Overview of the NCI Cancer Research Data Commons, Genomic Data Commons (GDC), and NCI Cloud Resources

Tanja Davidsen, PhD

NCI Center for Biomedical Informatics and IT

NIH NATIONAL CANCER INSTITUTE

The NCI Cancer Research Data Commons
June, 2018

The Cancer Data Ecosystem

The Beau Biden Cancer Moonshot^{sm}

Overarching goals – Jan, 2016

- Accelerate progress in cancer, including prevention & screening
  - From cutting edge basic research to wider uptake of standard of care
- Encourage greater cooperation and collaboration
  - Within and between academia, government, and private sector
- Enhance data sharing

Blue Ribbon Panel – October, 2016

- Network for Direct Patient Engagement
- Cancer Immunotherapy Translational Science Network
- Therapeutic Target Identification to Overcome Drug Resistance
- A National Cancer Data Ecosystem for Sharing and Analysis
- Fusion Oncoproteins in Childhood Cancers
- Symptom Management Research
- Prevention and Early Detection – Implementation of Evidence-based Approaches
- Retrospective Analysis of Biospecimens from Patients Treated with Standard of Care
- Generation of 3D Human Tumor Atlas
- Development of New Enabling Cancer Technologies
- Full report: www.cancer.gov/brp

National Cancer Data Ecosystem Recommendations

Overall goal: *"Enable all participants across the cancer research and care continuum to contribute, access, combine and analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer."*

Recommendations

- Build a National Cancer Data Ecosystem
  - Enhanced cloud-computing platforms.
  - Services that link disparate information, including clinical, image, and molecular data.
  - Essential underlying data science infrastructure, methods, and portals for the Cancer Data Ecosystem.
  - Establish sustainable data governance to ensure long-term health of the Ecosystem.
  - Develop standards and tools so that data are interoperable.

Enhanced Data Sharing Working Group Recommendation:
The Cancer Data Ecosystem

SBG CGC

Broad FireCloud

ISB CGC

Cancer Research Data Commons

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

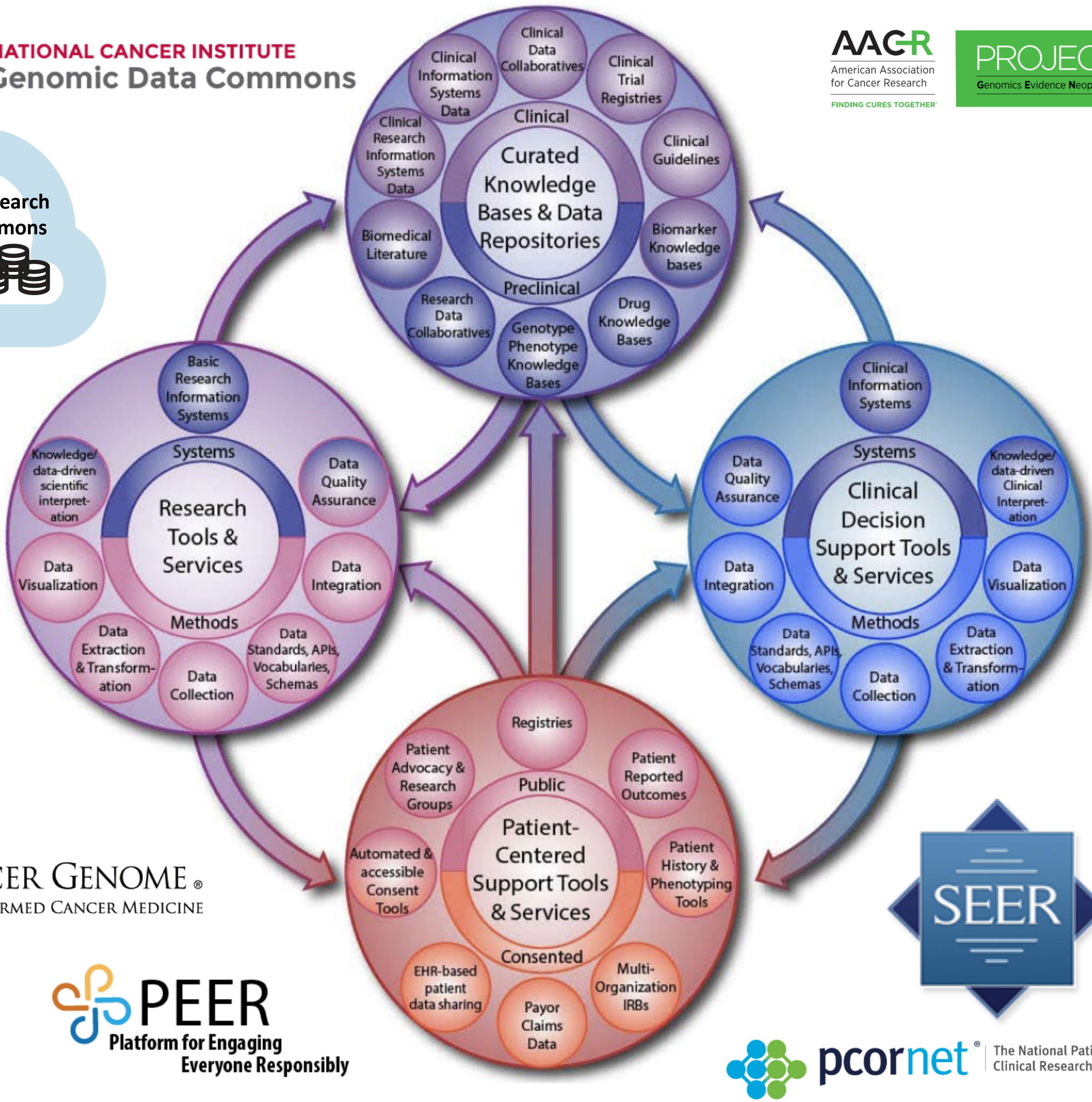
NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

monarch INITIATIVE

MY CANCER GENOME® GENETICALLY INFORMED CANCER MEDICINE

Sage BIONETWORKS

PEER Platform for Engaging Everyone Responsibly



Curated Knowledge Bases & Data Repositories
Clinical
Preclinical
Clinical Data Collaboratives
Clinical Trial Registries
Clinical Guidelines
Biomarker Knowledge bases
Drug Knowledge Bases
Genotype Phenotype Knowledge Bases
Research Data Collaboratives
Biomedical Literature
Clinical Research Information Systems Data
Clinical Information Systems Data
Research Tools & Services
Systems
Methods
Basic Research Information Systems
Data Quality Assurance
Data Integration
Data Standards, APIs, Vocabularies, Schemas
Data Collection
Data Extraction & Transform-ation
Data Visualization
Knowledge/ data-driven scientific interpret-ation
Clinical Decision Support Tools & Services
Systems
Methods
Clinical Information Systems
Knowledge/ data-driven Clinical Interpret-ation
Data Visualization
Data Extraction & Transform-ation
Data Collection
Data Standards, APIs, Vocabularies, Schemas
Data Integration
Data Quality Assurance
Patient-Centered Support Tools & Services
Public
Consented
Registries
Patient Reported Outcomes
Patient History & Phenotyping Tools
Multi-Organization IRBs
Payor Claims Data
EHR-based patient data sharing
Automated & accessible Consent Tools
Patient Advocacy & Research Groups

AACR American Association for Cancer Research FINDING CURES TOGETHER

PROJECTGENIE Genomics Evidence Neoplasia Information Exchange

ORIEN ONCOLOGY RESEARCH INFORMATION EXCHANGE NETWORK

ASCO CANCERLINQ™ Learning Intelligence Network for Quality

SEER

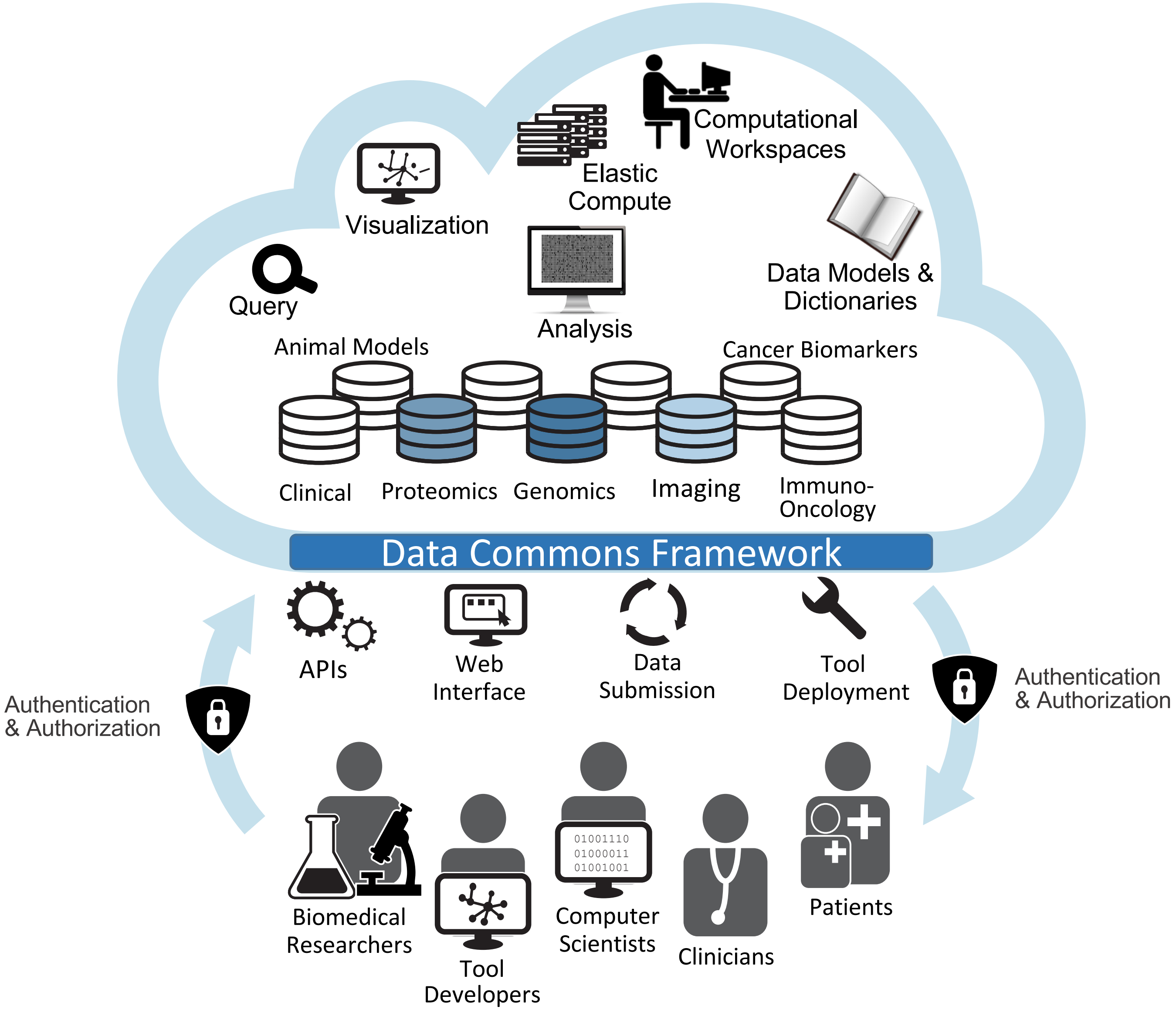
pcornet® The National Patient-Centered Clinical Research Network

The Cancer Research Data Commons (CRDC)

NCI Cancer Research Data Commons (CRDC) - Concept

NCI Scope: *“Create a data science infrastructure necessary to connect repositories, analytical tools, and knowledge bases”*

Data commons co-locate data, storage and computing infrastructure with commonly used services, tools & apps for analyzing and sharing data to create an interoperable resource for the research community.*



Data Contributors and Consumers

*Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016. Source of image: The CDIS, GDC, & OCC data commons infrastructure at the University of Chicago Kenwood Data Center.

Goals of the NCI CRDC

- Enable the cancer research community to share diverse data types across programs and institutions.

- Provide easy access to data, regardless of where they are stored.

- Provide mechanisms for innovative tool discovery, access, and usage, e.g., ITCR tools.

- Help Data Coordinating Centers share their data publicly.

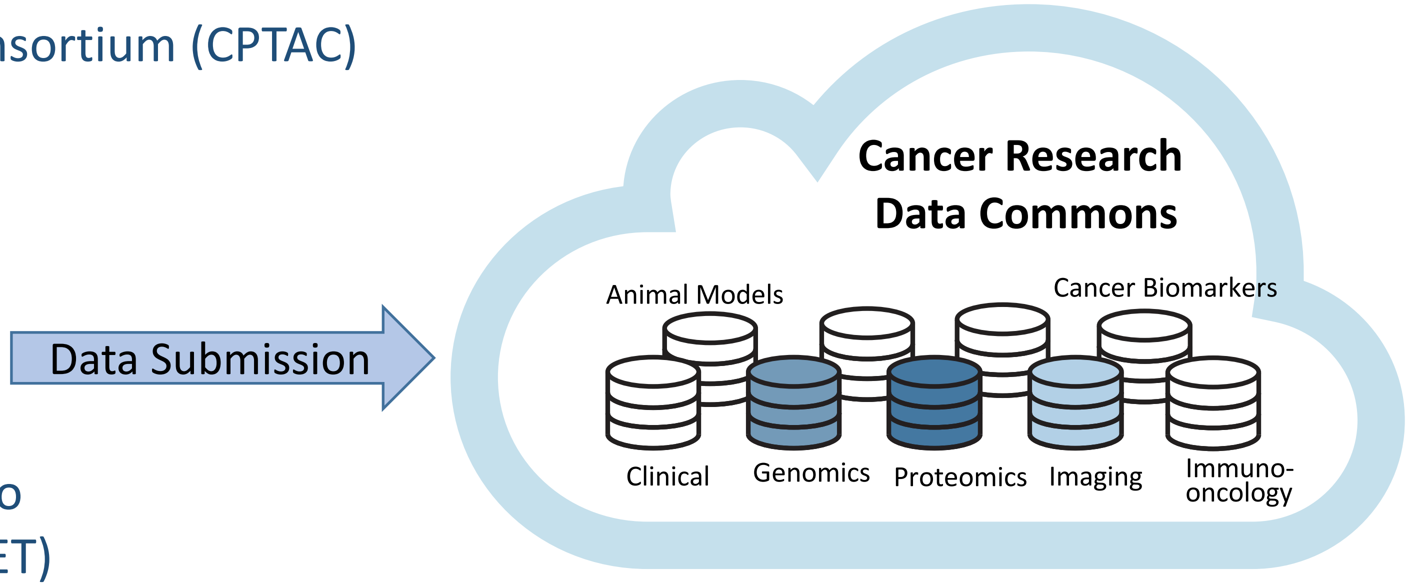
NCI Cancer Research Data Commons
Data Sources / Contributors (Examples)

The Cancer Genome Atlas (TCGA)

Clinical Proteomic Tumor Analysis Consortium (CPTAC)

The Cancer Imaging Archive (TCIA)

NCI Individual Labs / Grants / Contracts / Cancer Centers (GENIE)



Data Submission

Cancer Research Data Commons

Animal Models

Cancer Biomarkers

Clinical

Genomics

Proteomics

Imaging

Immuno-oncology

Therapeutically Applicable Research to Generate Effective Treatments (TARGET)

APOLLO
ICPC

Collaborative Programs: APOLLO (Applied Proteogenomic OrganizationaL Learning and Outcomes), ICPC (International Cancer Proteogenome Consortium)

3rd Party Programs: Foundation Medicine, Multiple Myeloma Research Foundation

Data Commons Framework – What Is It?

Reusable, expandable framework for a Data Commons

Core principles and structures for a Data Commons

Set of modular components that can be leveraged across Data Commons

Modular Components

- Secure user authentication and authorization
- Metadata validation and tools
- Domain-specific, extensible data models and dictionaries
- API and container environment for tools and pipelines
- Access to computational workspaces for storing data, tools, and results

NCI Cancer Research Data Commons

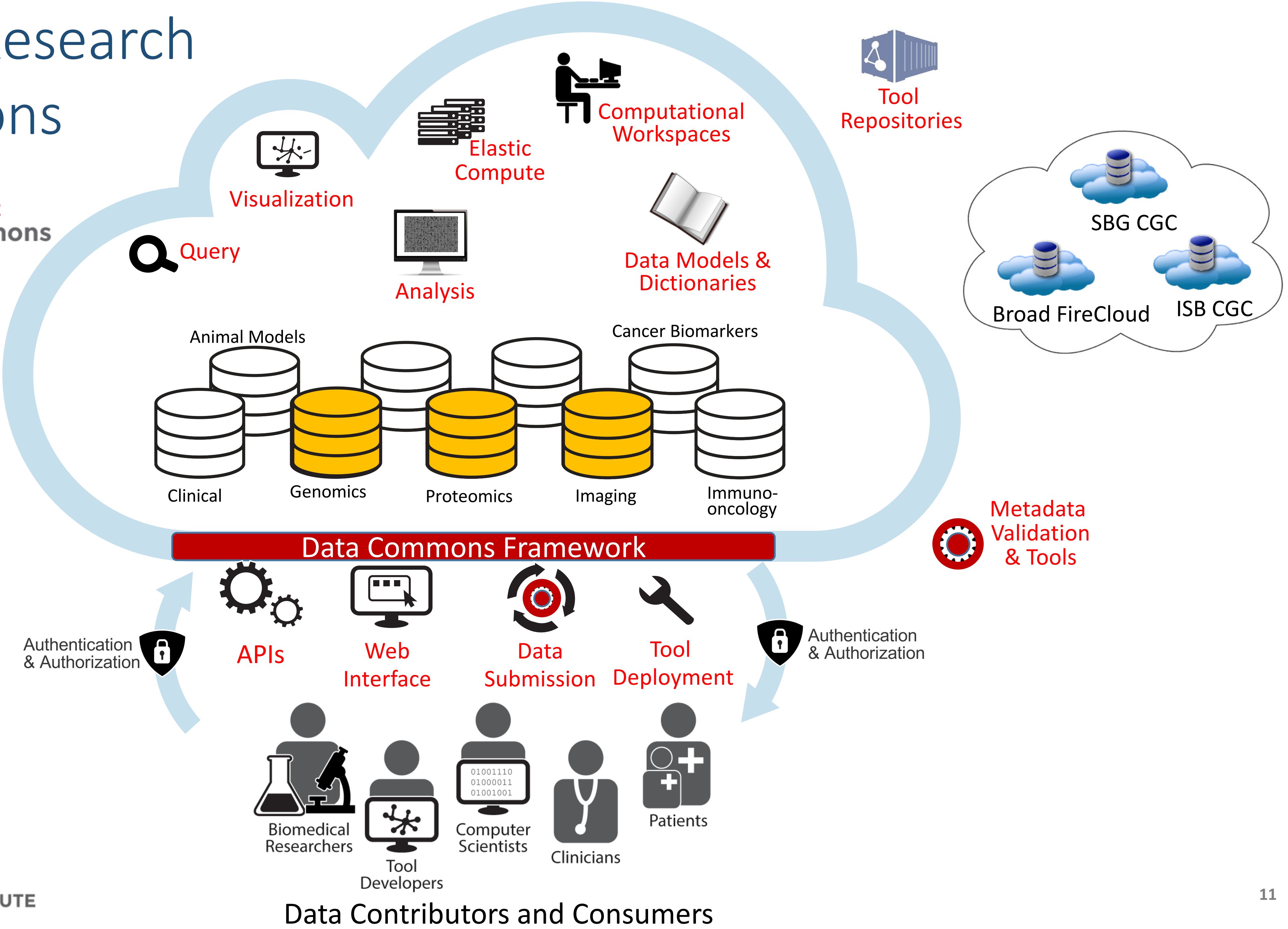
NATIONAL CANCER INSTITUTE
Genomic Data Commons

CPTAC

*Clinical Proteomics Tumor Analysis Consortium**

TCIA

*The Cancer Imaging Archive**



Visualization

Elastic Compute

Computational Workspaces

Tool Repositories

Query

Analysis

Data Models & Dictionaries

SBG CGC

Broad FireCloud

ISB CGC

Animal Models

Cancer Biomarkers

Clinical

Genomics

Proteomics

Imaging

Immuno-oncology

Data Commons Framework

Metadata Validation & Tools

Authentication & Authorization

APIs

Web Interface

Data Submission

Tool Deployment

Authentication & Authorization

Biomedical Researchers

Tool Developers

Computer Scientists

Clinicians

Patients

Data Contributors and Consumers

The NCI Genomic Data Commons

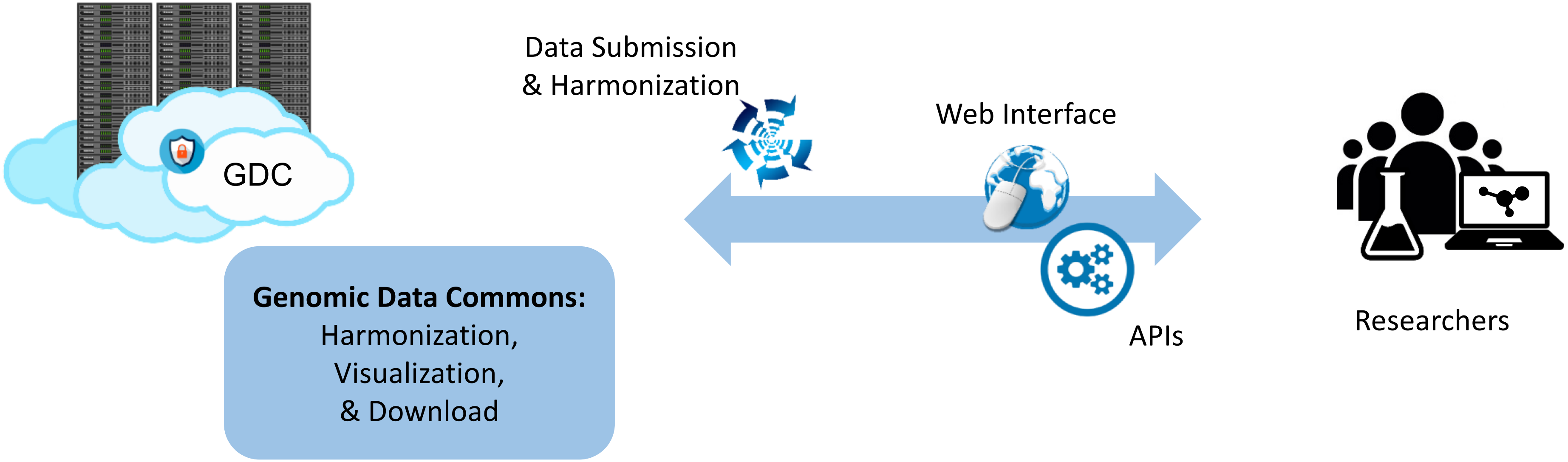
Provide the cancer research community with a ***unified data repository*** *that enables* ***data sharing*** *across* ***cancer genomic studies*** *in support of* ***precision medicine***

The NCI Genomic Data Commons

- Support the *receipt*, *quality control*, *integration*, *storage*, and *redistribution* of standardized genomic data sets derived from cancer research studies
  - Available data
    - NCI Funded cancer genomics datasets
    - User submissions
  - Data searching and retrieval/downloading
  - Harmonization of raw sequence (alignment and variant calling) of all GDC data
  - Application of state-of-the-art methods of generating derived data
- Developed, supported, and hosted by U. Chicago

Genomic Data Commons (GDC):

A unified data repository for the research community developed, supported, and hosted by U. Chicago



GDC

Genomic Data Commons:
Harmonization,
Visualization,
& Download

Data Submission
& Harmonization

Web Interface

APIs

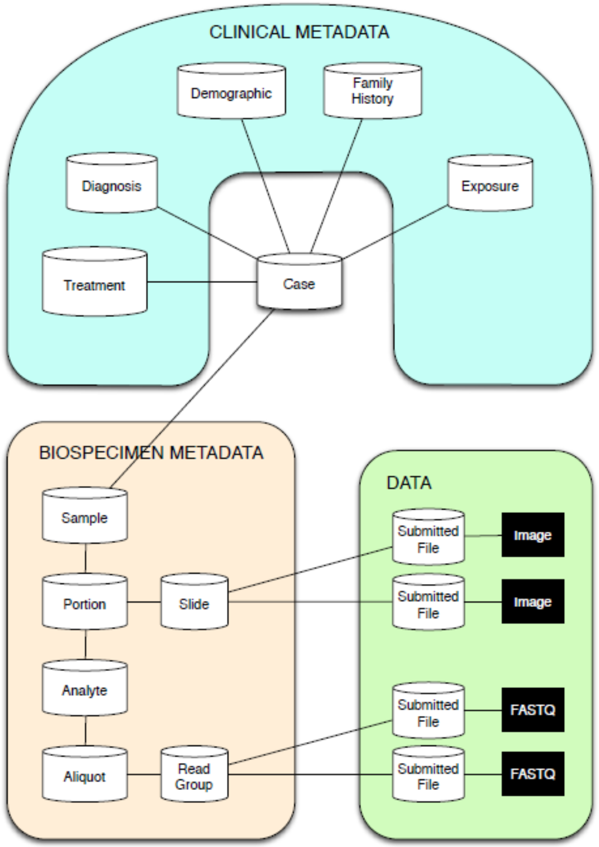
Researchers

https://gdc.cancer.gov

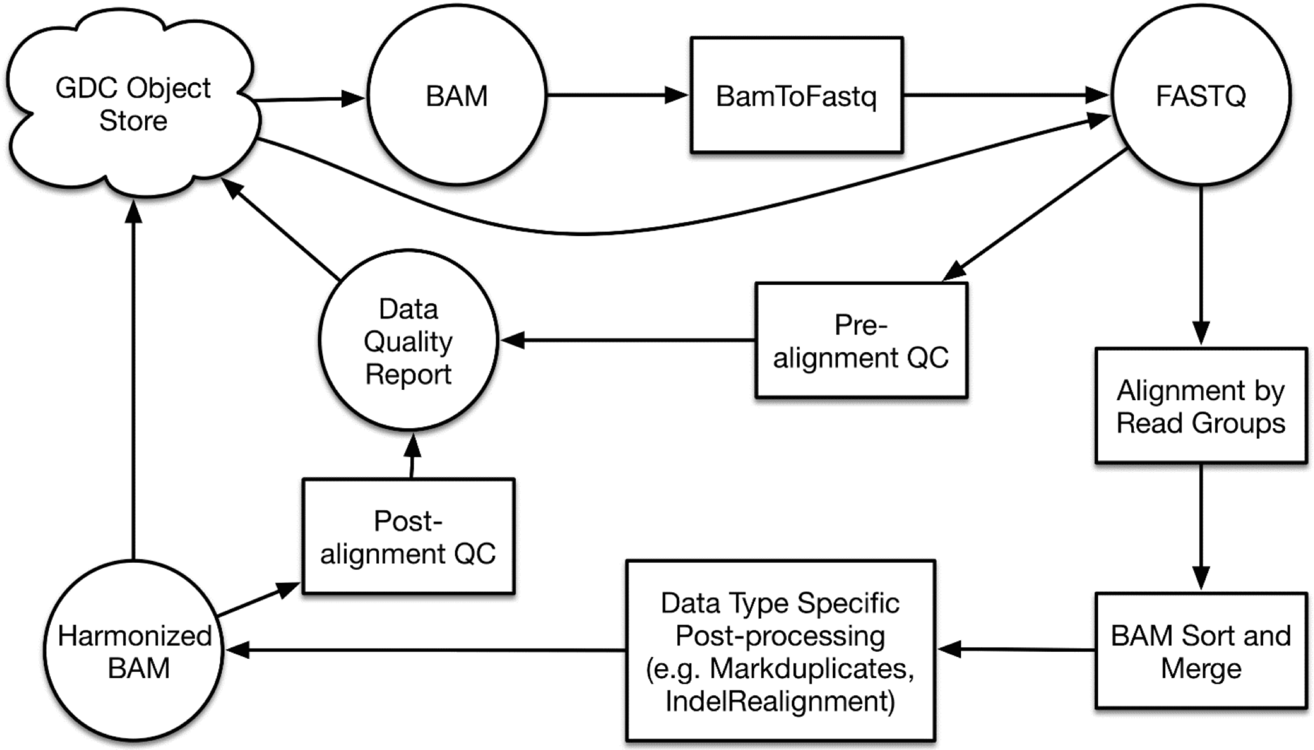
Authentication
& Authorization thru
eRA Commons & dbGaP

GDC: Data Submission & Harmonization

Data Submission



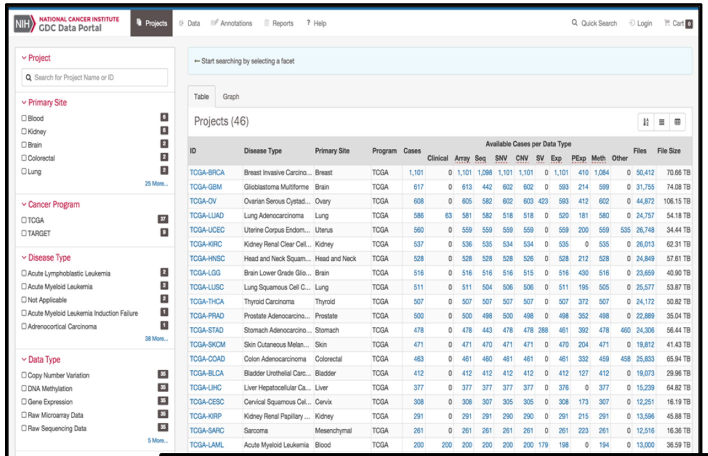
Data Harmonization



https://gdc.cancer.gov/

GDC: Data Retrieval

Data Portal



← Start searching by selecting a facet

Project

Search for Project Name or ID

Primary Site

- ☐ Blood 6
- ☐ Kidney 6
- ☐ Brain 2
- ☐ Colorectal 2
- ☐ Lung 2

25 More...

Cancer Program

- ☐ TCGA 37
- ☐ TARGET 9

Disease Type

- ☐ Acute Lymphoblastic Leukemia 2
- ☐ Acute Myeloid Leukemia 2
- ☐ Not Applicable 2
- ☐ Acute Myeloid Leukemia Induction Failure 1
- ☐ Adrenocortical Carcinoma 1

38 More...

Data Type

- ☐ Copy Number Variation 35
- ☐ DNA Methylation 35
- ☐ Gene Expression 35
- ☐ Raw Microarray Data 35
- ☐ Raw Sequencing Data 35

5 More...

Experimental Strategy

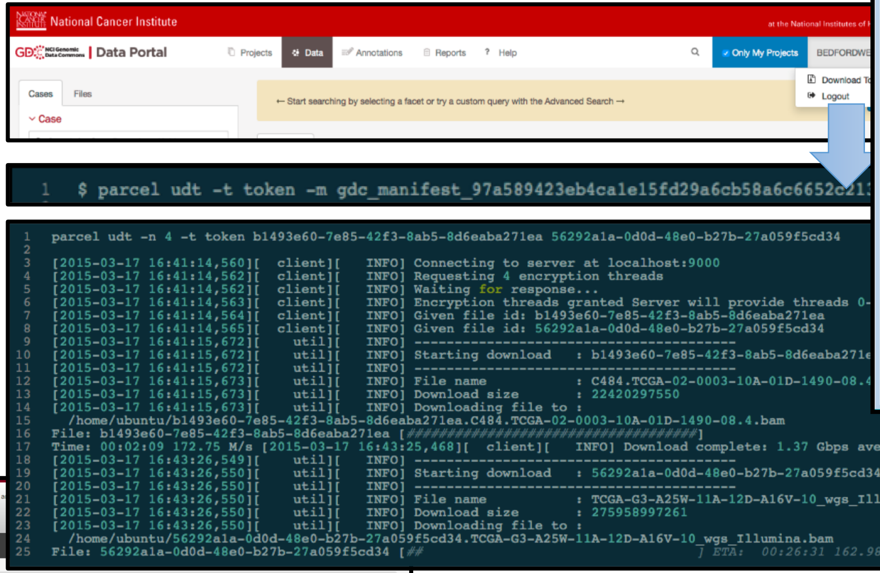
- ☐ Genotyping Array

Table | Graph

Projects (46)

ID	Disease Type	Primary Site	Program	Cases	Available Cases per Data Type										Files	File Size
					Clinical	Array	Seq	SNV	CNV	SV	Exp	PExp	Meth	Other		
TCGA-BRCA	Breast Invasive Carcino...	Breast	TCGA	1,101	0	1,101	1,098	1,101	1,101	0	1,101	410	1,084	0	50,412	70.66 TB
TCGA-GBM	Glioblastoma Multiforme	Brain	TCGA	617	0	613	442	602	602	0	593	214	599	0	31,755	74.08 TB
TCGA-OV	Ovarian Serous Cystad...	Ovary	TCGA	608	0	605	582	602	603	423	593	412	602	0	44,872	106.15 TB
TCGA-LUAD	Lung Adenocarcinoma	Lung	TCGA	586	63	581	582	518	518	0	520	181	580	0	24,757	54.18 TB
TCGA-UCEC	Uterine Corpus Endom...	Uterus	TCGA	560	0	559	559	559	559	0	559	200	559	535	26,748	34.44 TB
TCGA-KIRC	Kidney Renal Clear Cell...	Kidney	TCGA	537	0	536	535	534	534	0	535	0	535	0	26,013	62.31 TB
TCGA-HNSC	Head and Neck Squam...	Head and Neck	TCGA	528	0	528	528	528	526	0	528	212	528	0	24,849	57.61 TB
TCGA-LGG	Brain Lower Grade Glio...	Brain	TCGA	516	0	516	516	516	515	0	516	430	516	0	23,659	40.90 TB
TCGA-LUSC	Lung Squamous Cell C...	Lung	TCGA	511	0	511	504	506	506	0	511	195	505	0	25,577	53.87 TB
TCGA-THCA	Thyroid Carcinoma	Thyroid	TCGA	507	0	507	507	507	507	0	507	372	507	0	24,172	50.82 TB
TCGA-PRAD	Prostate Adenocarcino...	Prostate	TCGA	500	0	500	498	500	498	0	498	352	498	0	22,889	35.04 TB
TCGA-STAD	Stomach Adenocarcino...	Stomach	TCGA	478	0	478	443	478	478	288	461	392	478	460	24,306	56.44 TB
TCGA-SKCM	Skin Cutaneous Melan...	Skin	TCGA	471	0	471	470	471	471	0	470	204	471	0	19,812	41.43 TB
TCGA-COAD	Colon Adenocarcinoma	Colorectal	TCGA	463	0	461	460	461	461	0	461	332	459	458	25,833	65.94 TB
TCGA-BLCA	Bladder Urothelial Carc...	Bladder	TCGA	412	0	412	412	412	412	0	412	127	412	0	19,073	29.96 TB
TCGA-LIHC	Liver Hepatocellular Ca...	Liver	TCGA	377	0	377	377	377	377	0	376	0	377	0	15,239	64.82 TB
TCGA-CESC	Cervical Squamous Cel...	Cervix	TCGA	308	0	308	307	305	305	0	308	173	307	0	12,251	16.19 TB
TCGA-KIRP	Kidney Renal Papillary ...	Kidney	TCGA	291	0	291	291	290	290	0	291	215	291	0	13,596	45.88 TB
TCGA-SARC	Sarcoma	Mesenchymal	TCGA	261	0	261	261	261	261	0	261	223	261	0	12,516	16.36 TB
TCGA-LAML	Acute Myeloid Leukemia	Blood	TCGA	200	200	200	200	200	200	179	198	0	194	0	13,000	36.59 TB

Data Transfer Tool


National Cancer Institute
at the National Institutes of H
GDC NCI Genomic Data Commons | Data Portal
Projects
Data
Annotations
Reports
Help
Only My Projects
BEDFORDWE
Download To
Logout
Cases
Files
Case

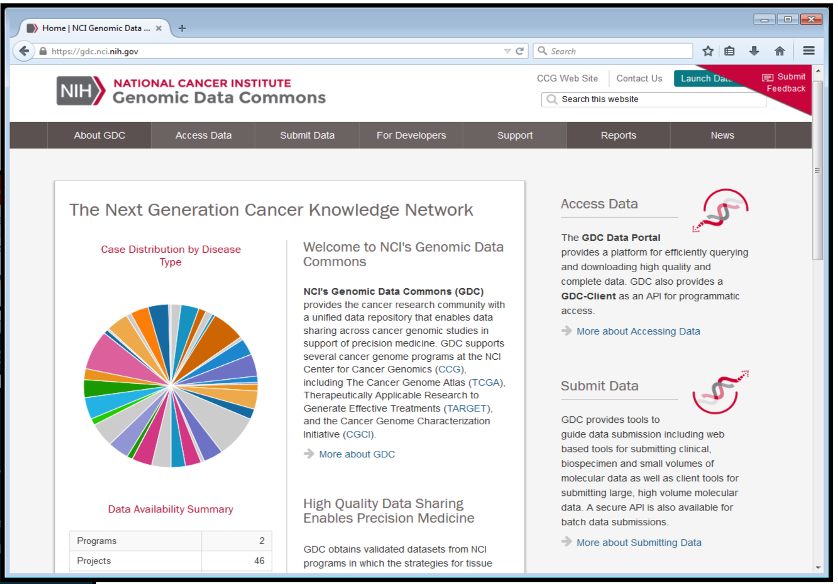
← Start searching by selecting a facet or try a custom query with the Advanced Search →

```
$ parcel udt -t token -m gdc_manifest_97a589423eb4ca1e15fd29a6cb58a6c6652c21
```

```
parcel udt -n 4 -t token b1493e60-7e85-42f3-8ab5-8d6eaba271ea 56292a1a-0d0d-48e0-b27b-27a059f5cd34

[2015-03-17 16:41:14,560][  client][   INFO] Connecting to server at localhost:9000
[2015-03-17 16:41:14,562][  client][   INFO] Requesting 4 encryption threads
[2015-03-17 16:41:14,562][  client][   INFO] Waiting for response...
[2015-03-17 16:41:14,563][  client][   INFO] Encryption threads granted Server will provide threads 0-
[2015-03-17 16:41:14,564][  client][   INFO] Given file id: b1493e60-7e85-42f3-8ab5-8d6eaba271ea
[2015-03-17 16:41:14,565][  client][   INFO] Given file id: 56292a1a-0d0d-48e0-b27b-27a059f5cd34
[2015-03-17 16:41:15,672][    util][   INFO] ------------------------------------
[2015-03-17 16:41:15,672][    util][   INFO] Starting download    : b1493e60-7e85-42f3-8ab5-8d6eaba271e
[2015-03-17 16:41:15,672][    util][   INFO] ------------------------------------
[2015-03-17 16:41:15,673][    util][   INFO] File name            : C484.TCGA-02-0003-10A-01D-1490-08.4
[2015-03-17 16:41:15,673][    util][   INFO] Download size        : 22420297550
[2015-03-17 16:41:15,673][    util][   INFO] Downloading file to  :
   /home/ubuntu/b1493e60-7e85-42f3-8ab5-8d6eaba271ea.C484.TCGA-02-0003-10A-01D-1490-08.4.bam
File: b1493e60-7e85-42f3-8ab5-8d6eaba271ea [####################################]
Time: 00:02:09 172.75 M/s [2015-03-17 16:43:25,468][  client][   INFO] Download complete: 1.37 Gbps average
[2015-03-17 16:43:26,549][    util][   INFO] ------------------------------------
[2015-03-17 16:43:26,550][    util][   INFO] Starting download    : 56292a1a-0d0d-48e0-b27b-27a059f5cd34
[2015-03-17 16:43:26,550][    util][   INFO] ------------------------------------
[2015-03-17 16:43:26,550][    util][   INFO] File name            : TCGA-G3-A25W-11A-12D-A16V-10_wgs_Illumina.bam
[2015-03-17 16:43:26,550][    util][   INFO] Download size        : 275958997261
[2015-03-17 16:43:26,550][    util][   INFO] Downloading file to  :
   /home/ubuntu/56292a1a-0d0d-48e0-b27b-27a059f5cd34.TCGA-G3-A25W-11A-12D-A16V-10_wgs_Illumina.bam
File: 56292a1a-0d0d-48e0-b27b-27a059f5cd34 [##                                  ] ETA:  00:26:31 162.98 M/s
```

GDC Website



The Next Generation Cancer Knowledge Network

Case Distribution by Disease Type

Data Availability Summary

Programs	2
Projects	46

Welcome to NCI's Genomic Data Commons

NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine. GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and the Cancer Genome Characterization Initiative (CGCI).

→ More about GDC

High Quality Data Sharing Enables Precision Medicine

GDC obtains validated datasets from NCI programs in which the strategies for tissue

Access Data

The **GDC Data Portal** provides a platform for efficiently querying and downloading high quality and complete data. GDC also provides a **GDC-Client** as an API for programmatic access.

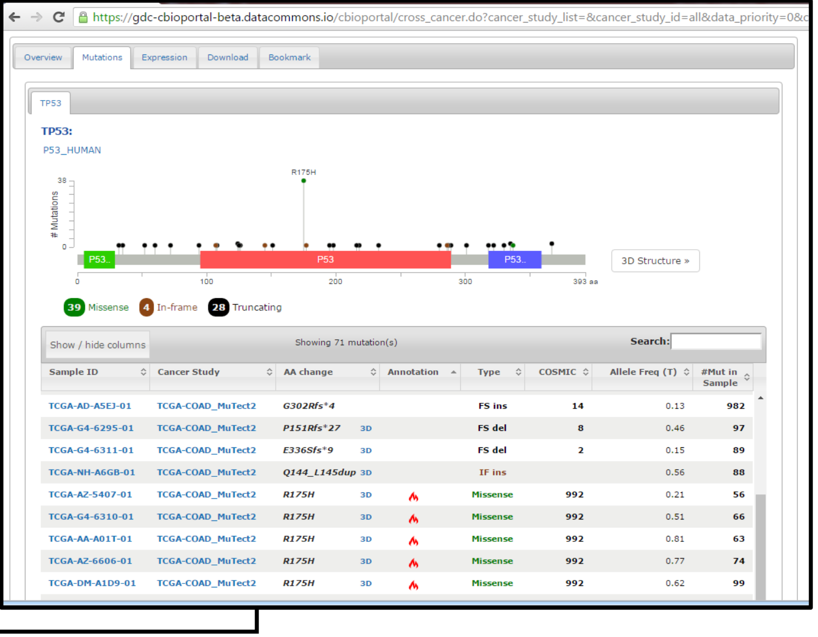
→ More about Accessing Data

Submit Data

GDC provides tools to guide data submission including web based tools for submitting clinical, biospecimen and small volumes of molecular data as well as client tools for submitting large, high volume molecular data. A secure API is also available for batch data submissions.

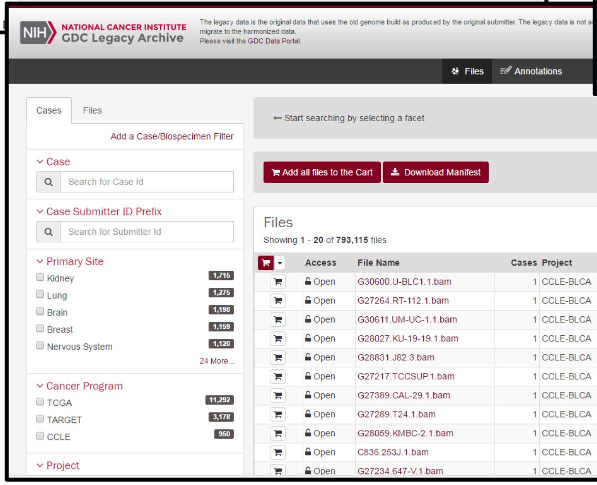
→ More about Submitting Data

Visualization Tools



Showing 71 mutation(s)

Sample ID	Cancer Study	AA change		Annotation	Type	COSMIC	Allele Freq (T)	#Mut in Sample
TCGA-AD-A5EJ-01	TCGA-COAD_MuTect2	G302Rfs*4			FS ins	14	0.13	982
TCGA-G4-6295-01	TCGA-COAD_MuTect2	P151Rfs*27	3D		FS del	8	0.46	97
TCGA-G4-6311-01	TCGA-COAD_MuTect2	E336Sfs*9	3D		FS del	2	0.15	89
TCGA-NH-A6GB-01	TCGA-COAD_MuTect2	Q144_L145dup	3D		IF ins		0.56	88
TCGA-AZ-5407-01	TCGA-COAD_MuTect2	R175H	3D	🔥	Missense	992	0.21	56
TCGA-G4-6310-01	TCGA-COAD_MuTect2	R175H	3D	🔥	Missense	992	0.51	66
TCGA-AA-A01T-01	TCGA-COAD_MuTect2	R175H	3D	🔥	Missense	992	0.81	63
TCGA-AZ-6606-01	TCGA-COAD_MuTect2	R175H	3D	🔥	Missense	992	0.77	74
TCGA-DM-A1D9-01	TCGA-COAD_MuTect2	R175H	3D	🔥	Missense	992	0.62	99


NIH NATIONAL CANCER INSTITUTE GDC Legacy Archive
Files
Annotations
Cases
Files
Add all files to the Cart
Download Manifest

The legacy data is the original data that uses the old genome build as produced by the original submitter. The legacy data is not a migrate to the harmonized data.
Please visit the GDC Data Portal.

← Start searching by selecting a facet

Add a Case/Biospecimen Filter

Case

Search for Case Id

Case Submitter ID Prefix

Search for Submitter Id

Primary Site

- ☐ Kidney 1,715
- ☐ Lung 1,275
- ☐ Brain 1,198
- ☐ Breast 1,159
- ☐ Nervous System 1,120

24 More...

Cancer Program

- ☐ TCGA 11,292
- ☐ TARGET 3,178
- ☐ CCLE 950

Project

Files

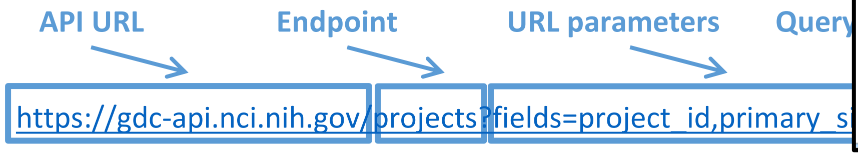
Showing **1** - **20** of **793,115** files

	Access	File Name	Cases	Project
	Open	G30600.U-BLC1.1.bam	1	CCLE-BLCA
	Open	G27264.RT-112.1.bam	1	CCLE-BLCA
	Open	G30611.UM-UC-1.1.bam	1	CCLE-BLCA
	Open	G28027.KU-19-19.1.bam	1	CCLE-BLCA
	Open	G28831.J82.3.bam	1	CCLE-BLCA
	Open	G27217.TCCSUP.1.bam	1	CCLE-BLCA
	Open	G27389.CAL-29.1.bam	1	CCLE-BLCA
	Open	G27289.T24.1.bam	1	CCLE-BLCA
	Open	G28059.KMBC-2.1.bam	1	CCLE-BLCA
	Open	C836.253J.1.bam	1	CCLE-BLCA
	Open	G27234.647-V.1.bam	1	CCLE-BLCA

Legacy Archive

GDC API

```
{
 "data": {
  "hits": [
    {"project_id": "TCGA-SKCM","primary_site": "Skin"}
   , {"project_id": "TCGA-PCPG","primary_site": "Nervous System"}
   , {"project_id": "TCGA-LAML","primary_site": "Blood"}
   , {"project_id": "TCGA-CNTL","primary_site": "Not Applicable"}
   , {"project_id": "TCGA-UVM","primary_site": "Eye"}
```

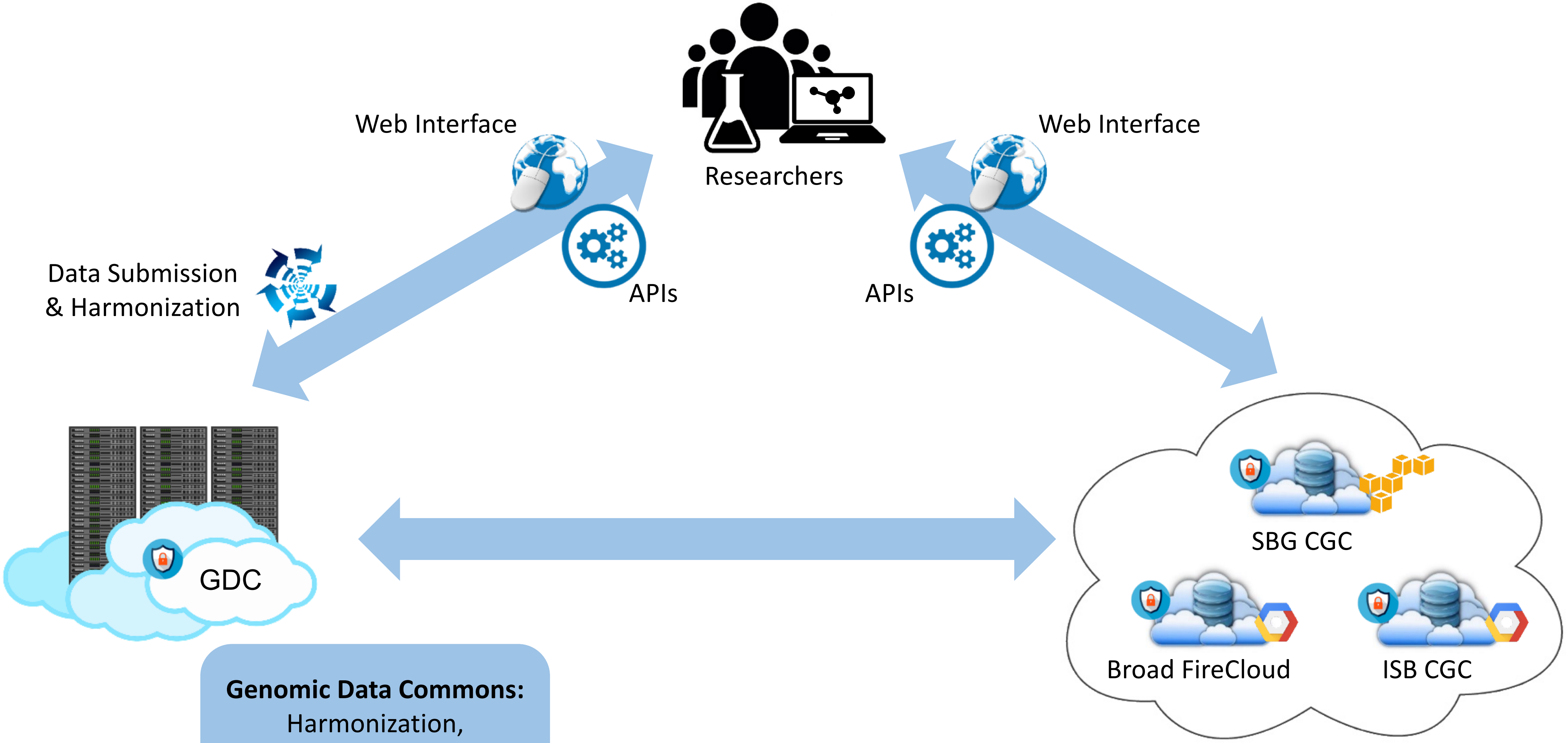


https://gdc-api.nci.nih.gov/projects?fields=project_id,primary_si

The NCI Cloud Resources

Understanding how to meet the research community's need to analyze large-scale cancer genomic and clinical data

GDC and the NCI Cloud Resources



Genomic Data Commons: Harmonization, Visualization, & Download

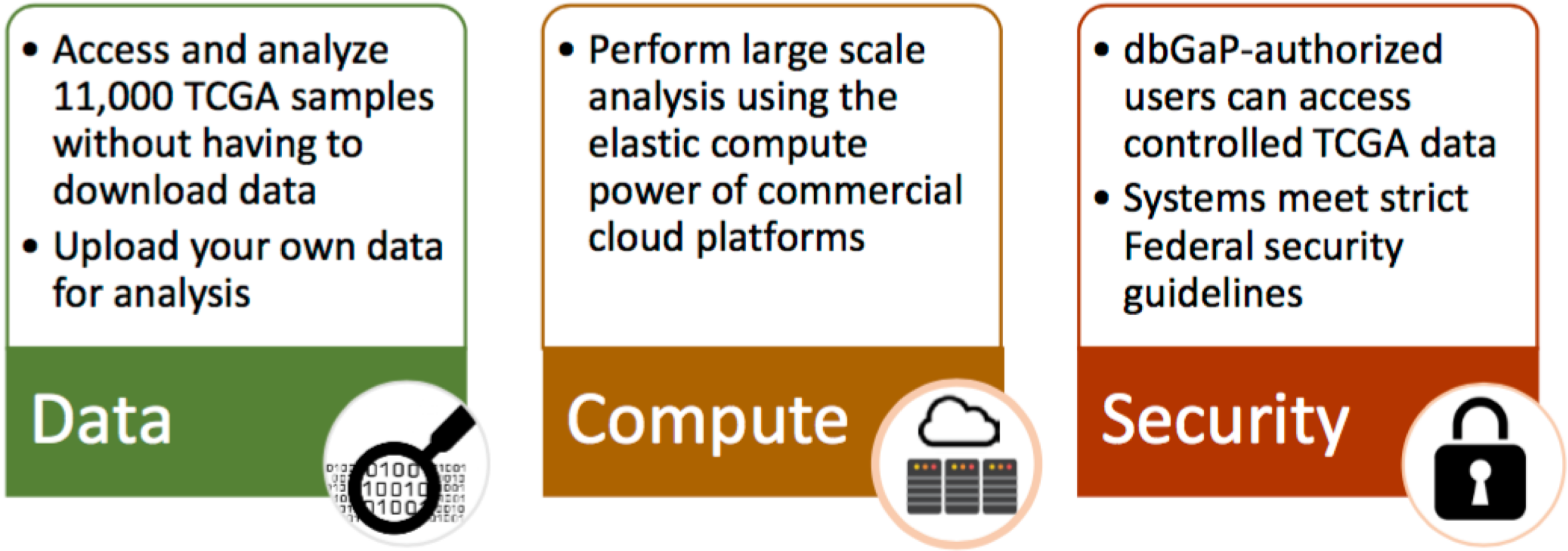
Cloud Resources: Compute, Pipelines, Workspaces

🛡 Authentication & Authorization thru eRA Commons & dbGaP

NCI Cloud Resources

The Cloud Resources provide:

- Access to large cancer data sets without need to download
- Access to popular analysis tools and pipelines
- Ability for researchers to bring their own data to the Cloud Resources
- Ability for researchers to bring their own tools and pipelines to the data
- Workspaces, for researchers to save and share their data and results of analyses



- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Data

- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Compute

- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

Security

Democratize access to cancer datasets, and to provide cost-effective computational capacity to the cancer research community.

#NCICloud

Three NCI Cloud Resources

Broad Institute

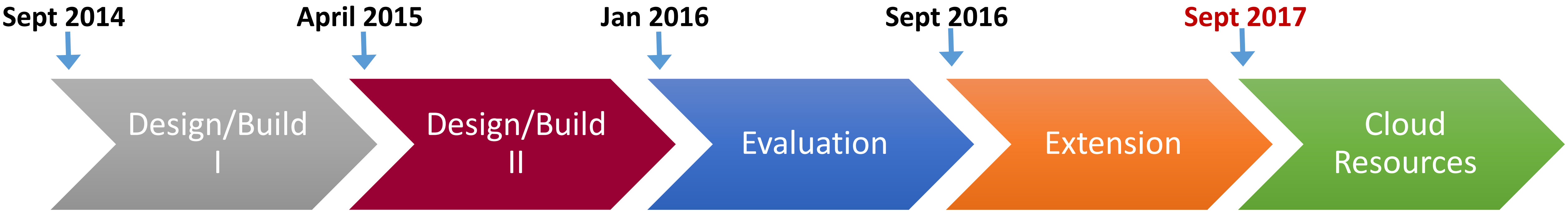
- PI: Anthony Philippakis
- Google Cloud
- Firehose in the cloud including Broad best practices workflows
- http://firecloud.org

Institute for Systems Biology

- PI: Ilya Shmulevich
- Google Cloud
- Leverage Google infrastructure; Novel query and visualization
- http://cgc.systemsbiology.net/

Seven Bridges Genomics

- PI: Brandi Davis-Dusenbery
- Amazon Web Services
- Interactive data exploration; > 30 public pipelines
- http://www.cancergenomicscloud.org



Sept 2014

Design/Build I

April 2015

Design/Build II

Jan 2016

Evaluation

Sept 2016

Extension

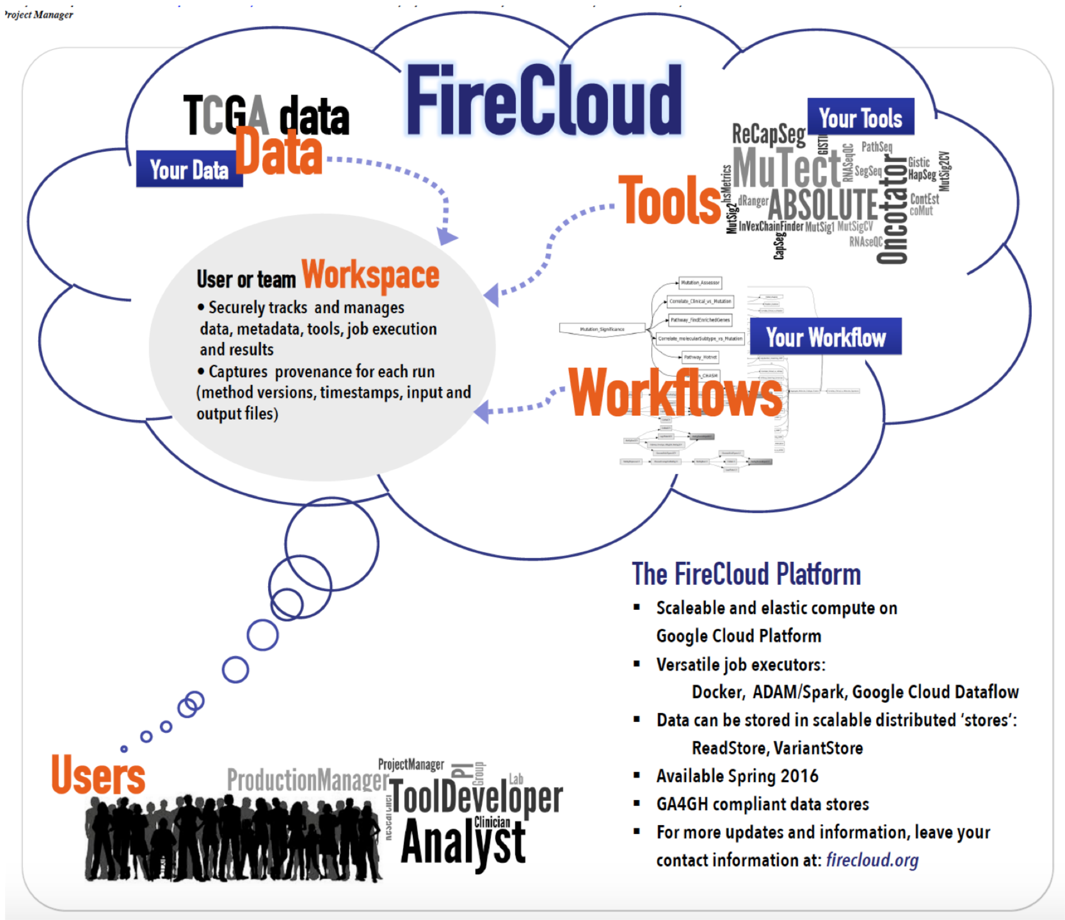
Sept 2017

Cloud Resources

Broad Institute Cloud Resource

- Targeted at users performing analyses at scale.
- Modeled after their Firehose analysis infrastructure developed for the TCGA program.
- Users can upload their own data and tools and/or run the Broad's best practice tools and pipelines on pre-loaded data.

http://firecloud.org



User or team Workspace

- Securely tracks and manages data, metadata, tools, job execution and results
- Captures provenance for each run (method versions, timestamps, input and output files)

The FireCloud Platform

- Scaleable and elastic compute on Google Cloud Platform
- Versatile job executors:
  Docker, ADAM/Spark, Google Cloud Dataflow
- Data can be stored in scalable distributed 'stores':
  ReadStore, VariantStore
- Available Spring 2016
- GA4GH compliant data stores
- For more updates and information, leave your contact information at: *firecloud.org*

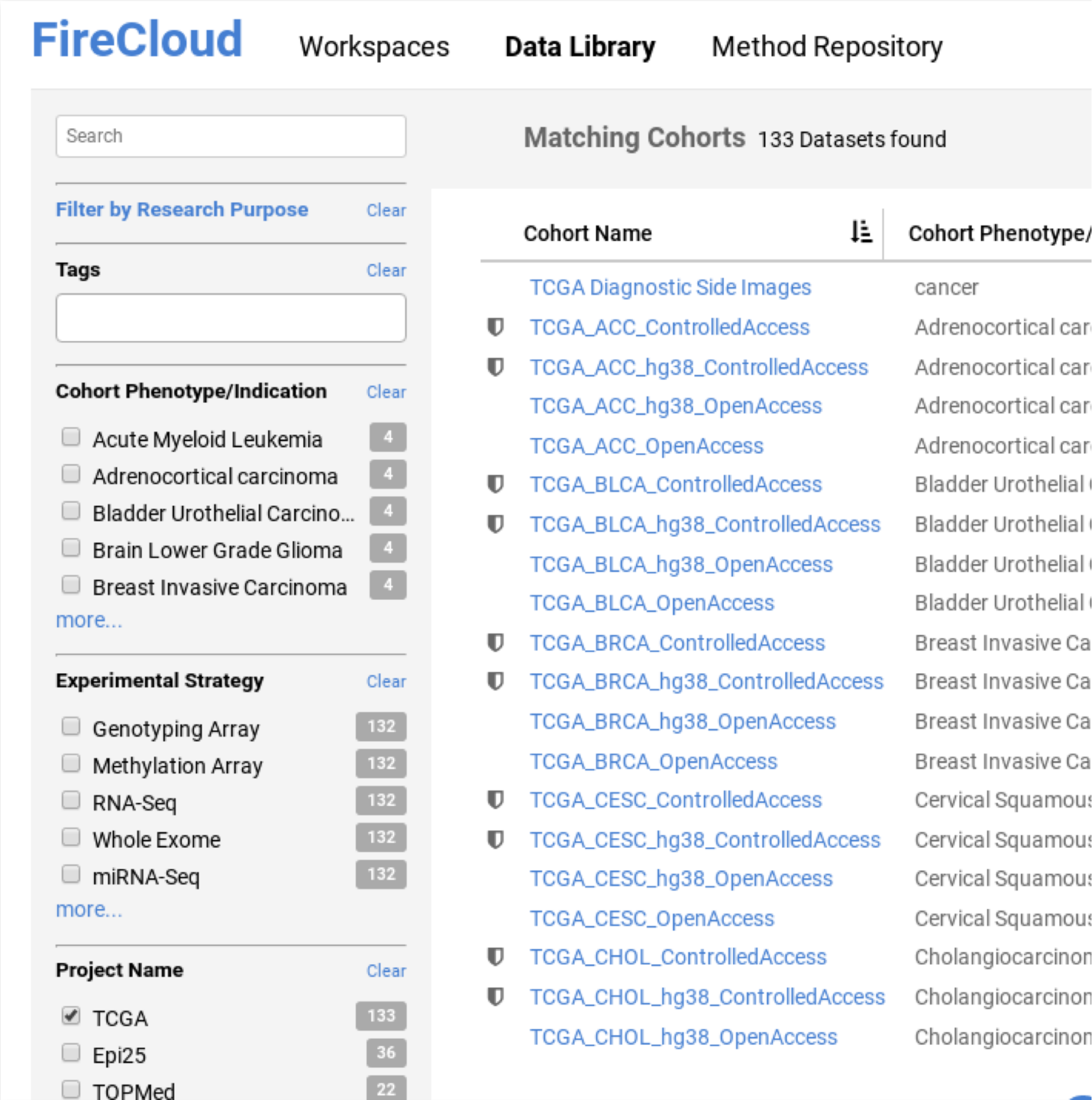
The Data Library

The primary tool for discovering datasets at Broad and beyond

Broad’s Genomics Platform has been delivering all WGS projects into FireCloud for a year.

Recently begun cataloguing all data into the Data Library for discovery and access.

Users can search the datasets and filter datasets by the data use restrictions.



Search

Filter by Research Purpose Clear

Tags Clear

Cohort Phenotype/Indication Clear

☐ Acute Myeloid Leukemia	4
☐ Adrenocortical carcinoma	4
☐ Bladder Urothelial Carcino...	4
☐ Brain Lower Grade Glioma	4
☐ Breast Invasive Carcinoma	4

more...

Experimental Strategy Clear

☐ Genotyping Array	132
☐ Methylation Array	132
☐ RNA-Seq	132
☐ Whole Exome	132
☐ miRNA-Seq	132

more...

Project Name Clear

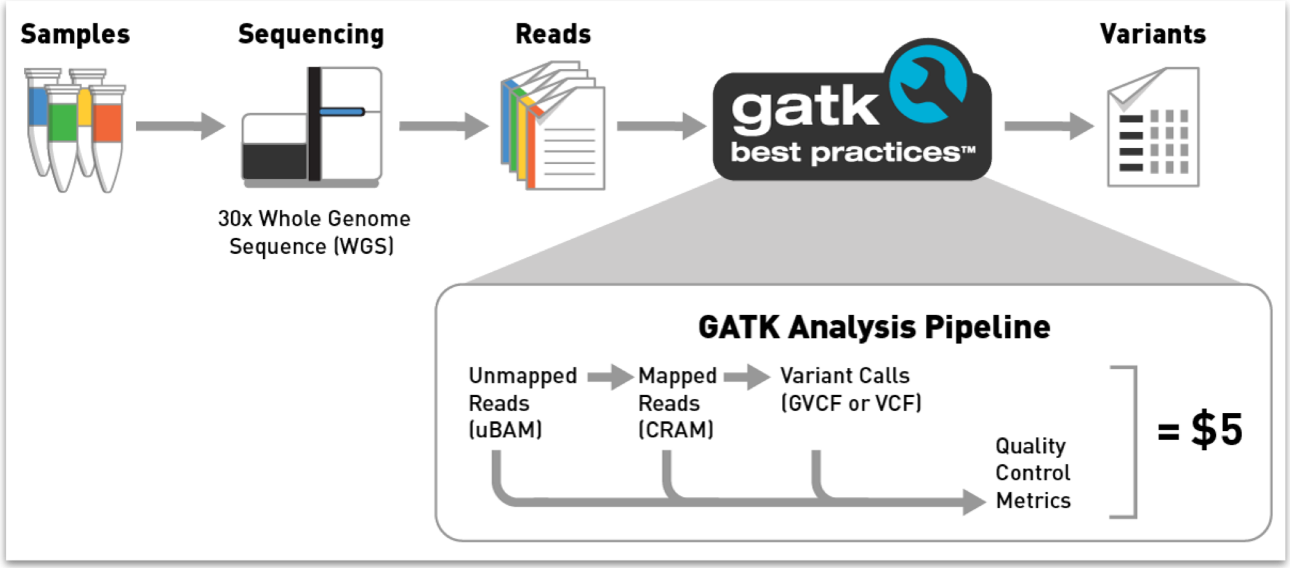
☑ TCGA	133
☐ Epi25	36
☐ TOPMed	22

Matching Cohorts 133 Datasets found

	Cohort Name	Cohort Phenotype/
	TCGA Diagnostic Side Images	cancer
🛡	TCGA_ACC_ControlledAccess	Adrenocortical car
🛡	TCGA_ACC_hg38_ControlledAccess	Adrenocortical car
	TCGA_ACC_hg38_OpenAccess	Adrenocortical car
	TCGA_ACC_OpenAccess	Adrenocortical car
🛡	TCGA_BLCA_ControlledAccess	Bladder Urothelial
🛡	TCGA_BLCA_hg38_ControlledAccess	Bladder Urothelial
	TCGA_BLCA_hg38_OpenAccess	Bladder Urothelial
	TCGA_BLCA_OpenAccess	Bladder Urothelial
🛡	TCGA_BRCA_ControlledAccess	Breast Invasive Ca
🛡	TCGA_BRCA_hg38_ControlledAccess	Breast Invasive Ca
	TCGA_BRCA_hg38_OpenAccess	Breast Invasive Ca
	TCGA_BRCA_OpenAccess	Breast Invasive Ca
🛡	TCGA_CESC_ControlledAccess	Cervical Squamous
🛡	TCGA_CESC_hg38_ControlledAccess	Cervical Squamous
	TCGA_CESC_hg38_OpenAccess	Cervical Squamous
	TCGA_CESC_OpenAccess	Cervical Squamous
🛡	TCGA_CHOL_ControlledAccess	Cholangiocarcinom
🛡	TCGA_CHOL_hg38_ControlledAccess	Cholangiocarcinom
	TCGA_CHOL_hg38_OpenAccess	Cholangiocarcinom

The $5 Genome: Pipeline Optimizations

- Broad is optimizing production pipelines with a commitment to openness, transparency, and continued improvements in cost and performance
- Example: Germline GATK best practices
  - $45/sample* in 2016
  - $13.50/sample in 2017
  - $5/sample in 2018
- Pipelines will be available to run in FireCloud and will also be in Dockstore.



Optimized somatic best practices coming soon!

* Cloud compute costs from Google Cloud Platform

FireCloud is part of the Data Biosphere

FireCloud will evolve into a citizen of an interoperable world through principles outlined across the *Data Biosphere.*

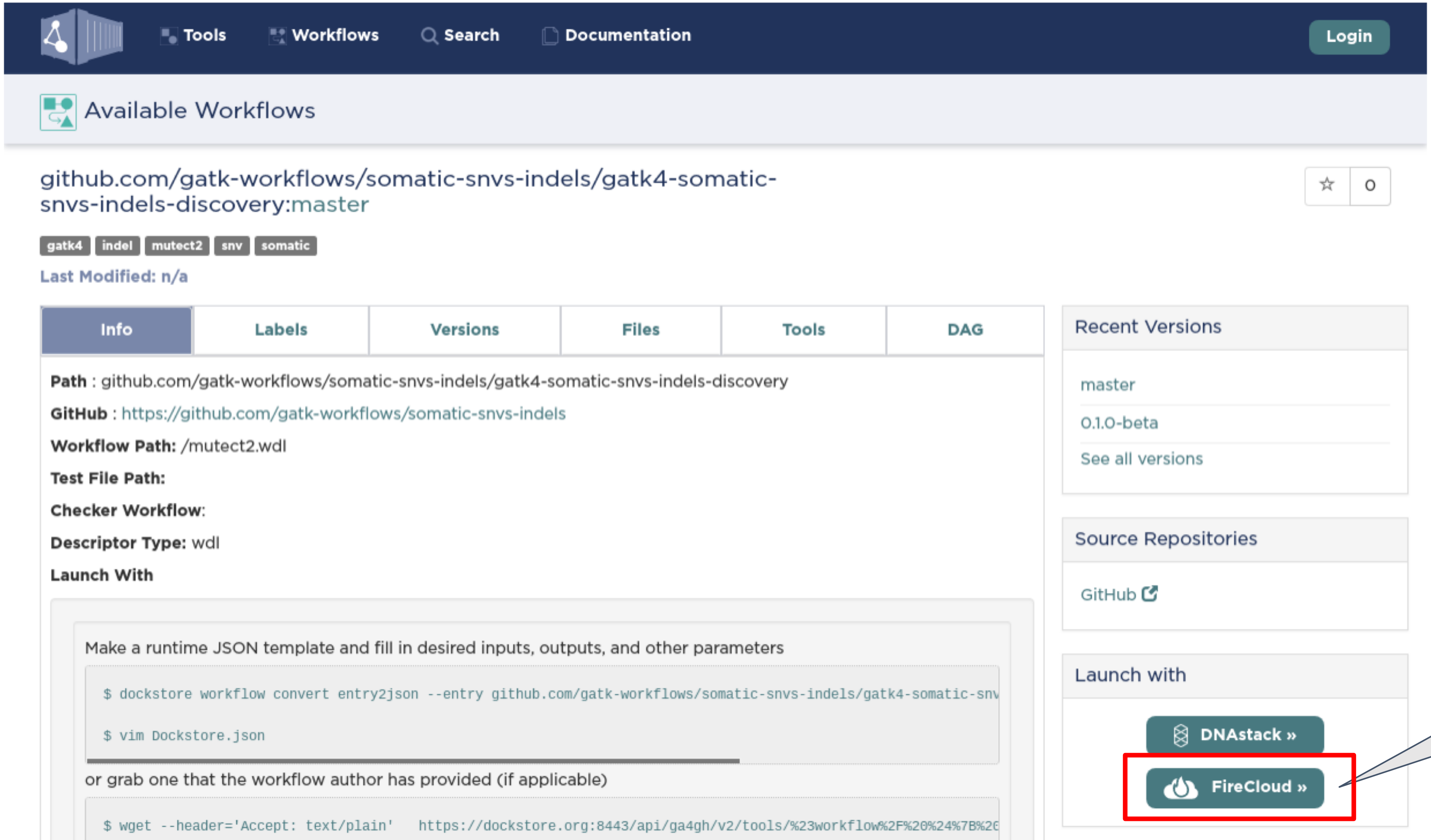
The Biosphere is a collaboration among institutions working on data platforms that will serve several large-scale, high-profile biomedical research projects.

Principles

- Open
- Standards Based
- Modular
- Community Driven

Initial collaborators in the Data Biosphere are building data platforms for the NCI Data Commons, NIH Data Commons, All of Us Research Program, Human Cell Atlas, Gabriella Miller Kids First, and others.

First integration with Data Biosphere: Dockstore



Available Workflows

github.com/gatk-workflows/somatic-snvs-indels/gatk4-somatic-snvs-indels-discovery:master

Last Modified: n/a

Path : github.com/gatk-workflows/somatic-snvs-indels/gatk4-somatic-snvs-indels-discovery
GitHub : https://github.com/gatk-workflows/somatic-snvs-indels
Workflow Path: /mutect2.wdl
Test File Path:
Checker Workflow:
Descriptor Type: wdl
Launch With

Make a runtime JSON template and fill in desired inputs, outputs, and other parameters

```
$ dockstore workflow convert entry2json --entry github.com/gatk-workflows/somatic-snvs-indels/gatk4-somatic-snv

$ vim Dockstore.json
```

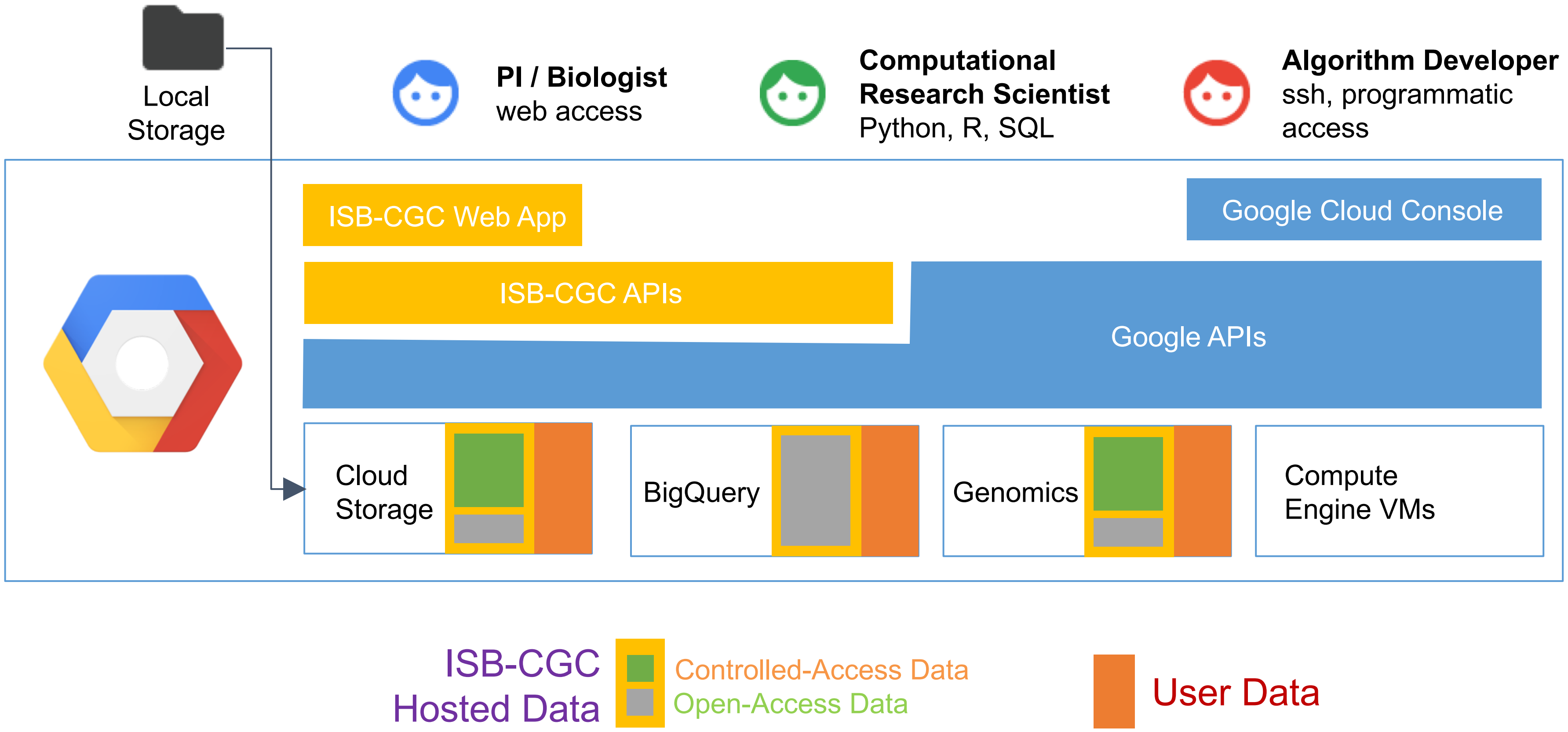
or grab one that the workflow author has provided (if applicable)

```
$ wget --header='Accept: text/plain'   https://dockstore.org:8443/api/ga4gh/v2/tools/%23workflow%2F%20%24%7B%2€
```

WDL workflows on Dockstore.org can now be launched with FireCloud!

ISB Cancer Genomics Cloud (ISB-CGC)

- Closely tied with Google Cloud Platform tools including BigQuery, App Engine, Cloud Datalab, Google Genomics, and Compute Engine



Local Storage

PI / Biologist
web access

Computational Research Scientist
Python, R, SQL

Algorithm Developer
ssh, programmatic access

ISB-CGC Web App

Google Cloud Console

ISB-CGC APIs

Google APIs

Cloud Storage

BigQuery

Genomics

Compute Engine VMs

ISB-CGC Hosted Data

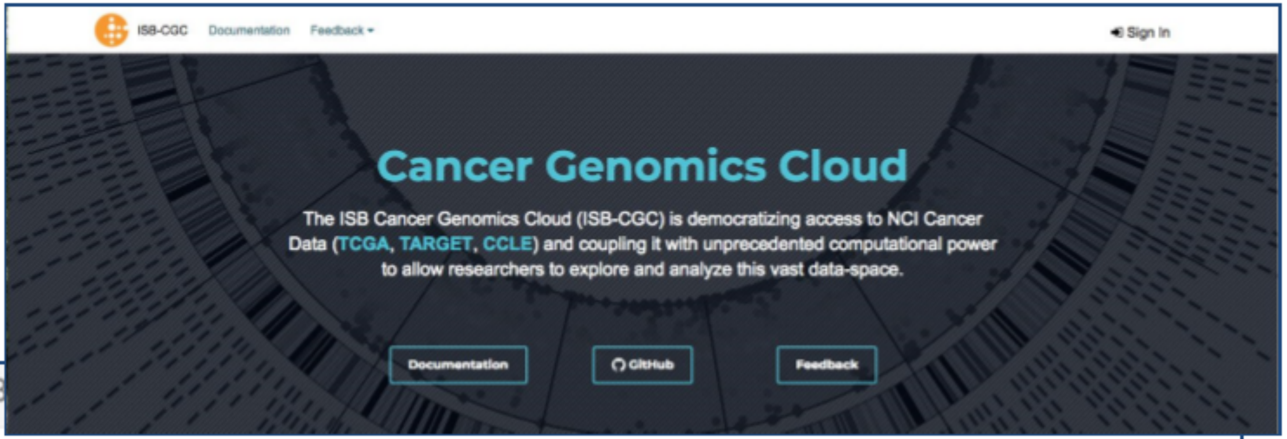
Controlled-Access Data

Open-Access Data

User Data

http://cgc.systemsbiology.net/

There are three primary ways of working on ISB-CGC.

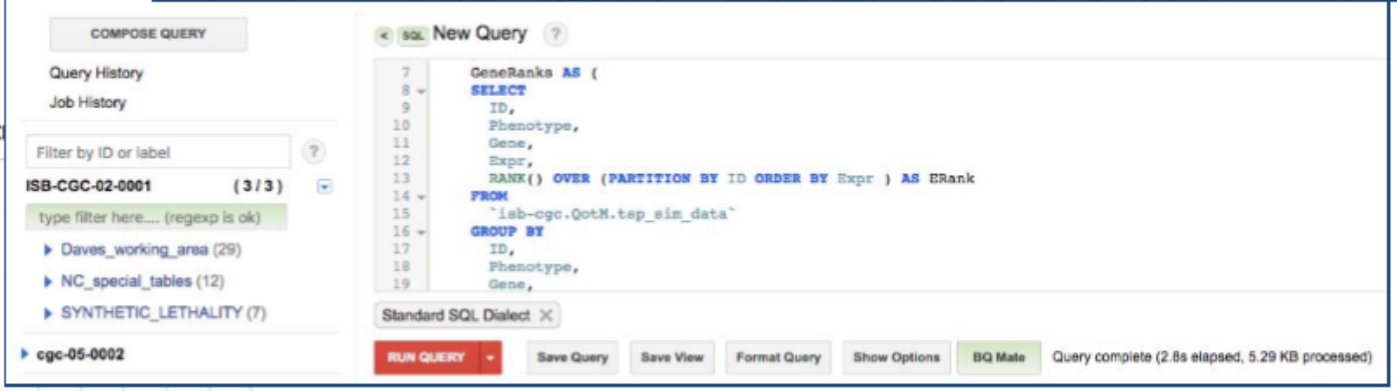


Cancer Genomics Cloud

The ISB Cancer Genomics Cloud (ISB-CGC) is democratizing access to NCI Cancer Data (TCGA, TARGET, CCLE) and coupling it with unprecedented computational power to allow researchers to explore and analyze this vast data-space.

ISB-CGC Web App

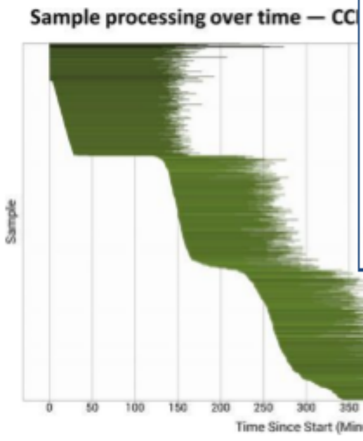
BigQuery



```
GeneRanks AS (
SELECT
  ID,
  Phenotype,
  Gene,
  Expr,
  RANK() OVER (PARTITION BY ID ORDER BY Expr ) AS ERank
FROM
  `isb-cgc.QotM.tsp_sim_data`
GROUP BY
  ID,
  Phenotype,
  Gene,
```

Query complete (2.8s elapsed, 5.29 KB processed)

Pipelines



What you choose depends on the question and what you're comfortable with

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Method 2: Working with BigQuery

R

bigrquery and bigQueryR

Shiny by RStudio

googleAuthR

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Pre-built VM images

IP[y]: IPython

jupyter

Cloud notebooks and workspaces.

Cloud Datalab

SQL

jupyter

Google BigQuery plays well with others.

Google BigQuery it's *great* for answering questions

Q: How many samples have a mutation in PARP1?

Use standard SQL to answer it:

```
SELECT
   project_short_name,
   COUNT(DISTINCT(sample_barcode_tumor)) AS n
FROM
   `isb-cgc.TCGA_hg38_data_v0.Somatic_Mutation_DR10`
WHERE
   Hugo_Symbol = 'PARP1'
GROUP BY
   project_short_name
ORDER BY
   n DESC
```

Easy to join tables on any shared variable.

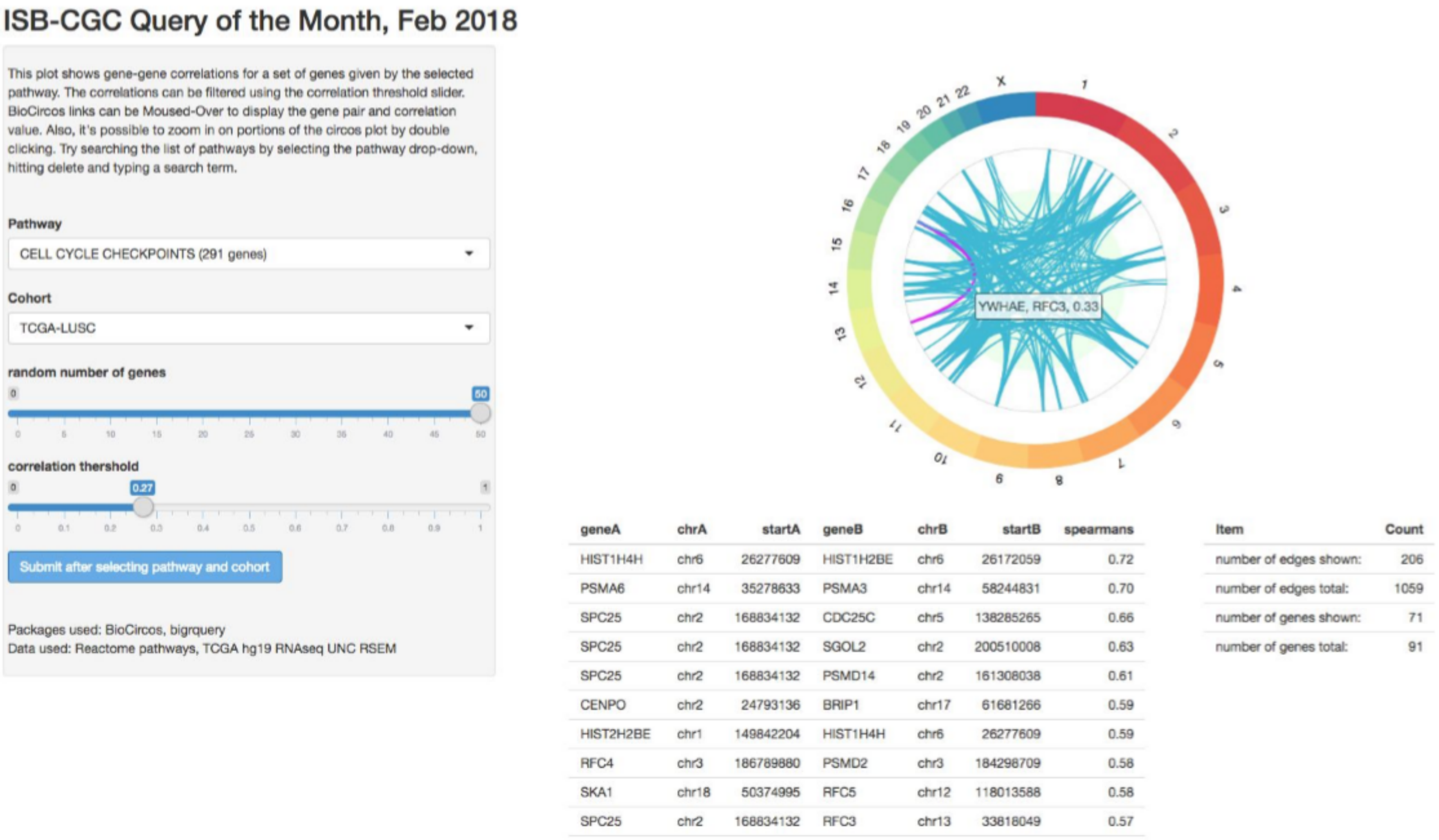
Lots of built in functions for math, string processing, etc

Can process massive amounts of data in parallel.

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Query Of The Month Club

Spearman correlations using RNA-seq data and pathway definitions



ISB-CGC Query of the Month, Feb 2018

This plot shows gene-gene correlations for a set of genes given by the selected pathway. The correlations can be filtered using the correlation threshold slider. BioCircos links can be Moused-Over to display the gene pair and correlation value. Also, it's possible to zoom in on portions of the circos plot by double clicking. Try searching the list of pathways by selecting the pathway drop-down, hitting delete and typing a search term.

Pathway

CELL CYCLE CHECKPOINTS (291 genes)

Cohort

TCGA-LUSC

random number of genes

0 — 50

correlation thershold

0 — 0.27 — 1

Submit after selecting pathway and cohort

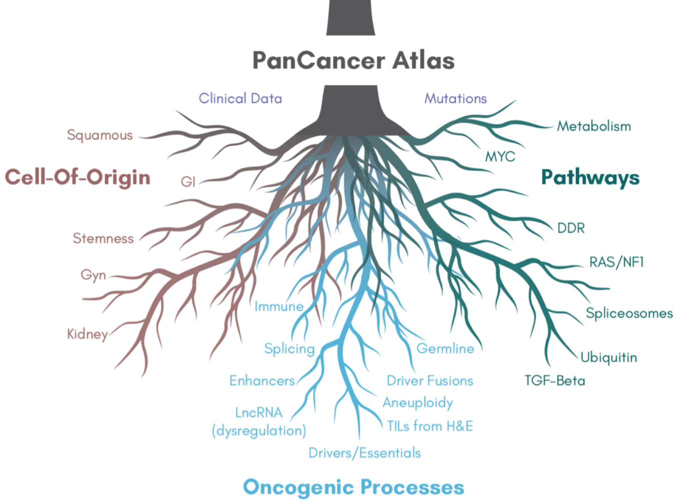
Packages used: BioCircos, bigrquery
Data used: Reactome pathways, TCGA hg19 RNAseq UNC RSEM

geneA	chrA	startA	geneB	chrB	startB	spearmans
HIST1H4H	chr6	26277609	HIST1H2BE	chr6	26172059	0.72
PSMA6	chr14	35278633	PSMA3	chr14	58244831	0.70
SPC25	chr2	168834132	CDC25C	chr5	138285265	0.66
SPC25	chr2	168834132	SGOL2	chr2	200510008	0.63
SPC25	chr2	168834132	PSMD14	chr2	161308038	0.61
CENPO	chr2	24793136	BRIP1	chr17	61681266	0.59
HIST2H2BE	chr1	149842204	HIST1H4H	chr6	26277609	0.59
RFC4	chr3	186789880	PSMD2	chr3	184298709	0.58
SKA1	chr18	50374995	RFC5	chr12	118013588	0.58
SPC25	chr2	168834132	RFC3	chr13	33818049	0.57

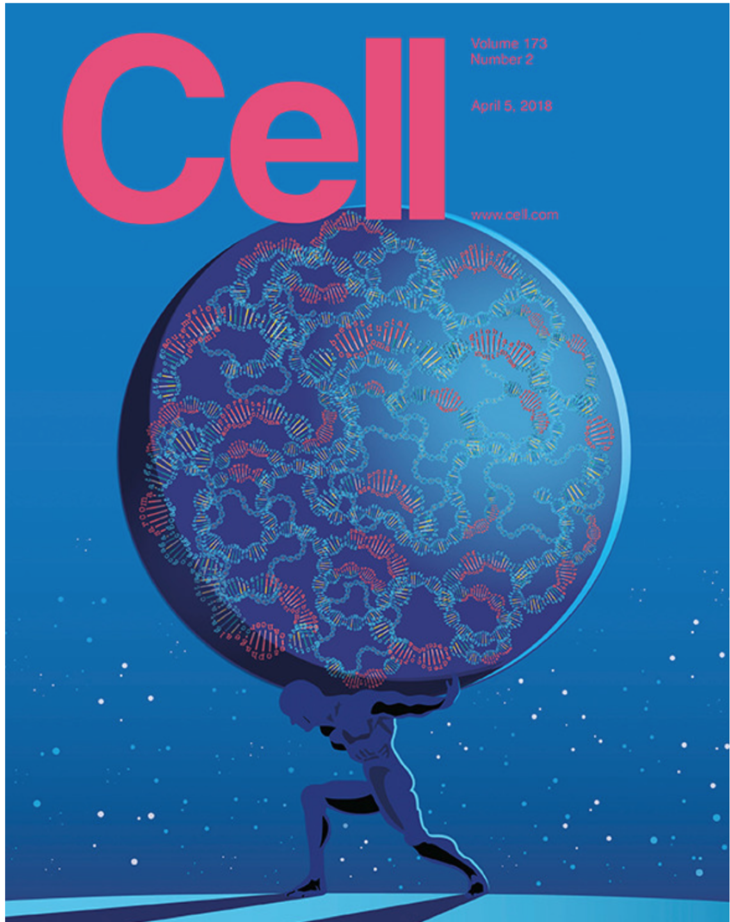
Item	Count
number of edges shown:	206
number of edges total:	1059
number of genes shown:	71
number of genes total:	91

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

ISB-CGC a key resource
for
TCGA #PanCancerAtlas



- Germline, Fusion, and Immune Response papers used ISB-CGC to access and compute on TCGA sequence data on the Google Cloud Platform
- Immune, MYC, and DDR papers used BigQuery and ISB-CGC data tables
- #PanCancerAtlas open-access tables now available in BigQuery (referenced from GDC page)
- The availability of PanCaner Atlas data in BigQuery enables easy integration with other public datasets through BigQuery



The Seven Bridges Cancer Genomics Cloud (CGC)

- A user-friendly, web-based portal for collaborative analysis of petabytes of multi-omic data alongside private data
- Built upon the SBG commercial cloud-based genomics platform
- For cancer genomics research and beyond

Easy data management

Scalable computation

Secure collaboration

Optimized bioinformatics algorithms

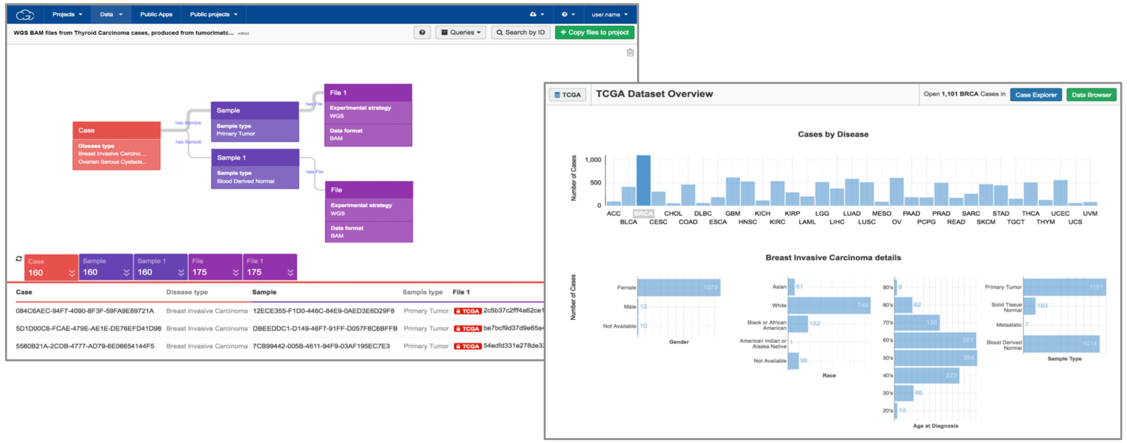
Flexible & fully reproducible methods

Extensible & developer-friendly platform

http://www.cancergenomicscloud.org

Available Resources

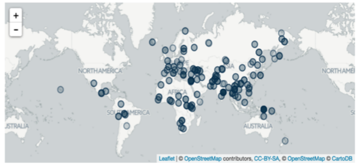
- Access **3$^+$ PB** of multi-omic public data through interactive query tools & APIs.
- Upload private data for analysis.
- Collaborate securely with colleagues anywhere.



WGS BAM files from Thyroid Carcinoma cases, produced from tumorimatc... edited

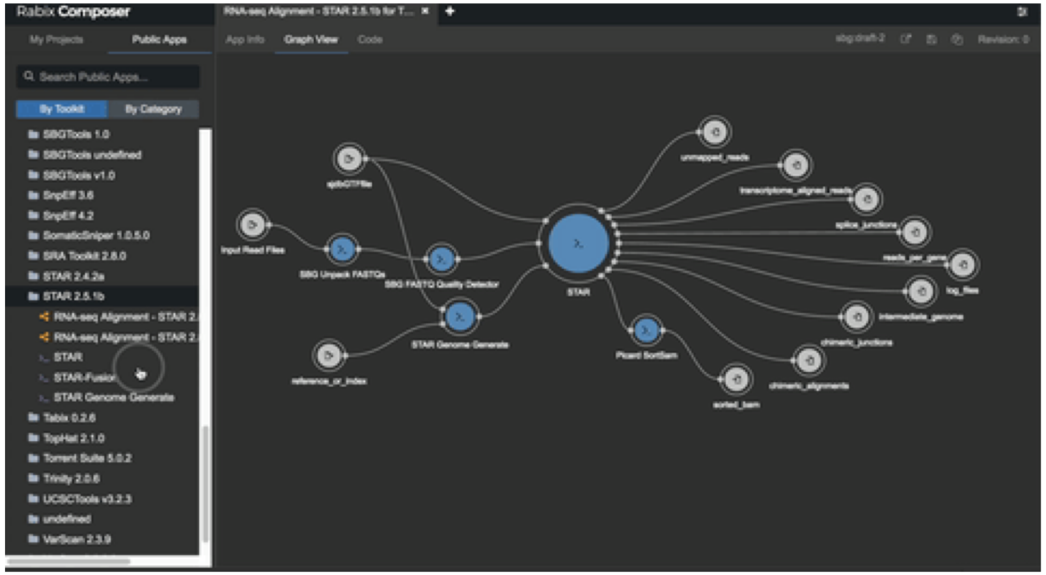
Case	Disease type	Sample	Sample type	File 1
084C6AEC-94F7-4090-8F3F-59FA9E89721A	Breast Invasive Carcinoma	12ECE355-F1D0-446C-84E9-0AED3E6D29F8	Primary Tumor	TCGA 2c5b37c2fff4a62ce1
5D1D00C6-FCAE-479E-AE1E-DE76EFD41D98	Breast Invasive Carcinoma	DBEEDDC1-D149-46F7-91FF-D057F8C6BFFB	Primary Tumor	TCGA be7bcf9d37d9e65a
5580B21A-2CDB-4777-AD79-6E06654144F5	Breast Invasive Carcinoma	7CB99442-005B-4611-94F9-03AF195EC7E3	Primary Tumor	TCGA 54edfd331e278de3

TCGA
THE CANCER IMAGING ARCHIVE
International Cancer Genome Consortium
PCAWG PanCancer Analysis OF WHOLE GENOMES
SIMONS FOUNDATION


NORTH AMERICA
SOUTH AMERICA
AUSTRALIA
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap © CartoDB

TARGET
CPTAC
CCLE Cancer Cell Line Encyclopedia

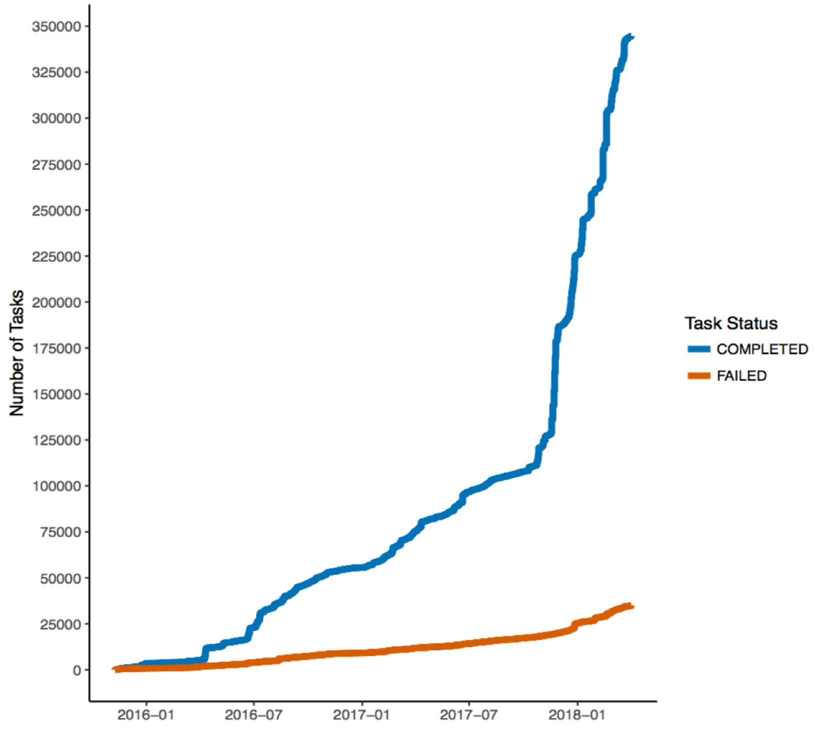
- Use the **360$^+$** cloud- and cost-optimized tools in the Public Apps library.
- Deploy custom tools using SDK (**Rabix**) & Jupyter notebook (**Data Cruncher**).
- Consult with **200$^+$** expert support staff.



Usage by the Research Community

3,100$^+$ users from **60**$^+$ countries have used the CGC to run **347,000**$^+$ computational tasks representing **465**$^+$ years of total compute time to:

- Detect aberrant splice junctions and splicing profiles across patient populations
- Identify neoantigens arising from novel gene fusion events
- Profile miRNA expression across patient populations
- Conduct HLA typing to identify neoantigens
- Compare viral infection patterns across patient populations
- Detect novel gene fusions from RNA-Seq data
- Identify cis-regulatory region variants across patient populations
- ...and much more



Scalable, Cost-Effective Research

Case Study #1: TCGA Immune Response Working Group

- Collaborative analysis with members of the Immune Response Working Group of The Cancer Genome Atlas (TCGA) Research Network
- Outcome: cost-optimized (<$0.30/sample), high-throughput HLA typing across ~9,000 TCGA RNA-Seq (fastq) files

Case Study #2: PanCancer Analysis of Whole Genomes (PCAWG) Study from International Cancer Genome Consortium (ICGC)

- High-throughput, harmonized analysis by Seven Bridges of all tumor and matched genomes in the dataset (~1,350)
- Outcome: rapid generation of ~65,000 output files (including ~5,000 VCFs) totaling 725 TB

Case Study #3: Independent Analysis on 45,000 Bacterial Genomes

- High-throughput analysis of 45,000 bacterial genomes accessed from SRA via API and analyzed using a custom workflow
- Outcome: analysis completed in ~1 week by a novice CGC user with no substantive assistance from the CGC team

The Seven Bridges Cloud Ecosystem:
Interoperable Data Access and Analysis to Drive Precision Medicine

Infrastructure

CANCER GENOMICS CLOUD
SEVEN BRIDGES

PDX

CAVATICA

FAIR4CURES

Interoperability

COMMON
WORKFLOW
LANGUAGE

Rabix
[Reproducible Analysis for Bioinformatics]

BioCompute
Objects

Global Alliance
for Genomics & Health

Partnerships

NIH

NATIONAL® CANCER INSTITUTE

MVP

Genomics
england

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

TCGA

PCAWG
PanCancer Analysis
OF WHOLE GENOMES

Bristol-Myers Squibb

Celgene

NEON
THERAPEUTICS

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol