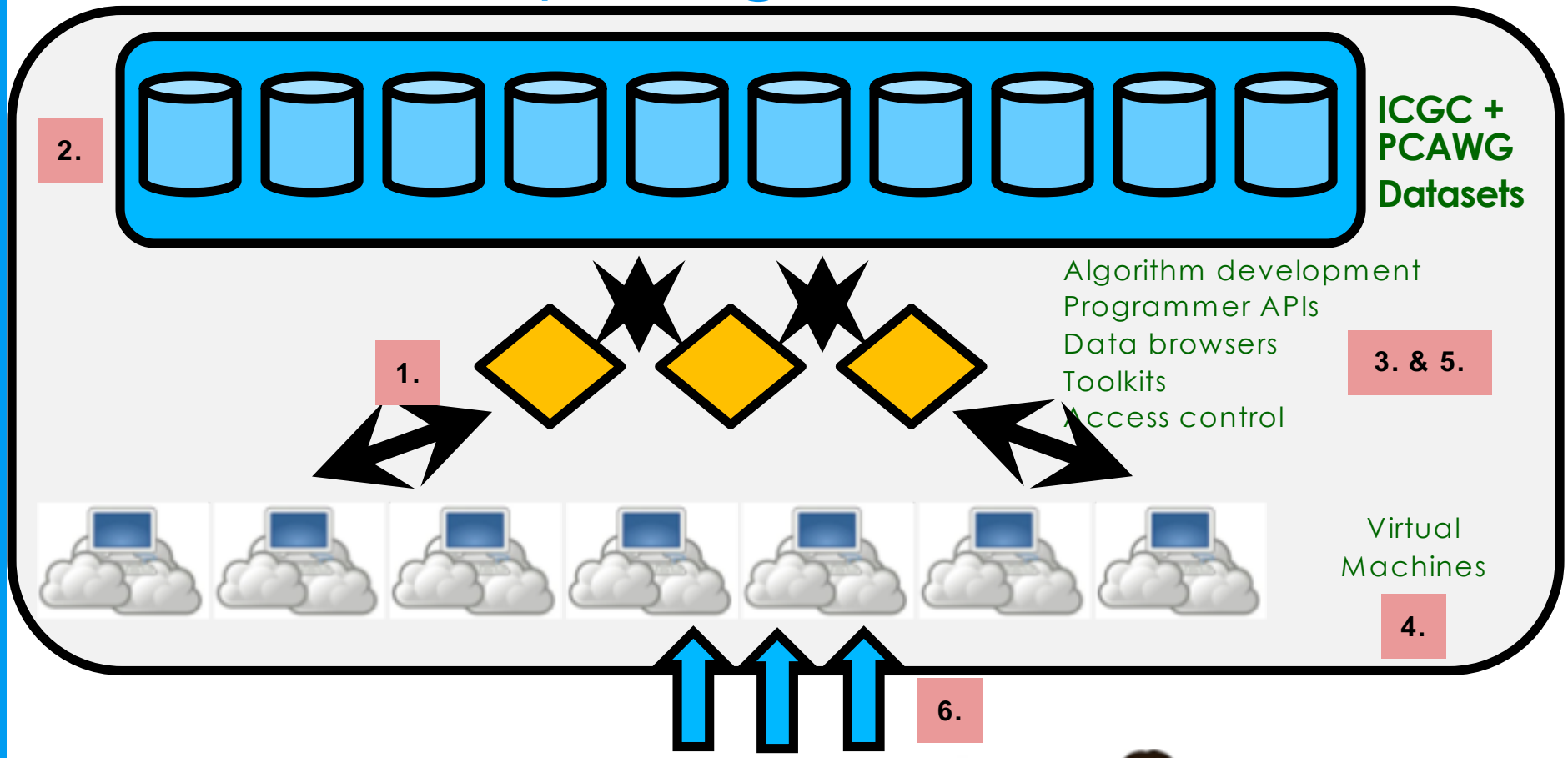
Few research labs have large enough storage and compute capacities

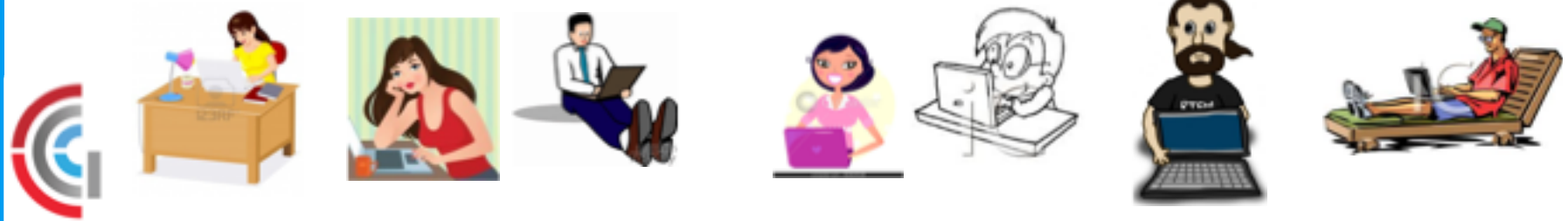
Cancer Genome COLLABORATORY

Cloud computing for collaborative research

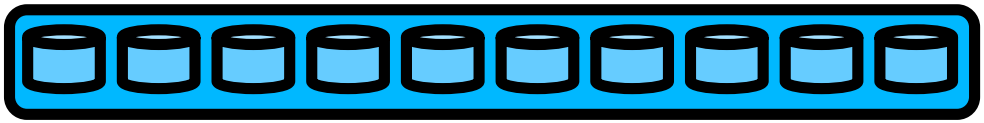
http://www.cancercollaboratory.org

Cloud Computing as a Solution





Collaboratory Infrastructure, Data & Usage



Hardware & Data

- 2,592 CPU cores, 4.4 PB storage (raw)
- 1,949 PCAWG donors released in Collaboratory + 885 PCAWG donors released in Chicago
- Sequencing data for ICGC from EGA is being copied into Collaboratory

Usage

- Collaboratory is open to user enrollment
- User account and project registration system is in place
- 50 enrolled users across 20 projects around the globe

cancercollaboratory.org

Request an Account | Getting Started | Collaboratory Console | Collaboratory Repository

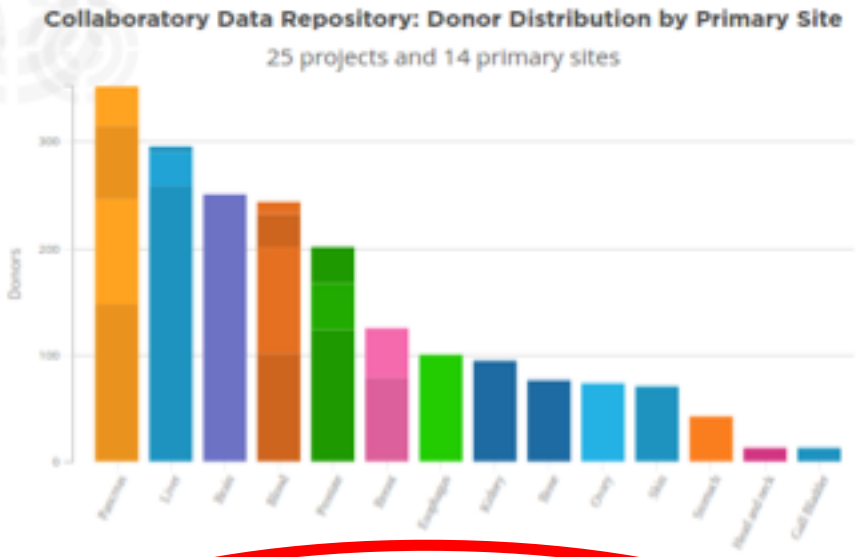
Cancer Genome COLLABORATORY
Cloud computing for collaborative research

About Us | Our Services | Our Research | Support | Contact Us | Search

Cloud Computing for BIG DATA Genomics

Welcome to the **Cancer Genome Collaboratory**, an academic compute cloud resource that allows researchers to run complex analysis operations across large ICGC cancer genome data sets.

ABOUT OUR SERVICES →

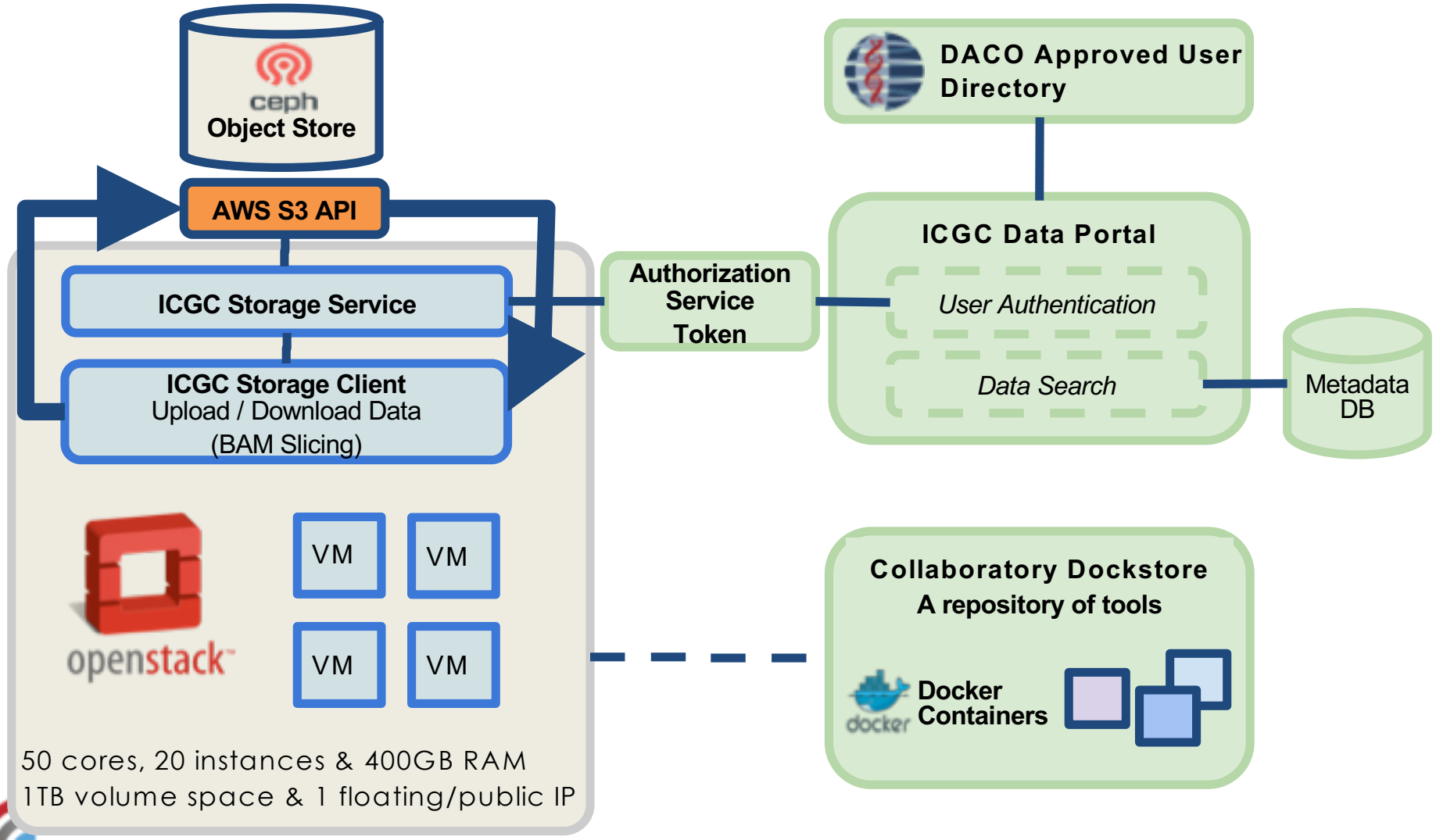


The Collaboratory data consists of:

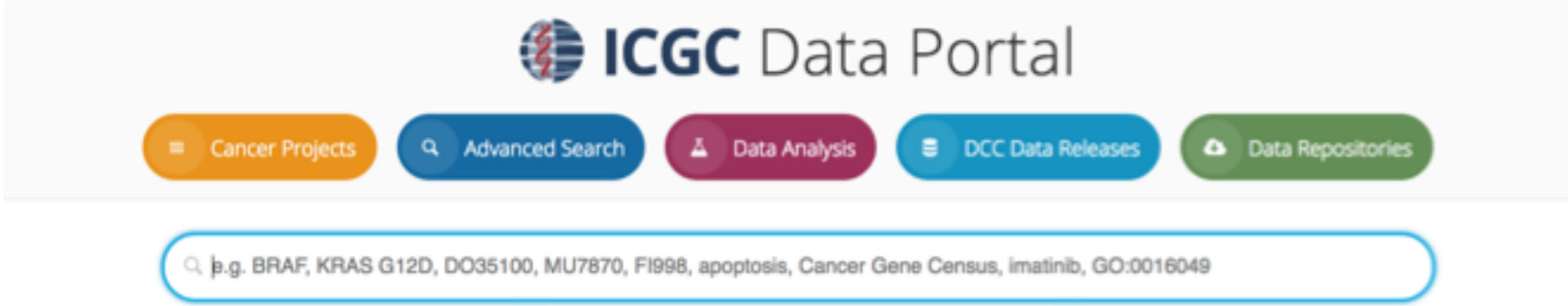
Collaboratory - Toronto	1,949 donors	48,317 files	546.8 TB
PDC - Chicago	885 donors	5,184 files	254.19 TB
Total	2,834 donors	53,501 files	801 TB

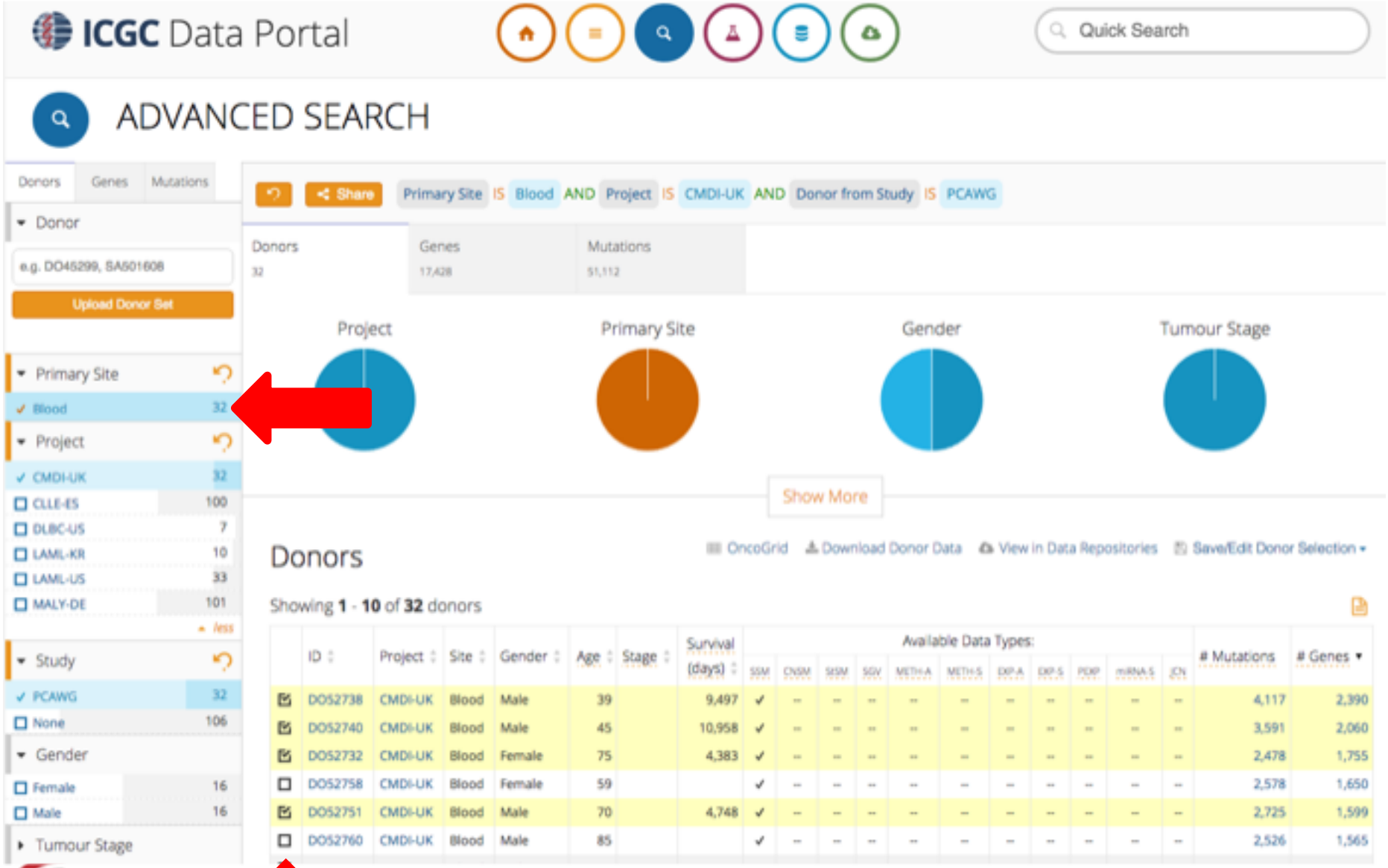
Collaboratory and PDC together hold the entire PCAWG dataset

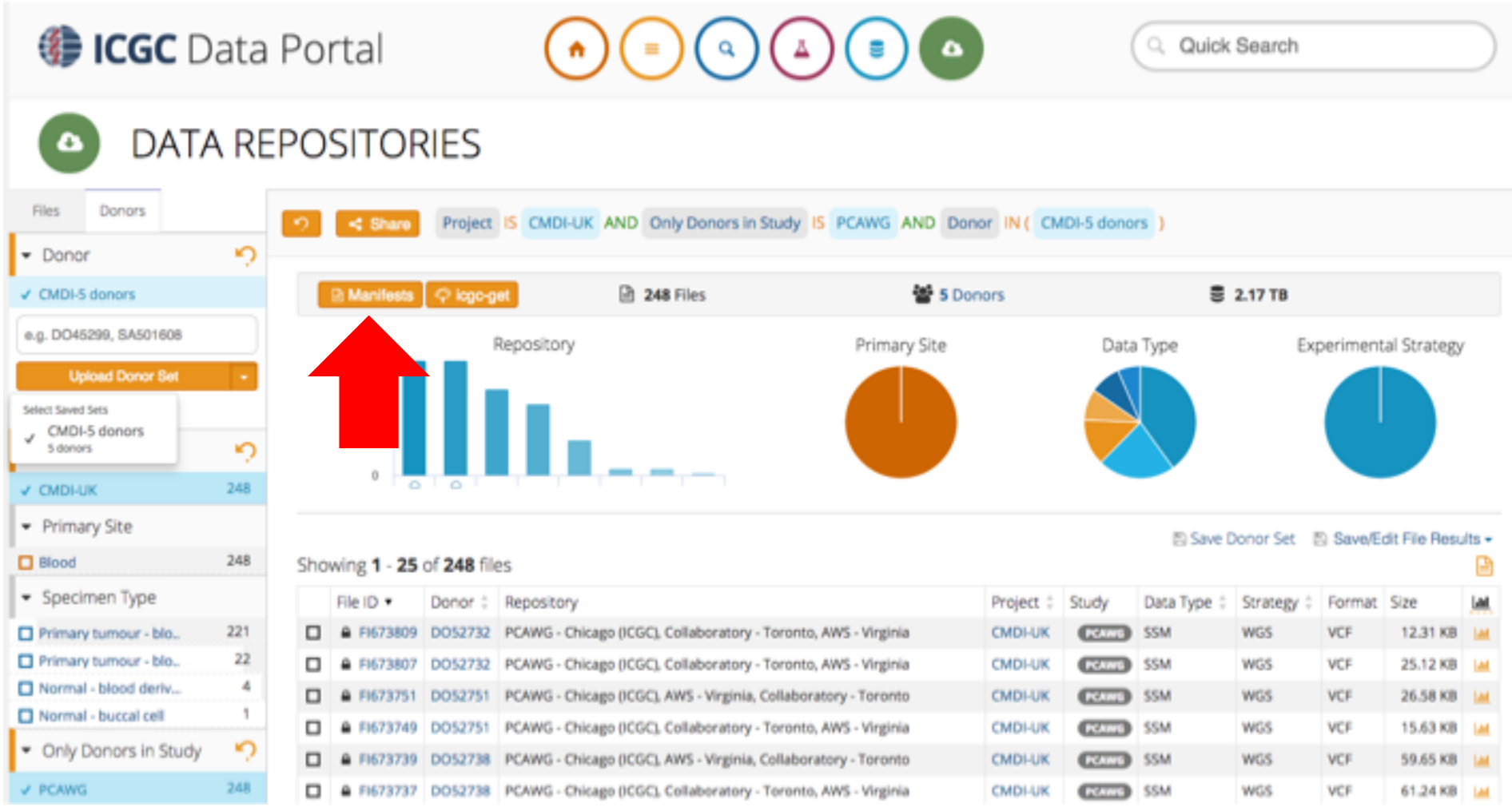
Collaboratory System Architecture



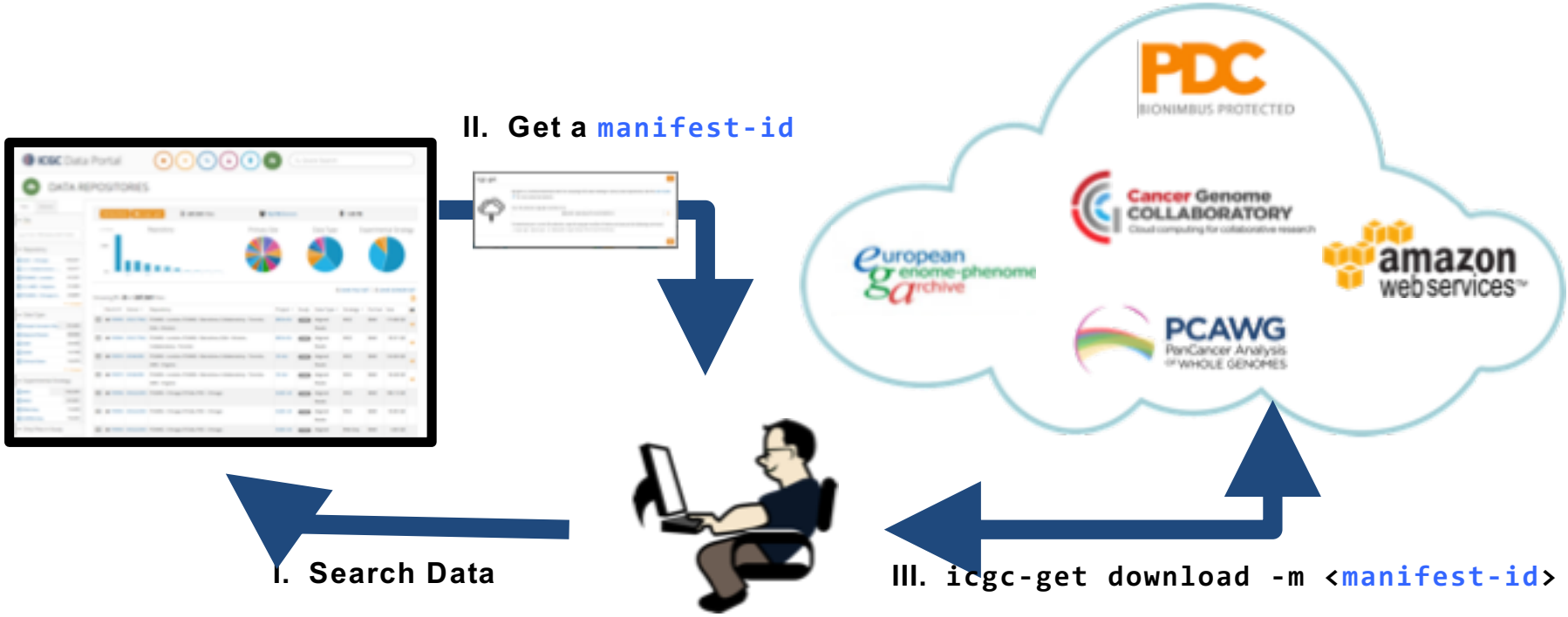
ICGC Data Portal (dcc.icgc.org)




ICGC Data Portal
Quick Search
ADVANCED SEARCH
Donors
Genes
Mutations
Donor
e.g. DO45299, SA501608
Upload Donor Set
Primary Site
Blood 32
Project
CMDI-UK 32
CLLE-ES 100
DLBC-US 7
LAML-KR 10
LAML-US 33
MALY-DE 101
less
Study
PCAWG 32
None 106
Gender
Female 16
Male 16
Tumour Stage
Share
Primary Site IS Blood AND Project IS CMDI-UK AND Donor from Study IS PCAWG
Donors 32
Genes 17,428
Mutations 51,112
Project
Primary Site
Gender
Tumour Stage
Show More
Donors
OncoGrid
Download Donor Data
View in Data Repositories
Save/Edit Donor Selection
Showing 1 - 10 of 32 donors
ID
Project
Site
Gender
Age
Stage
Survival (days)
Available Data Types:
Mutations
Genes
DO52738 CMDI-UK Blood Male 39 9,497 4,117 2,390
DO52740 CMDI-UK Blood Male 45 10,958 3,591 2,060
DO52732 CMDI-UK Blood Female 75 4,383 2,478 1,755
DO52758 CMDI-UK Blood Female 59 2,578 1,650
DO52751 CMDI-UK Blood Male 70 4,748 2,725 1,599
DO52760 CMDI-UK Blood Male 85 2,526 1,565


ICGC Data Portal
Quick Search
DATA REPOSITORIES
Files
Donors
Donor
CMDI-5 donors
e.g. DO45299, SA501608
Upload Donor Set
Select Saved Sets
CMDI-5 donors
5 donors
CMDI-UK 248
Primary Site
Blood 248
Specimen Type
Primary tumour - blo.. 221
Primary tumour - blo.. 22
Normal - blood deriv... 4
Normal - buccal cell 1
Only Donors in Study
PCAWG 248
Share
Project IS CMDI-UK AND Only Donors in Study IS PCAWG AND Donor IN (CMDI-5 donors)
Manifests
icgc-get
248 Files
5 Donors
2.17 TB
Repository
Primary Site
Data Type
Experimental Strategy
Save Donor Set
Save/Edit File Results
Showing 1 - 25 of 248 files
File ID
Donor
Repository
Project
Study
Data Type
Strategy
Format
Size
FI673809 DO52732 PCAWG - Chicago (ICGC), Collaboratory - Toronto, AWS - Virginia CMDI-UK PCAWG SSM WGS VCF 12.31 KB
FI673807 DO52732 PCAWG - Chicago (ICGC), Collaboratory - Toronto, AWS - Virginia CMDI-UK PCAWG SSM WGS VCF 25.12 KB
FI673751 DO52751 PCAWG - Chicago (ICGC), AWS - Virginia, Collaboratory - Toronto CMDI-UK PCAWG SSM WGS VCF 26.58 KB
FI673749 DO52751 PCAWG - Chicago (ICGC), Collaboratory - Toronto, AWS - Virginia CMDI-UK PCAWG SSM WGS VCF 15.63 KB
FI673739 DO52738 PCAWG - Chicago (ICGC), AWS - Virginia, Collaboratory - Toronto CMDI-UK PCAWG SSM WGS VCF 59.65 KB
FI673737 DO52738 PCAWG - Chicago (ICGC), Collaboratory - Toronto, AWS - Virginia CMDI-UK PCAWG SSM WGS VCF 61.24 KB

icgc-get:
A Universal Download Client for ICGC Data

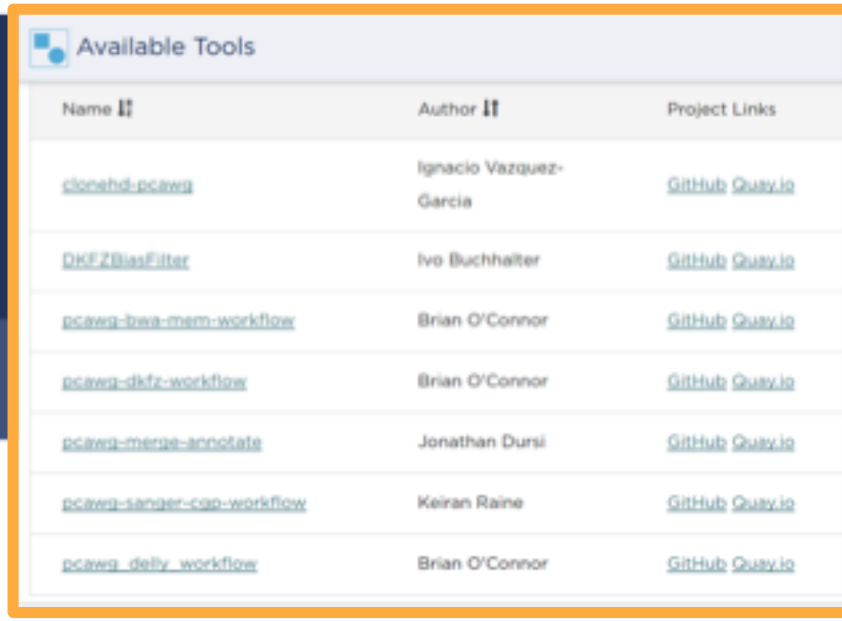


How it works:

- Uses DCC Portal to define / serve your manifest (manifest-id)
- Checks you are authorised to access the requested data
- Downloads files from repositories in a chosen order of preference (e.g. eliminates duplicates)

Dockstore.org for Analysis Workflows

- Open platform for sharing Docker-based tools using Common Workflow Language used by GA4GH
  - Website and command line tools for registering containers
  - Builds upon Docker containers (docker.com)
- "Dockerized" all workflows (alignment and variant calling pipelines) used by PCAWG working groups



Cloud Computing Training

Bioinformatics.ca

- Provide training on use of the Cancer Genome Collaboratory cloud

Topics covered:

- Cloud computing
- Principles of virtual machine management
- Creation of portable software packages using Docker containers
- Use of Dockstore to co-package binaries & computational workflow descriptions

Benefits of Collaboratory

The Cancer Genome Collaboratory offers cancer researchers:

- Compute capacity
- Individual project tenancy
- Protected cancer sequencing data sets
- Tools to facilitate easy data access
- Tools for exporting and sharing analysis workflows
- Easy to understand pricing model (linear for # cores used)
- Responsive helpdesk support team

cancercollaboratory.org

Looking Forward in Collaboratory

- Collect more cancer datasets from other repositories
- Automate the enrollment process for DACO-approved researchers
- Train users in securely using cloud computing for analysis of protected data
- Increase the capacity of the distributed storage infrastructure, while maintaining performance & stability
- Develop APIs and tools for analysis of streaming data, rather than downloaded data

Co-Investigators of Cores & Research Modules

Robert Grossman, PhD
Core 1 - Infrastructure
University of Chicago

+

Vincent Ferretti, PhD
Core 1 - Infrastructure
OICR

Paul Boutros, PhD
Core 2 - Benchmarking
OICR & Univ. Toronto

Francis Ouellette
Core 3 - Training
OICR & Univ. Toronto

Lincoln Stein, MD, PhD
Lead & Core 4 - Administration
OICR & Univ. Toronto

Michelle Brazas
Project Manager
OICR

Cenk Sahinalp, PhD
Res. Module 1 - Search, Indexing & Compression
Simon Fraser Univ.

Guillaume Bourque, PhD
Res. Module 2 - Variant Consequence Prediction
McGill Univ.

Sohrab Shah, PhD
Res. Module 3 - Tumour Heterogeneity
Univ. British Columbia

Lincoln Stein, MD, PhD
Res. Module 4 - Drug Target Identification
OICR & Univ. Toronto

Khaled El Emam, PhD
Res. Module 5 - Bioethics
McGill Univ.

Bartha Knoppers, PhD
Res. Module 5 - Bioethics
McGill Univ.

Acknowledgements

Annai Systems
Michael Ainsworth
Scott McIntee
Thomas Schlumpberger
Francisco De La Vega

Broad Institute
Kristian Cibulskis
Gad Getz
Julian Hess
Ignaty Leshchiner
Dimitri Livitz
Esther Reinbay
Mara Roseberg
Gordon Saksena
Chip Stewart
Grace Tiao
Jeremiah Wala

BSC
Romina Royo
Javier Bartolome
Josep Gelpi
David Torrents

DKFZ
Ivo Buchhalter
Michael Heinold
Kortine Kleinheinz
Rolf Kabbe
Matthias Schlesner

EBI
Rich Boyce
Alvis Brazma
Andy Cafferkey
Jordi Rambla De Argila
Nuno Fonseca
Paul Flicek

EMBL
Jan Korbel
Sergei Iakhnin
Ivica Letunic
Joachim Weischenfeldt

ETRI
Youngchon Hyung-Lae
Jonghui Hong
Sung-Soo Keunchil
Kang-ho Kim
Hyunghwan Kim
Jongsun Wan
Jeon Seung-Hyub
Youngwook Kim

iDASH
Ashley Williams
Tony Chen
Jihoon Kim
Olivier Harismendy

OHSU
Kyle Ellrott

IMSUT+RIKEN
Keith Boroevich
Akihiro Fujimoto
Satoru Miyano
Naoki Miyoshi
Hidewaki Nakagawa
Kazuhiro Ohi

OICR
Michelle Brazas
Niall Byrne
Jonathan Dursi
Nodirjon Fayzullaev
Vincent Ferretti
Wei Jiao
Jerry Lam
Sheldon McKay
George Mihaiescu
Jared Baker
Francois Gerthoffert
Robert Tisma
Hardeep Nahal
Francis Ouellette
Marc Perry
Solomon Shorser
Jared Simpson
Lincoln Stein
Bob Tiernay
Adam Wright
Linda Xiang
Andy Yang
Denis Yuen
Christina Yung
Junjun Zhang

ONTARIO INSTITUTE FOR CANCER RESEARCH

NIH
Carolyn Hutter
Todd Pihl
Heidi Sophia

SAGE Bionetworks
Larsson Omberg

Sanger
Keiran Raine
Adam Butler
Yilong Li
Peter Campbell

Seven Bridges Genomics
Brandi Davis Dusenbery
Petar Radovic
Milena Kovacevic
Nebojsa Tijanic
Deniz Kural

UCSC
David Haussler
Brian O'Connor
Benedict Paten
Linda Rosewood
Josh Stuart
Chris Wilks

University of Chicago
Allison Heath
Jonathan Spring
Michael Ford
Robert Grossman

Funding for the Ontario Institute for Cancer Research is provided by the Government of Ontario

OICR
Ontario Institute for Cancer Research

Ontario

Funding for the Cancer Genome Collaboratory is provided by

GenomeCanada

NSERC
CRSNG

OICR
Ontario Institute for Cancer Research

Ontario

INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

CIHR IRSC
Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

Thank you

Questions?