The Cancer Genomics Cloud: a secure and scalable cloud-based platform to access, share and analyze multi-omics datasets

April 6, 2021

Sai Lakshmi Subramanian
Program Manager

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Acknowledgements

The Global Seven Bridges Team

CGC users and collaborators

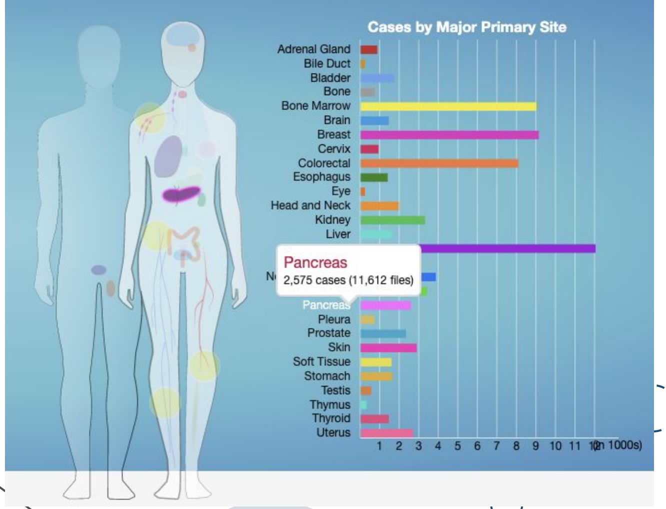
NATIONAL CANCER INSTITUTE

The Seven Bridges Cancer Research Data Commons Cloud Resource has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I and 75N91019D00024.

Institute for Systems Biology

BROAD INSTITUTE

THE UNIVERSITY OF CHICAGO



Pancreas
2,575 cases (11,612 files)

How do I identify the mutational burden (top mutated genes and somatic signature) in pancreatic adenocarcinoma cases?

I need access to different datasets, tools and workflows!

I need to share the results with my collaborators!

I am worried about security of analyzing my data in the cloud!

I am worried about cloud costs and billing!

I need to re-run my analysis using a different set of parameters!

My data is only available on XYZ cloud!

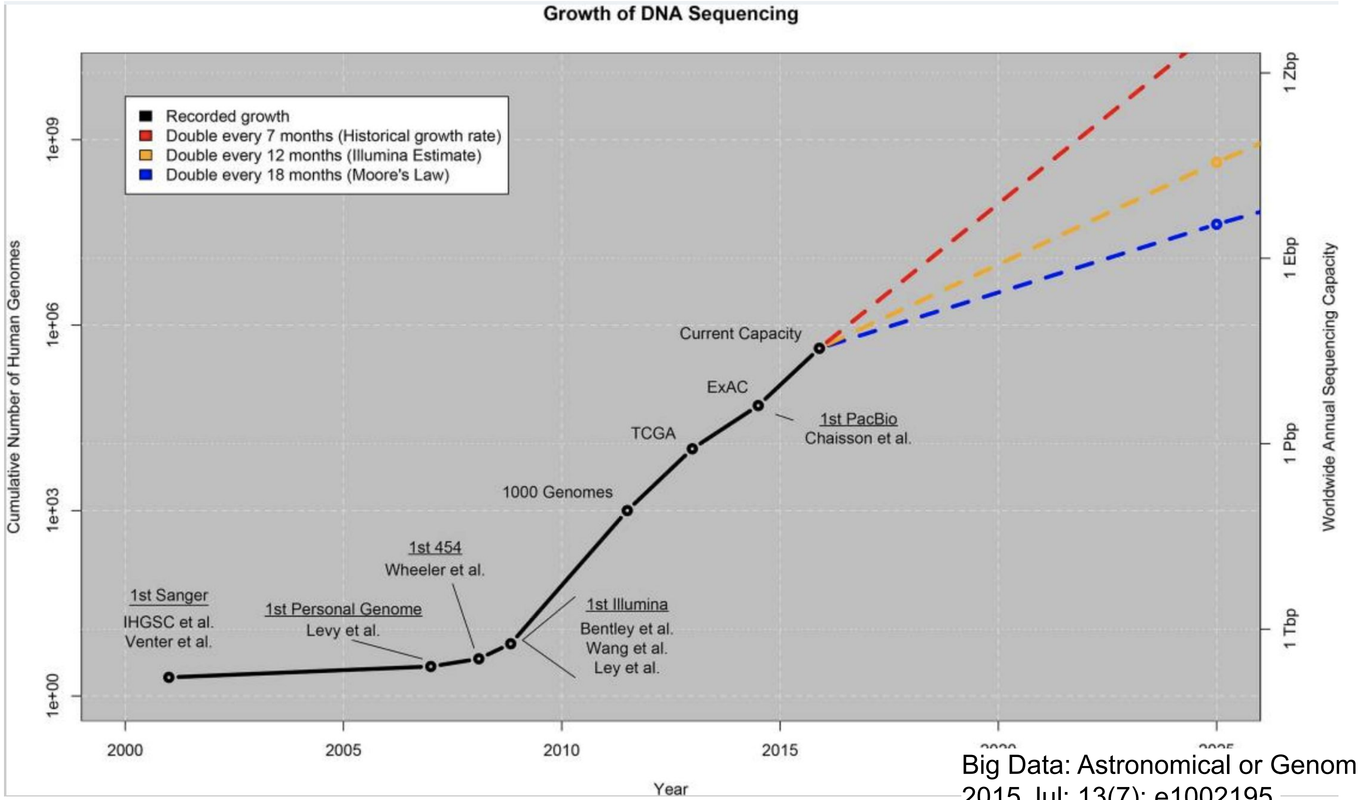
Image courtesy: https://portal.gdc.cancer.gov/

Agenda

- Background
- Access multi-omics datasets in the CGC
- Use Case 1: Identifying mutational burden in cancer - Analysis of TCGA datasets
- Use Case 2: microRNA biogenesis in cancer
- Questions/Discussion

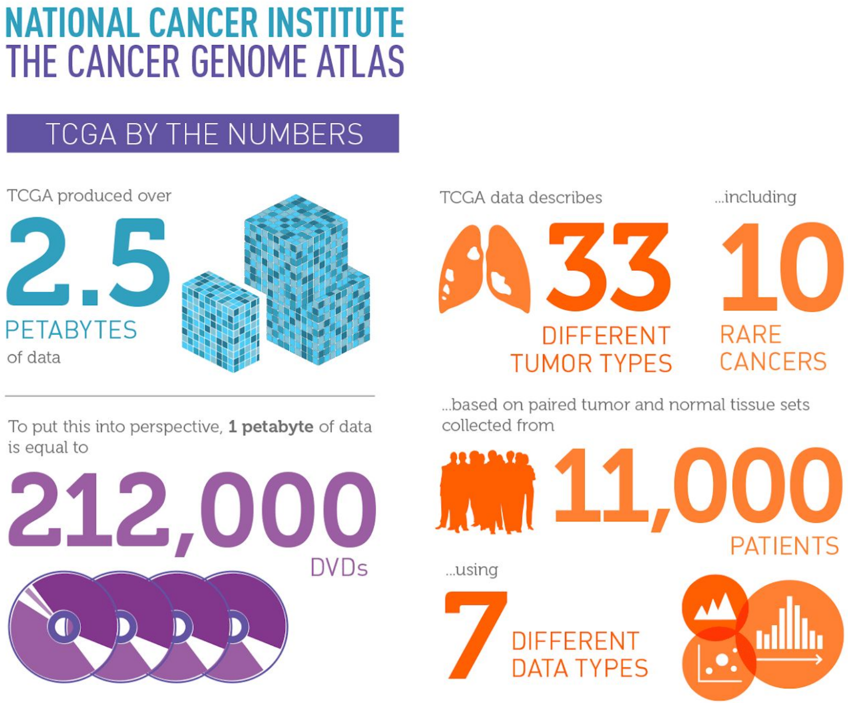
Background

Explosion of genomics data with ease of sequencing



Big Data: Astronomical or Genomical? Stephens et al; PLoS Biol. 2015 Jul; 13(7): e1002195.

Increasingly large datasets bring challenges to data analysis



NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs

TCGA data describes **33** DIFFERENT TUMOR TYPES

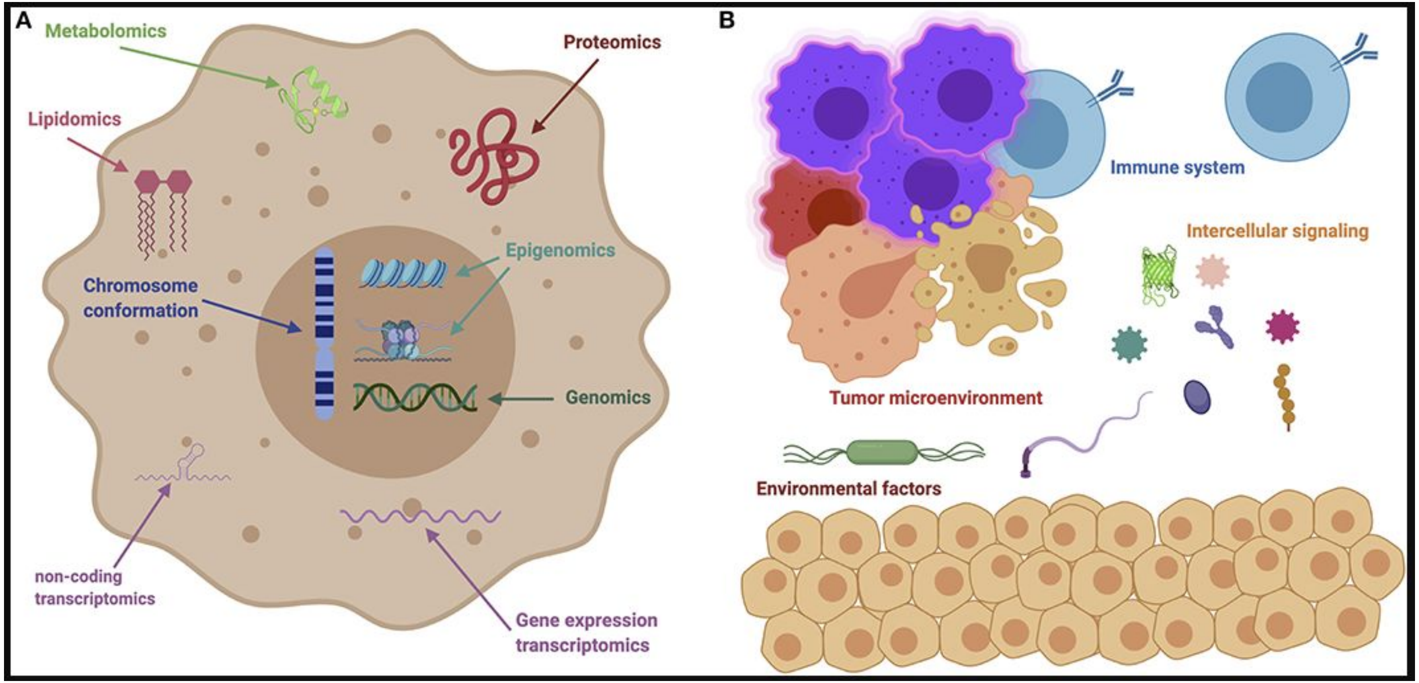
...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

www.cancer.gov/ccg

Multi-omic data is critical for cancer research



Cancer is a complex disease!

Comprehensively understanding the full picture of a research question requires examining multiple modalities

Guillermo de Anda-Jáuregui and Enrique Hernández-Lemus, Computational Oncology in the Multi-Omics Era: State of the Art. Front. Oncol., 07 April 2020 | https://doi.org/10.3389/fonc.2020.00423

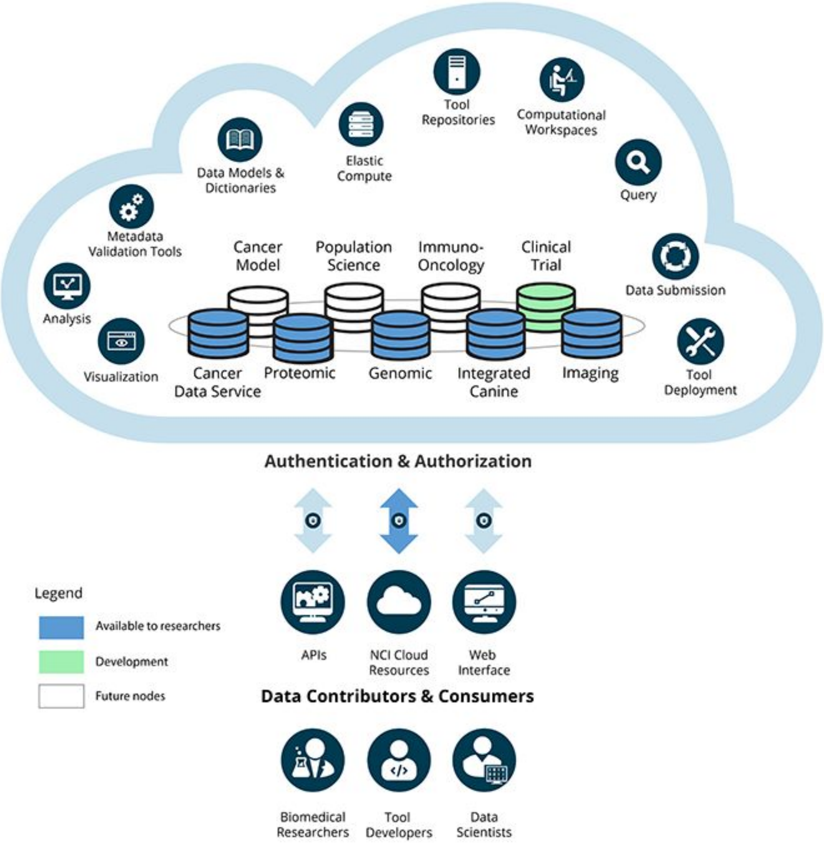
The Seven Bridges Cancer Genomics Cloud (CGC)

CANCER GENOMICS CLOUD
SEVEN BRIDGES

A Cloud Resource within the NCI Cancer Research Data Commons for secure storage, sharing & analysis of petabytes of public, multi-omic cancer datasets

https://datacommons.cancer.gov/cancer-research-data-commons

NCI Cancer Research Data Commons (CRDC)



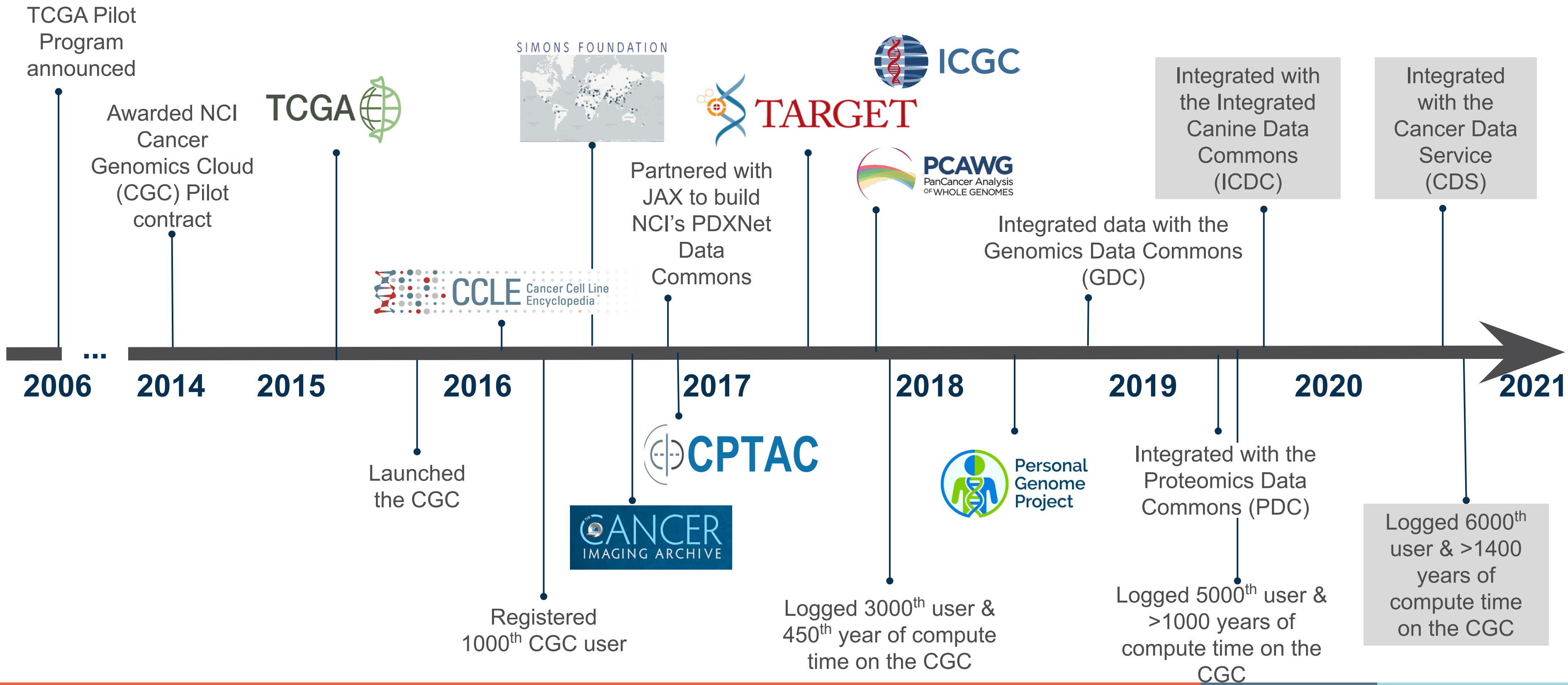
Legend

🟦 Available to researchers

🟩 Development

□ Future nodes

Growth of the Cancer Genomics Cloud Ecosystem


2006
...
2014
2015
2016
2017
2018
2019
2020
2021
TCGA
CCLE Cancer Cell Line Encyclopedia
SIMONS FOUNDATION
THE CANCER IMAGING ARCHIVE
CPTAC
TARGET
ICGC
PCAWG PanCancer Analysis OF WHOLE GENOMES
Personal Genome Project

TCGA Pilot Program announced

Awarded NCI Cancer Genomics Cloud (CGC) Pilot contract

Launched the CGC

Registered 1000th CGC user

Partnered with JAX to build NCI's PDXNet Data Commons

Logged 3000th user & 450th year of compute time on the CGC

Integrated data with the Genomics Data Commons (GDC)

Integrated with the Proteomics Data Commons (PDC)

Logged 5000th user & >1000 years of compute time on the CGC

Integrated with the Integrated Canine Data Commons (ICDC)

Integrated with the Cancer Data Service (CDS)

Logged 6000th user & >1400 years of compute time on the CGC

The CGC democratizes complex analyses in a FAIR data ecosystem

- A stable, secure, and highly customizable cloud storage and computing platform
- Promotes a **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR) data ecosystem
- A user-friendly portal for collaborative analysis of petabytes of public data alongside private data
- An optimized venue for reproducible data analysis using validated tools and pipelines

Easy data management

Secure collaboration & managed billing

Flexible & fully reproducible methods

Optimized bioinformatics algorithms

Scalable computation

Extensible & developer friendly tools

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016) doi:10.1038/sdata.2016.18

Accelerating cancer research

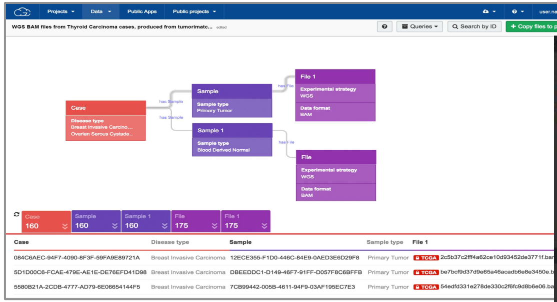
- Detect aberrant splice junctions and splicing profiles across patient populations
- Identify neoantigens arising from novel gene fusion events
- Profile miRNA expression across patient populations
- Conduct HLA typing to identify neoantigens
- Compare viral infection patterns across patient populations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

CANCER GENOMICS CLOUD
SEVEN BRIDGES

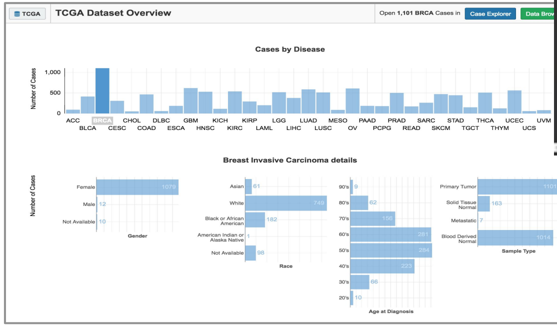
CGC provides an easy way to find and analyze data

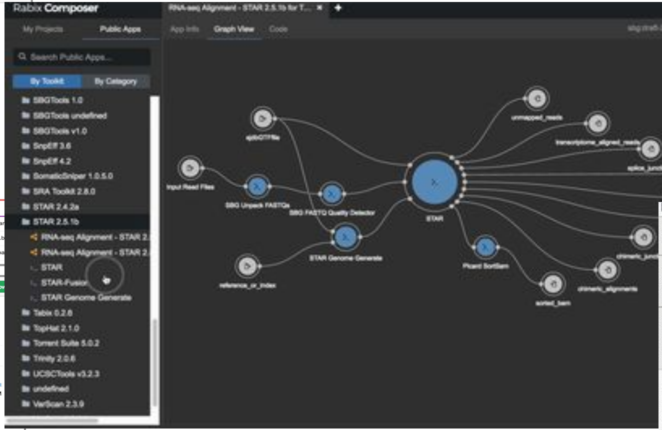
Visually explore and access 3$^+$ **PB** of multi-omic public data through interactive query tools & APIs.

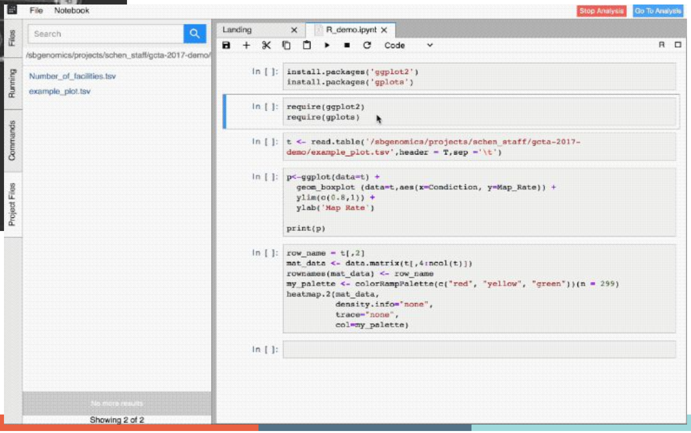
Use the **500$^+$** cloud- and cost-optimized tools in our Public Apps library OR deploy custom tools using **Rabix Composer**, Jupyter notebooks or R packages



Case	Disease type	Sample	Sample type	File 1
084C6AEC-94F7-4090-8F3F-59FA9E89721A	Breast Invasive Carcinoma	12ECE355-F1D0-446C-84E9-0AED3E6D29F8	Primary Tumor	TCGA 2c5b37c2fff4a62ce10d93452de3771f.bar
5D1D00C6-FCAE-479E-AE1E-DE76EFD41D98	Breast Invasive Carcinoma	DBEEDDC1-D149-46F7-91FF-D057F8C6BFFB	Primary Tumor	TCGA be7bcf9d37d9e65a46acadb6e8e3450e.b
5580B21A-2CDB-4777-AD79-6E06654144F5	Breast Invasive Carcinoma	7CB99442-005B-4611-94F9-03AF195EC7E3	Primary Tumor	TCGA 54edfd331e278de330c2f6fc9d8b6e06.ba







```
In [ ]: install.packages('ggplot2')
        install.packages('gplots')

In [ ]: require(ggplot2)
        require(gplots)

In [ ]: t <- read.table('/sbgenomics/projects/schen_staff/gcta-2017-
        demo/example_plot.tsv',header = T,sep ='\t')

In [ ]: p<-ggplot(data=t) +
          geom_boxplot (data=t,aes(x=Condiction, y=Map_Rate)) +
          ylim(c(0.8,1)) +
          ylab('Map Rate')

        print(p)

In [ ]: row_name = t[,2]
        mat_data <- data.matrix(t[,4:ncol(t)])
        rownames(mat_data) <- row_name
        my_palette <- colorRampPalette(c("red", "yellow", "green"))(n = 299)
        heatmap.2(mat_data,
                  density.info="none",
                  trace="none",
                  col=my_palette)

In [ ]:
```

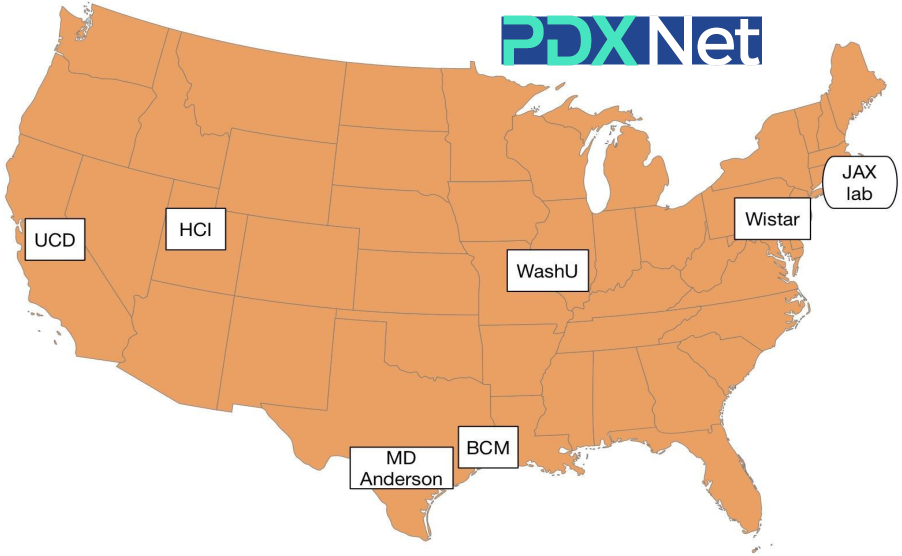
Empowering a coordinating center on the CGC

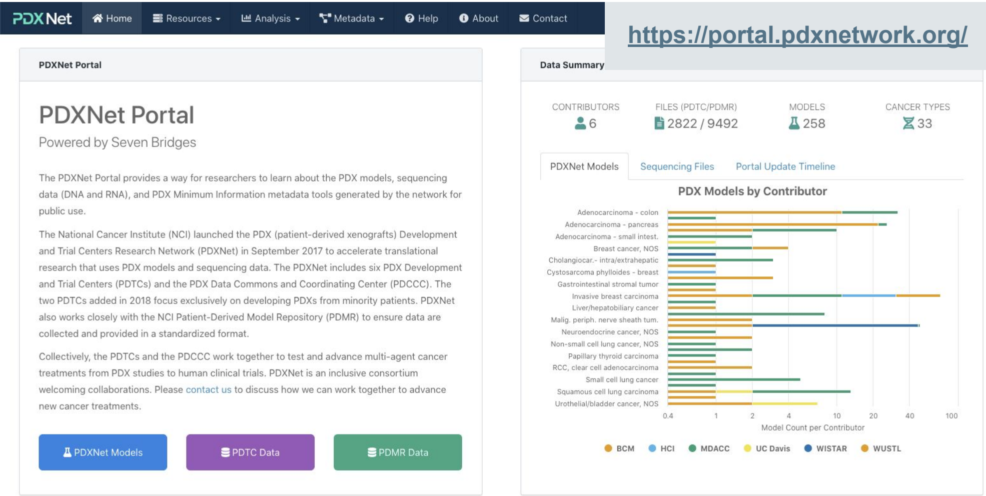
The Jackson Laboratory

PDX Data Commons and Coordination Center JAX-Seven Bridges

Collaborative and large-scale development and pre-clinical testing of targeted therapeutic agents in patient-derived models to advance the vision of cancer precision medicine.

- Data harmonized and securely shared
- Developed standardized PDX DNA-seq and RNA-seq workflows, available on the CGC
- Diverse models, metadata, and omics included





https://portal.pdxnetwork.org/

PDXNet Portal

PDXNet Portal

Powered by Seven Bridges

The PDXNet Portal provides a way for researchers to learn about the PDX models, sequencing data (DNA and RNA), and PDX Minimum Information metadata tools generated by the network for public use.

The National Cancer Institute (NCI) launched the PDX (patient-derived xenografts) Development and Trial Centers Research Network (PDXNet) in September 2017 to accelerate translational research that uses PDX models and sequencing data. The PDXNet includes six PDX Development and Trial Centers (PDTCs) and the PDX Data Commons and Coordinating Center (PDCCC). The two PDTCs added in 2018 focus exclusively on developing PDXs from minority patients. PDXNet also works closely with the NCI Patient-Derived Model Repository (PDMR) to ensure data are collected and provided in a standardized format.

Collectively, the PDTCs and the PDCCC work together to test and advance multi-agent cancer treatments from PDX studies to human clinical trials. PDXNet is an inclusive consortium welcoming collaborations. Please contact us to discuss how we can work together to advance new cancer treatments.

Data Summary

CONTRIBUTORS	FILES (PDTC/PDMR)	MODELS	CANCER TYPES
6	2822 / 9492	258	33

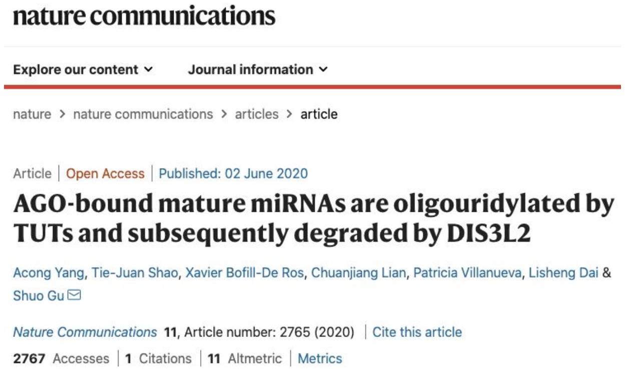
PDX Models by Contributor

Adenocarcinoma - colon
Adenocarcinoma - pancreas
Adenocarcinoma - small intest.
Breast cancer, NOS
Cholangiocar.- intra/extrahepatic
Cystosarcoma phylloides - breast
Gastrointestinal stromal tumor
Invasive breast carcinoma
Liver/hepatobiliary cancer
Malig. periph. nerve sheath tum.
Neuroendocrine cancer, NOS
Non-small cell lung cancer, NOS
Papillary thyroid carcinoma
RCC, clear cell adenocarcinoma
Small cell lung cancer
Squamous cell lung carcinoma
Urothelial/bladder cancer, NOS
0.4
1
2
4
10
20
40
100
Model Count per Contributor
BCM
HCI
MDACC
UC Davis
WISTAR
WUSTL

Enabled multiple high-impact publications

- ➔ Systematic Establishment of Robustness and Standards in Patient-Derived Xenograft Experiments and Analysis. *Cancer Research, March 2020*
- ➔ Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. *Nature Genetics, January 2021*

High impact publications on the CGC



nature communications

Explore our content ⌄ Journal information ⌄

nature › nature communications › articles › article

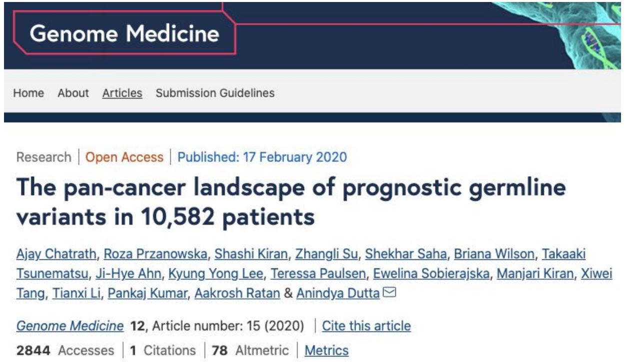
Article | Open Access | Published: 02 June 2020

AGO-bound mature miRNAs are oligouridylated by TUTs and subsequently degraded by DIS3L2

Acong Yang, Tie-Juan Shao, Xavier Bofill-De Ros, Chuanjiang Lian, Patricia Villanueva, Lisheng Dai & Shuo Gu ✉

Nature Communications **11**, Article number: 2765 (2020) | Cite this article

2767 Accesses | **1** Citations | **11** Altmetric | Metrics



Genome Medicine

Home About Articles Submission Guidelines

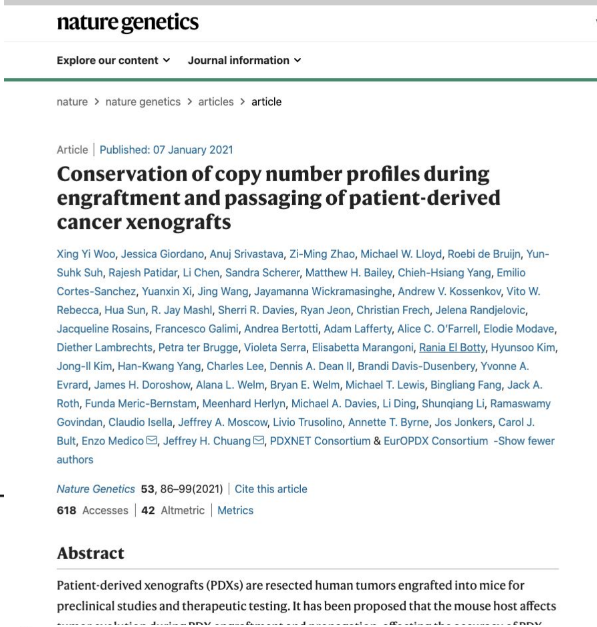
Research | Open Access | Published: 17 February 2020

The pan-cancer landscape of prognostic germline variants in 10,582 patients

Ajay Chatrath, Roza Przanowska, Shashi Kiran, Zhangli Su, Shekhar Saha, Briana Wilson, Takaaki Tsunematsu, Ji-Hye Ahn, Kyung Yong Lee, Teressa Paulsen, Ewelina Sobierajska, Manjari Kiran, Xiwei Tang, Tianxi Li, Pankaj Kumar, Aakrosh Ratan & Anindya Dutta ✉

Genome Medicine **12**, Article number: 15 (2020) | Cite this article

2844 Accesses | **1** Citations | **78** Altmetric | Metrics



nature genetics

Explore our content ⌄ Journal information ⌄

nature › nature genetics › articles › article

Article | Published: 07 January 2021

Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts

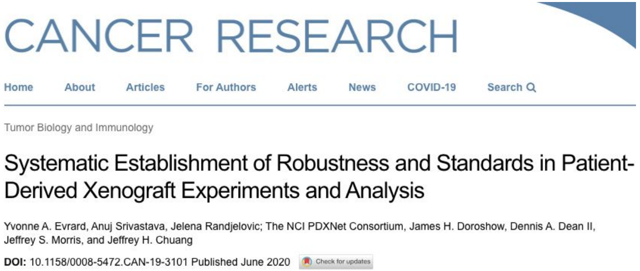
Xing Yi Woo, Jessica Giordano, Anuj Srivastava, Zi-Ming Zhao, Michael W. Lloyd, Roebi de Bruijn, Yun-Suhk Suh, Rajesh Patidar, Li Chen, Sandra Scherer, Matthew H. Bailey, Chieh-Hsiang Yang, Emilio Cortes-Sanchez, Yuanxin Xi, Jing Wang, Jayamanna Wickramasinghe, Andrew V. Kossenkov, Vito W. Rebecca, Hua Sun, R. Jay Mashl, Sherri R. Davies, Ryan Jeon, Christian Frech, Jelena Randjelovic, Jacqueline Rosains, Francesco Galimi, Andrea Bertotti, Adam Lafferty, Alice C. O'Farrell, Elodie Modave, Diether Lambrechts, Petra ter Brugge, Violeta Serra, Elisabetta Marangoni, Rania El Botty, Hyunsoo Kim, Jong-Il Kim, Han-Kwang Yang, Charles Lee, Dennis A. Dean II, Brandi Davis-Dusenbery, Yvonne A. Evrard, James H. Doroshow, Alana L. Welm, Bryan E. Welm, Michael T. Lewis, Bingliang Fang, Jack A. Roth, Funda Meric-Bernstam, Meenhard Herlyn, Michael A. Davies, Li Ding, Shunqiang Li, Ramaswamy Govindan, Claudio Isella, Jeffrey A. Moscow, Livio Trusolino, Annette T. Byrne, Jos Jonkers, Carol J. Bult, Enzo Medico ✉, Jeffrey H. Chuang ✉, PDXNET Consortium & EurOPDX Consortium -Show fewer authors

Nature Genetics **53**, 86–99(2021) | Cite this article

618 Accesses | **42** Altmetric | Metrics

Abstract

Patient-derived xenografts (PDXs) are resected human tumors engrafted into mice for preclinical studies and therapeutic testing. It has been proposed that the mouse host affects tumor evolution during PDX engraftment and propagation, affecting the accuracy of PDX



CANCER RESEARCH

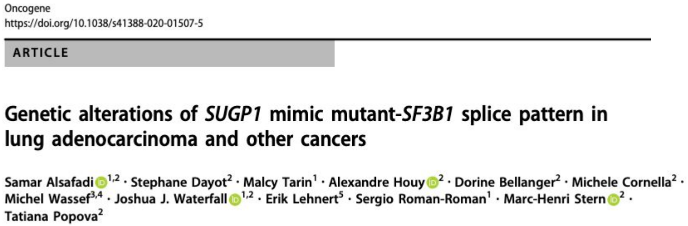
Home About Articles For Authors Alerts News COVID-19 Search

Tumor Biology and Immunology

Systematic Establishment of Robustness and Standards in Patient-Derived Xenograft Experiments and Analysis

Yvonne A. Evrard, Anuj Srivastava, Jelena Randjelovic; The NCI PDXNet Consortium, James H. Doroshow, Dennis A. Dean II, Jeffrey S. Morris, and Jeffrey H. Chuang

DOI: 10.1158/0008-5472.CAN-19-3101 Published June 2020 Check for updates



Oncogene
https://doi.org/10.1038/s41388-020-01507-5

ARTICLE

Genetic alterations of *SUGP1* mimic mutant-*SF3B1* splice pattern in lung adenocarcinoma and other cancers

Samar Alsafadi[1,2] · Stephane Dayot[2] · Malcy Tarin[1] · Alexandre Houy[2] · Dorine Bellanger[2] · Michele Cornella[2] · Michel Wassef[3,4] · Joshua J. Waterfall[1,2] · Erik Lehnert[5] · Sergio Roman-Roman[1] · Marc-Henri Stern[2] · Tatiana Popova[2]

Participating in open standards groups helps make us more FAIR

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

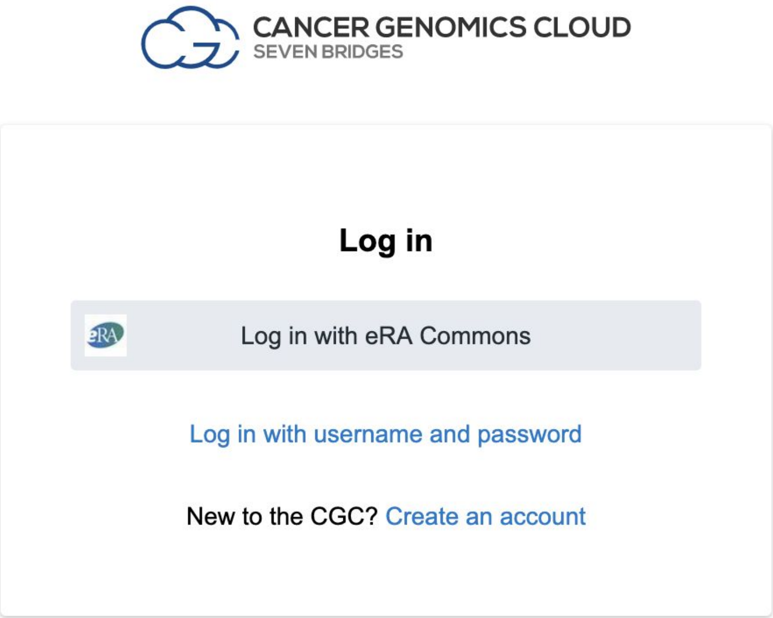
BioCompute
Objects

COMMON
WORKFLOW
LANGUAGE

docker

How do I get an account on the CGC?

- Sign up with your email
  - **https://www.cancergenomicscloud.org/**
- Option to connect with eRA Commons to access controlled data
- **$300 of pilot funding** to get your project started
- Comprehensive online documentation and training resources
- Technical support from a team of scientists, bioinformaticians, and engineers


CANCER GENOMICS CLOUD
SEVEN BRIDGES
Log in
eRA
Log in with eRA Commons
Log in with username and password
New to the CGC? Create an account

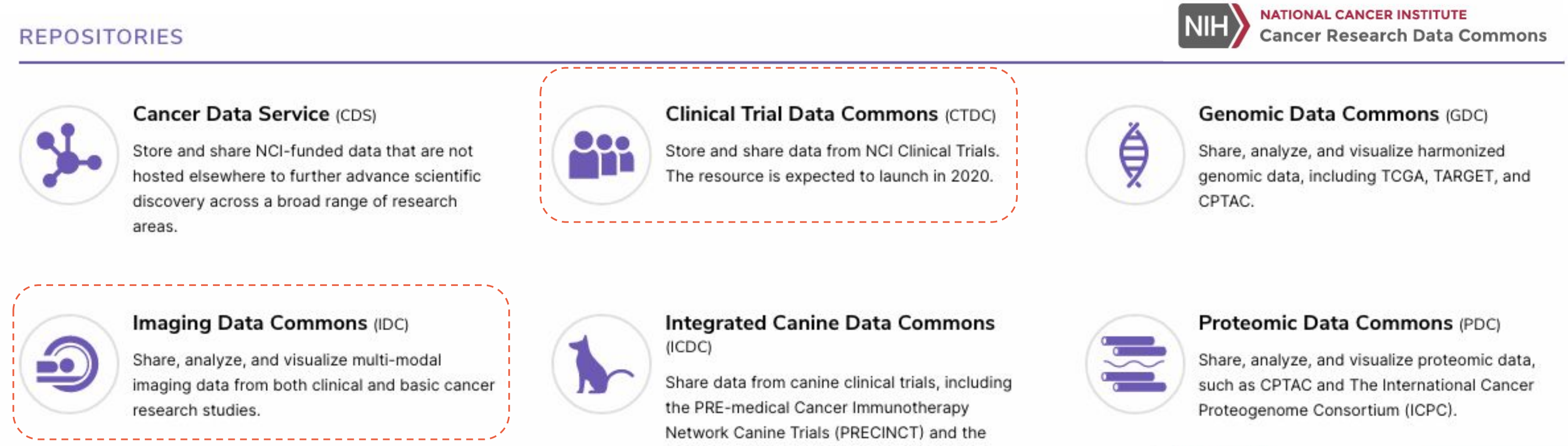
Access multi-omics datasets in the CGC

Access and search large public datasets on the CGC

Dataset	Description	Experimental setup	File types
TCGA	Rich dataset of tumor and normal tissues from 11,000 patients, covering 33 cancer types	WES, RNAseq, miRNAseq, methylation, genotyping, ATACseq, imaging, WGS, ..	BAM, VCF, MAF, TXT, TSV, SVS, XML
TARGET	Dataset of genomic changes in childhood cancers	RNASeq, WGS, WES, miRNAseq	BAM, MAF, TSV, VCF, XLSX, TXT
CANCER IMAGING ARCHIVE	Imaging data from many 21 tumor types	Imaging	DCM
CPTAC	Proteomics of 10 tumor types and associated genomic data	Proteomics, WGS, WES, RNAseq	BAM, TSV, VCF, mzML.gz, mzid.gz, raw, tar.gz
International Cancer Genome Consortium	Consortium of many datasets, 20 studies on CGC	WGS, RNASeq	BAM, VCF
CCLE Cancer Cell Line Encyclopedia	Dataset of 1457 cancer cell lines	WGS, WES, RNAseq	BAM
SIMONS FOUNDATION	Genome sequencing of 130 populations	WGS	BAM, VCF
Personal Genome Project	Crowdsourced genomics, datasets from 10 individuals	WGS, WGBS, RNAseq, methylation	BAM, FASTQ, IDAT, TBI, VCF
HUMAN CELL ATLAS	Single-cell genomics of healthy tissues	scRNASeq	FASTQ

CGC connects with several CRDC data repositories

Coming soon!



NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

REPOSITORIES

Cancer Data Service (CDS)

Store and share NCI-funded data that are not hosted elsewhere to further advance scientific discovery across a broad range of research areas.

Clinical Trial Data Commons (CTDC)

Store and share data from NCI Clinical Trials. The resource is expected to launch in 2020.

Genomic Data Commons (GDC)

Share, analyze, and visualize harmonized genomic data, including TCGA, TARGET, and CPTAC.

Imaging Data Commons (IDC)

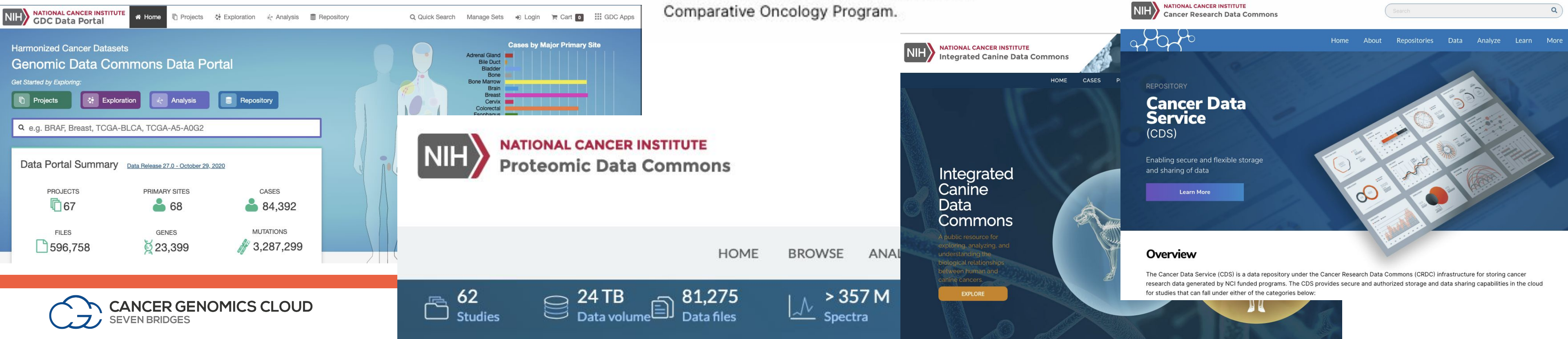
Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies.

Integrated Canine Data Commons (ICDC)

Share data from canine clinical trials, including the PRE-medical Cancer Immunotherapy Network Canine Trials (PRECINCT) and the Comparative Oncology Program.

Proteomic Data Commons (PDC)

Share, analyze, and visualize proteomic data, such as CPTAC and The International Cancer Proteogenome Consortium (ICPC).



NIH NATIONAL CANCER INSTITUTE
GDC Data Portal

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Data Portal Summary Data Release 27.0 - October 29, 2020

PROJECTS	PRIMARY SITES	CASES
67	68	84,392
FILES	GENES	MUTATIONS
596,758	23,399	3,287,299

NIH NATIONAL CANCER INSTITUTE
Proteomic Data Commons

62 Studies

24 TB Data volume

81,275 Data files

> 357 M Spectra

NIH NATIONAL CANCER INSTITUTE
Integrated Canine Data Commons

Integrated Canine Data Commons

A public resource for exploring, analyzing, and understanding the biological relationships between human and canine cancers.

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

REPOSITORY

Cancer Data Service (CDS)

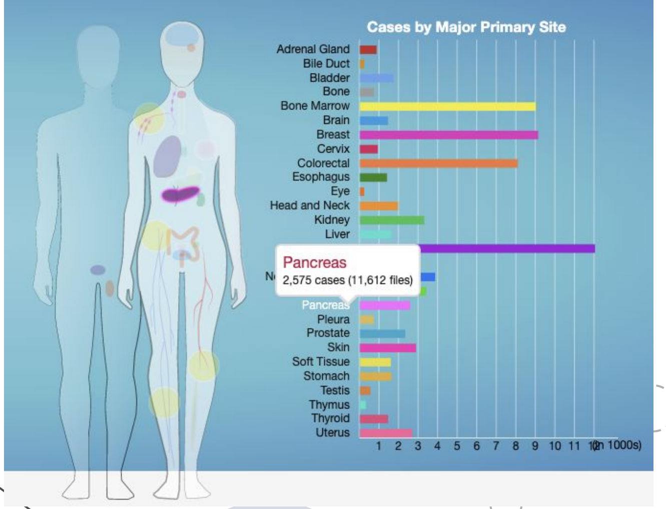
Enabling secure and flexible storage and sharing of data

Overview

The Cancer Data Service (CDS) is a data repository under the Cancer Research Data Commons (CRDC) infrastructure for storing cancer research data generated by NCI funded programs. The CDS provides secure and authorized storage and data sharing capabilities in the cloud for studies that can fall under either of the categories below:

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Use Case 1: Identifying mutational burden in cancer - Analysis of TCGA datasets



How do I identify the mutational burden (top mutated genes and somatic signature) in pancreatic adenocarcinoma cases?

I need access to different datasets, tools and workflows!

I need to share the results with my collaborators!

I am worried about security of analyzing my data in the cloud!

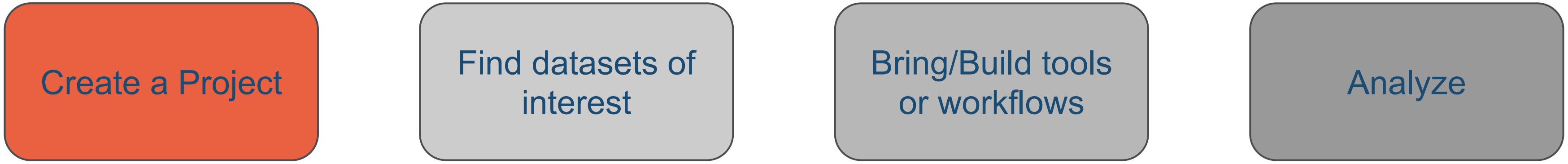
I am worried about cloud costs and billing!

I need to re-run my analysis using a different set of parameters!

My data is only available on XYZ cloud!

Image courtesy: https://portal.gdc.cancer.gov/

Typical User Flow



Create a Project

Organizational unit within the CGC

Find datasets of interest

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Bring/Build tools or workflows

Tools, workflows, and software packages

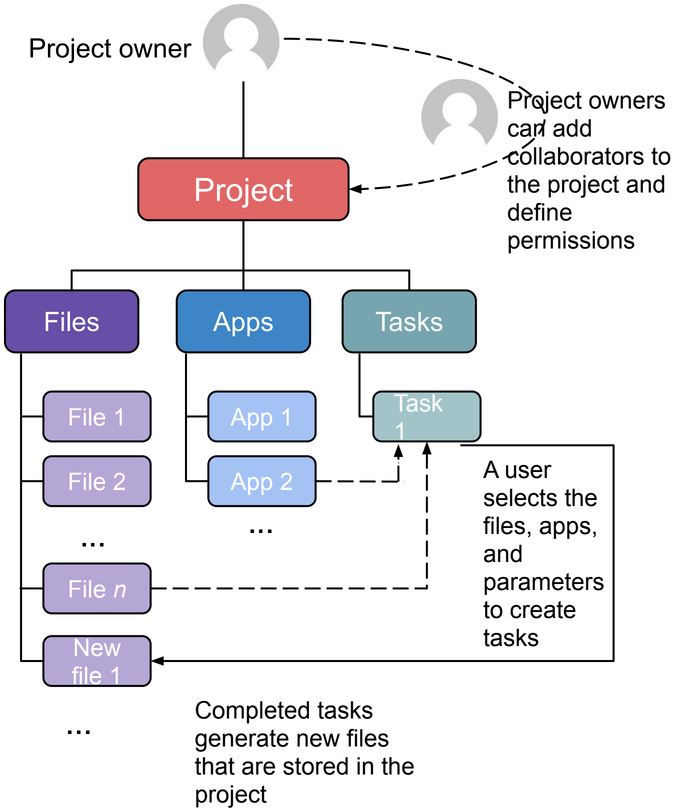
- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Analyze

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

Projects organize files, methods, and results



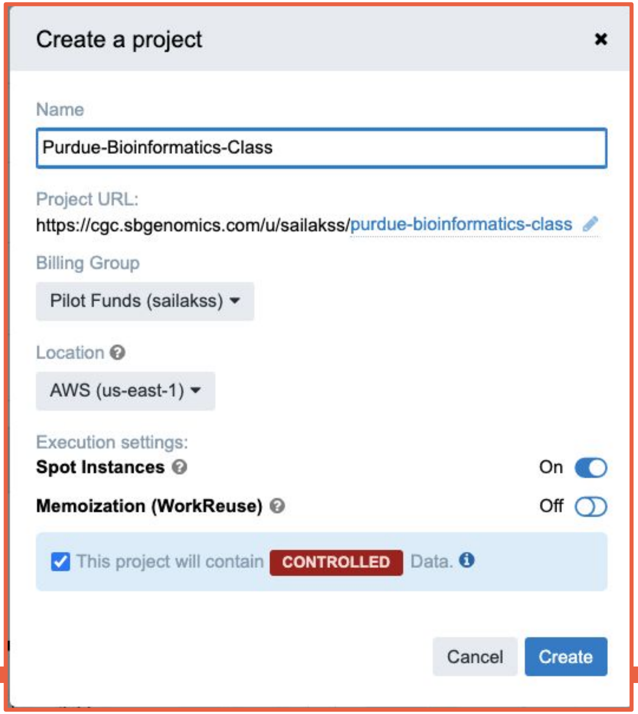
Project owners can add collaborators to the project and define permissions

A user selects the files, apps, and parameters to create tasks

Completed tasks generate new files that are stored in the project

Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions



Name: Purdue-Bioinformatics-Class

Project URL: https://cgc.sbgenomics.com/u/sailakss/purdue-bioinformatics-class

Billing Group: Pilot Funds (sailakss)

Location: AWS (us-east-1)

Execution settings:

Spot Instances On

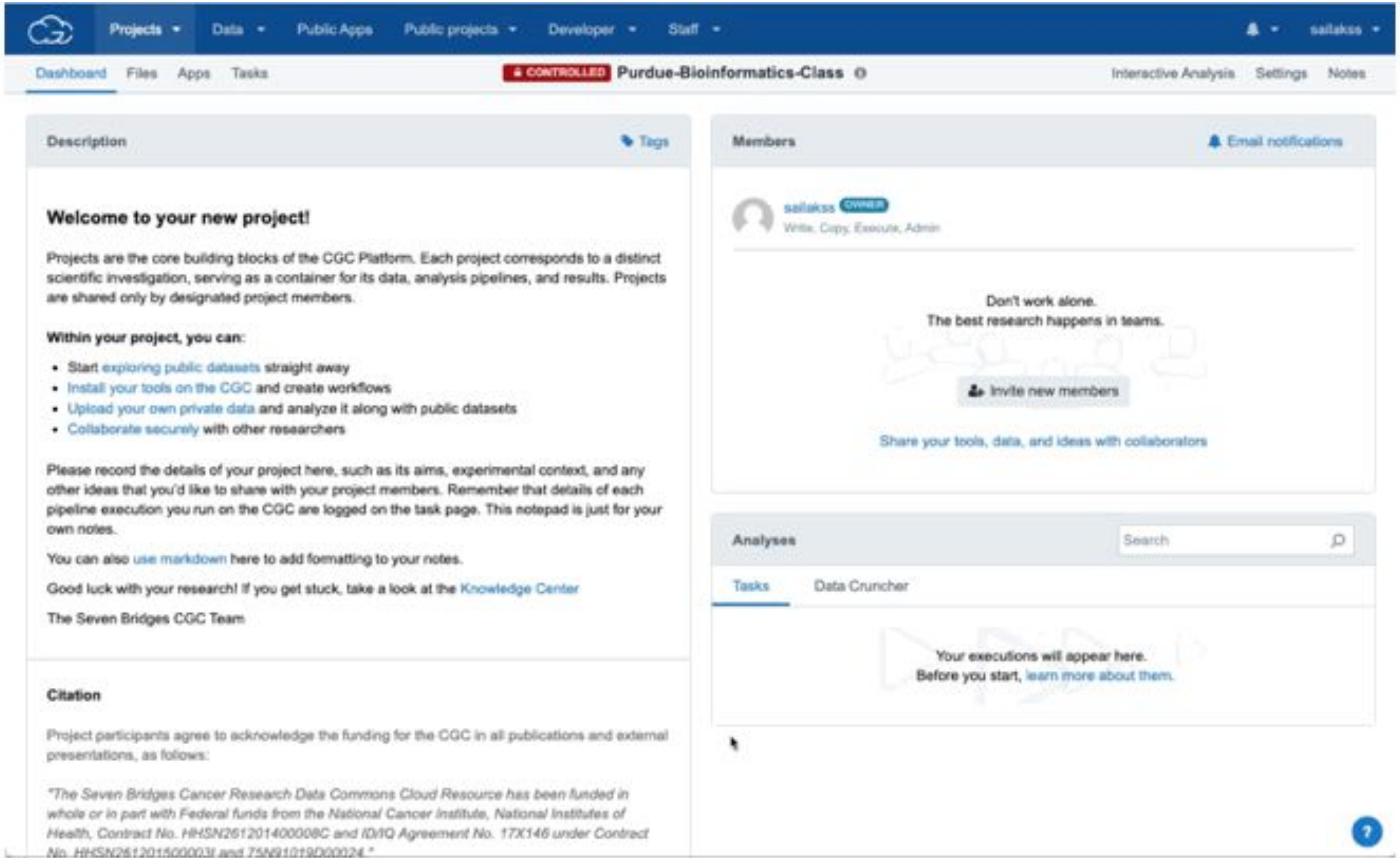
Memoization (WorkReuse) Off

☑ This project will contain CONTROLLED Data.

Projects are configurable, e.g.

- Customizable billing group - where costs should be attributed
- Cloud resources (AWS or GCP)
- **Spot** (or **preemptible**) instances
- Memoization - Intermediate file retention
- Using S3 or Glacier storage

Collaborate and share results quickly and easily



Purdue-Bioinformatics-Class

Description

Welcome to your new project!

Projects are the core building blocks of the CGC Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects are shared only by designated project members.

Within your project, you can:

- Start exploring public datasets straight away
- Install your tools on the CGC and create workflows
- Upload your own private data and analyze it along with public datasets
- Collaborate securely with other researchers

Please record the details of your project here, such as its aims, experimental context, and any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the CGC are logged on the task page. This notepad is just for your own notes.

You can also use markdown here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the Knowledge Center

The Seven Bridges CGC Team

Citation

Project participants agree to acknowledge the funding for the CGC in all publications and external presentations, as follows:

"The Seven Bridges Cancer Research Data Commons Cloud Resource has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I and 75N91019D00024."

Members

sailakss OWNER
Write, Copy, Execute, Admin

Don't work alone.
The best research happens in teams.

Share your tools, data, and ideas with collaborators

Analyses

Your executions will appear here.
Before you start, learn more about them.

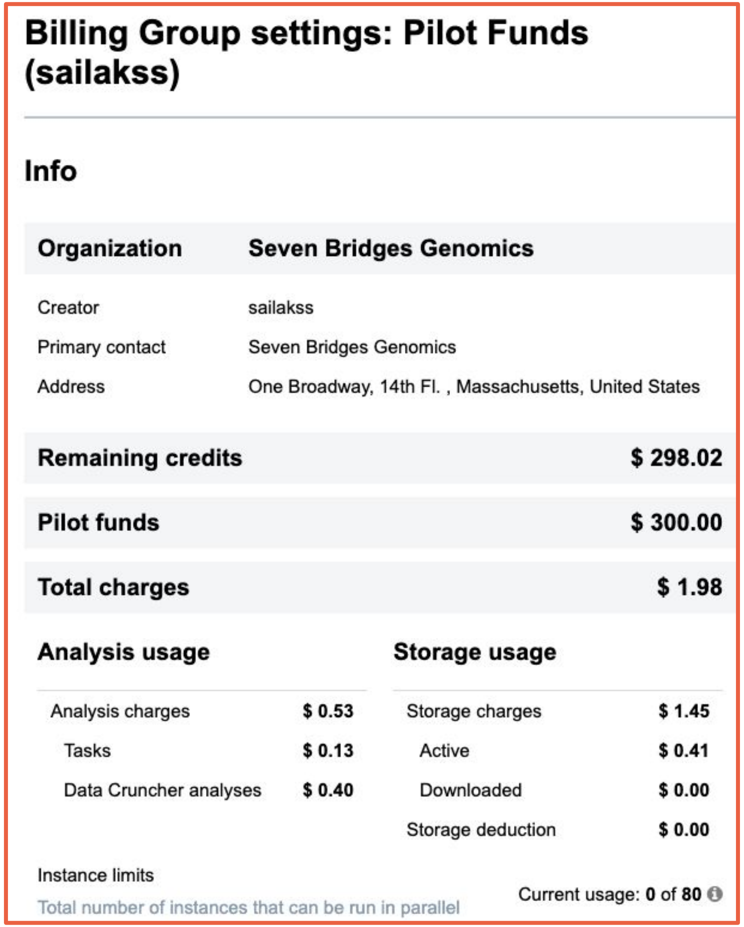
Billing groups

Clear advantages for collaboration and interoperability. Aligned to temporal dynamics of research funding.

Allow users to distribute costs appropriately per function, topic, lab, etc

Use different funding sources (e.g. R24, Pilot Funds, credit card)

SB can reimburse for task failure due to external factors



Billing Group settings: Pilot Funds (sailakss)

Info

Organization	**Seven Bridges Genomics**
Creator	sailakss
Primary contact	Seven Bridges Genomics
Address	One Broadway, 14th Fl. , Massachusetts, United States

Remaining credits	**$ 298.02**
Pilot funds	**$ 300.00**
Total charges	**$ 1.98**

Analysis usage

Analysis charges	$ 0.53
Tasks	$ 0.13
Data Cruncher analyses	$ 0.40

Storage usage

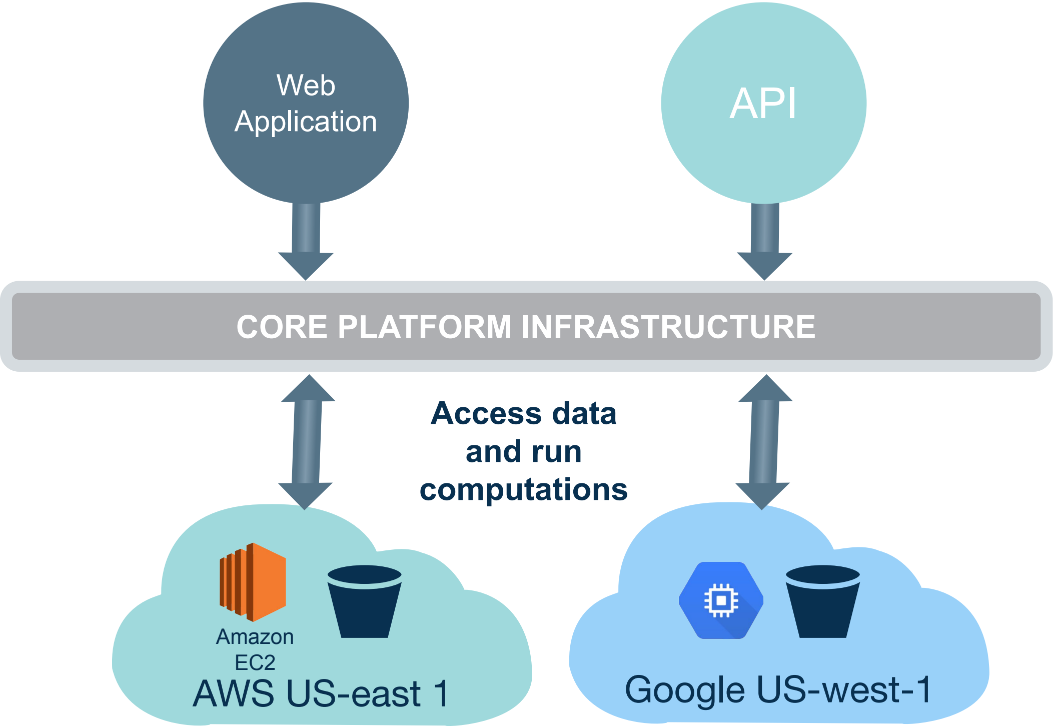
Storage charges	$ 1.45
Active	$ 0.41
Downloaded	$ 0.00
Storage deduction	$ 0.00

Instance limits

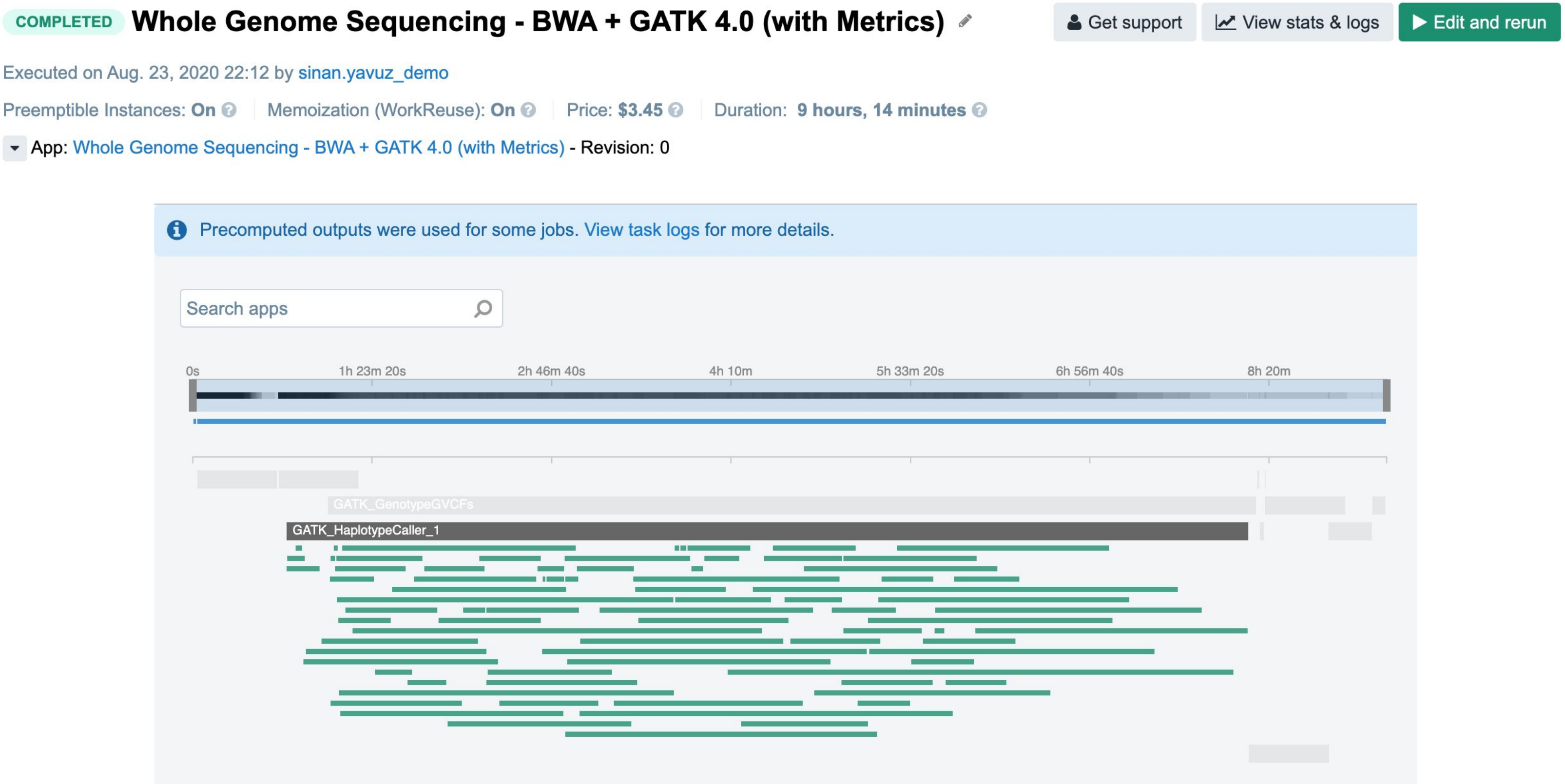
Total number of instances that can be run in parallel

Current usage: 0 of 80

Multi-cloud implementation on the CGC


Web Application
API
CORE PLATFORM INFRASTRUCTURE
Access data and run computations
Amazon EC2
AWS US-east 1
Google US-west-1

Memoization allows use of previously computed results



COMPLETED **Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)**

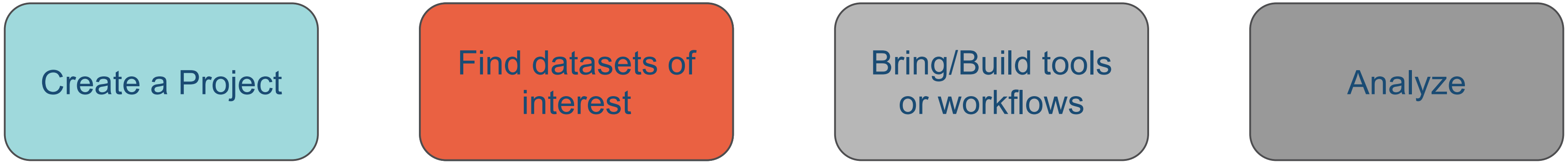
Executed on Aug. 23, 2020 22:12 by sinan.yavuz_demo

Preemptible Instances: **On** | Memoization (WorkReuse): **On** | Price: **$3.45** | Duration: **9 hours, 14 minutes**

App: Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) - Revision: 0

Precomputed outputs were used for some jobs. View task logs for more details.

User Flow



Create a Project

Find datasets of interest

Bring/Build tools or workflows

Analyze

Organizational unit within the CGC

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

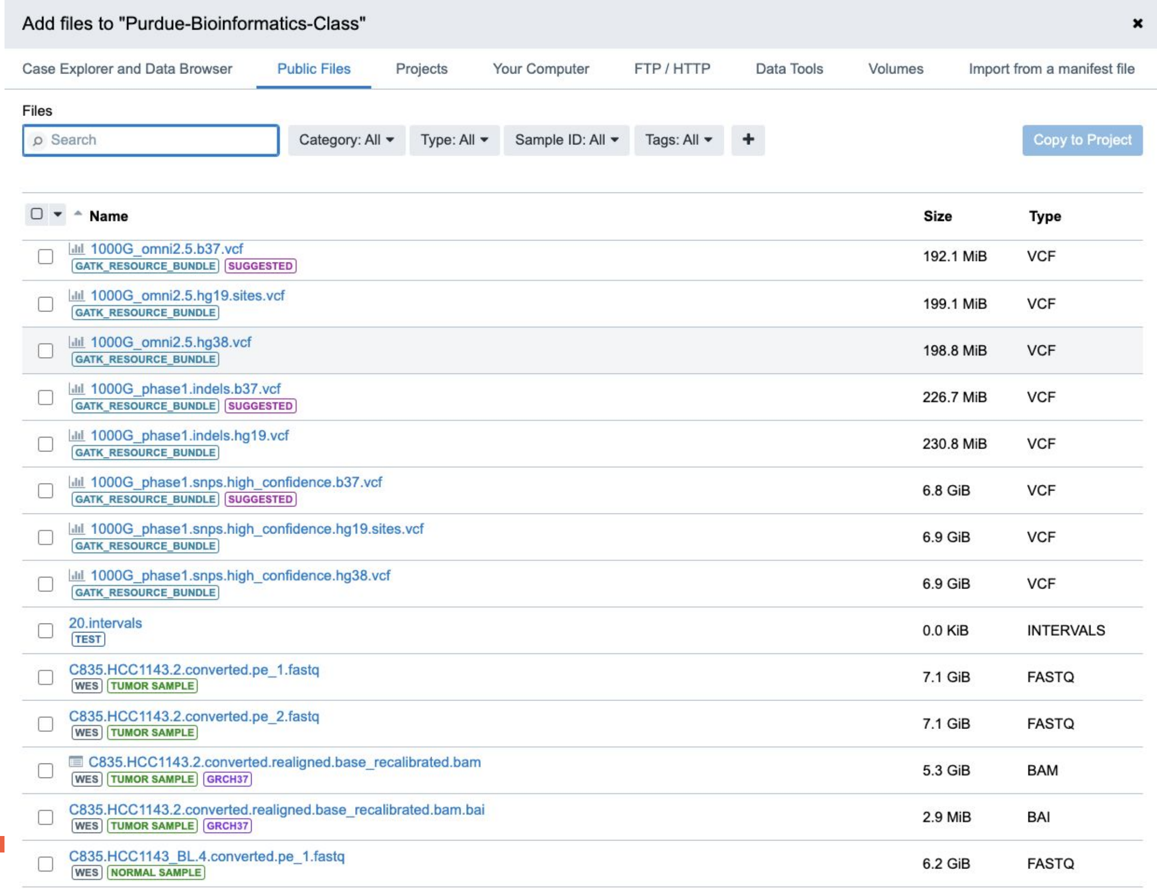
Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

Different options to bring data



Add files to "Purdue-Bioinformatics-Class" ✖

Case Explorer and Data Browser | Public Files | Projects | Your Computer | FTP / HTTP | Data Tools | Volumes | Import from a manifest file

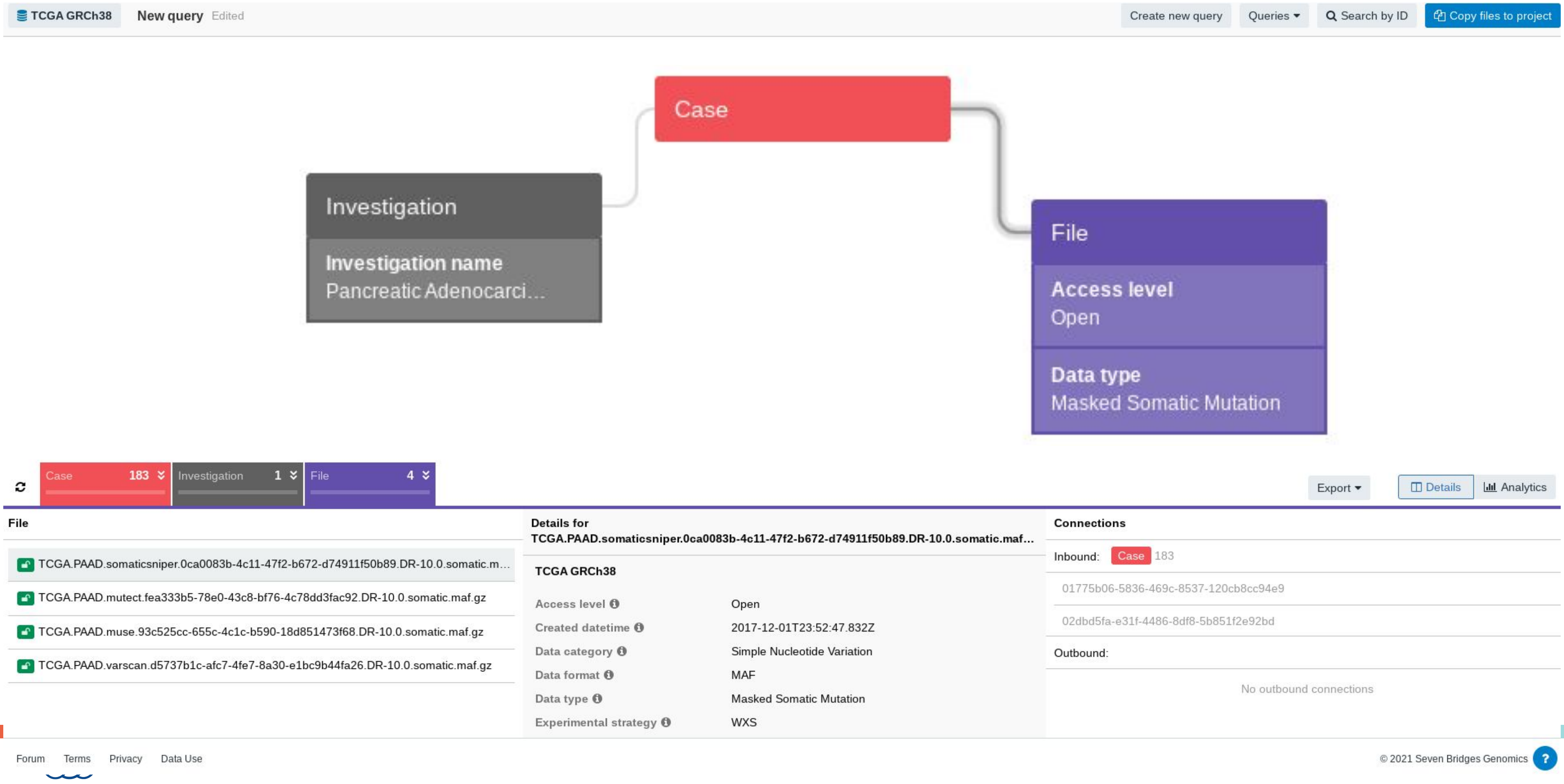
Files

Search | Category: All ▾ | Type: All ▾ | Sample ID: All ▾ | Tags: All ▾ | ✚ | Copy to Project

□	Name	Size	Type
□	1000G_omni2.5.b37.vcf GATK_RESOURCE_BUNDLE SUGGESTED	192.1 MiB	VCF
□	1000G_omni2.5.hg19.sites.vcf GATK_RESOURCE_BUNDLE	199.1 MiB	VCF
□	1000G_omni2.5.hg38.vcf GATK_RESOURCE_BUNDLE	198.8 MiB	VCF
□	1000G_phase1.indels.b37.vcf GATK_RESOURCE_BUNDLE SUGGESTED	226.7 MiB	VCF
□	1000G_phase1.indels.hg19.vcf GATK_RESOURCE_BUNDLE	230.8 MiB	VCF
□	1000G_phase1.snps.high_confidence.b37.vcf GATK_RESOURCE_BUNDLE SUGGESTED	6.8 GiB	VCF
□	1000G_phase1.snps.high_confidence.hg19.sites.vcf GATK_RESOURCE_BUNDLE	6.9 GiB	VCF
□	1000G_phase1.snps.high_confidence.hg38.vcf GATK_RESOURCE_BUNDLE	6.9 GiB	VCF
□	20.intervals TEST	0.0 KiB	INTERVALS
□	C835.HCC1143.2.converted.pe_1.fastq WES TUMOR SAMPLE	7.1 GiB	FASTQ
□	C835.HCC1143.2.converted.pe_2.fastq WES TUMOR SAMPLE	7.1 GiB	FASTQ
□	C835.HCC1143.2.converted.realigned.base_recalibrated.bam WES TUMOR SAMPLE GRCH37	5.3 GiB	BAM
□	C835.HCC1143.2.converted.realigned.base_recalibrated.bam.bai WES TUMOR SAMPLE GRCH37	2.9 MiB	BAI
□	C835.HCC1143_BL.4.converted.pe_1.fastq WES NORMAL SAMPLE	6.2 GiB	FASTQ

- * Public files
- * Case Explorer & Data Browser
- * Projects (that you are a member of)
- * FTP/HTTP (signed URLs)
- * Data tools
  - Command Line Uploader
  - Desktop Uploader
  - SBFS: Seven Bridges File System
  - API upload
- * Volumes
- * Import from manifest: ICDC/PDC

Find open access TCGA data with Data Browser



TCGA GRCh38 | **New query** Edited

Create new query | Queries ▾ | Search by ID | Copy files to project

Case

Investigation

Investigation name
Pancreatic Adenocarci…

File

Access level
Open

Data type
Masked Somatic Mutation

Case 183 | Investigation 1 | File 4

Export ▾ | Details | Analytics

File

- TCGA.PAAD.somaticsniper.0ca0083b-4c11-47f2-b672-d74911f50b89.DR-10.0.somatic.m…
- TCGA.PAAD.mutect.fea333b5-78e0-43c8-bf76-4c78dd3fac92.DR-10.0.somatic.maf.gz
- TCGA.PAAD.muse.93c525cc-655c-4c1c-b590-18d851473f68.DR-10.0.somatic.maf.gz
- TCGA.PAAD.varscan.d5737b1c-afc7-4fe7-8a30-e1bc9b44fa26.DR-10.0.somatic.maf.gz

Details for
TCGA.PAAD.somaticsniper.0ca0083b-4c11-47f2-b672-d74911f50b89.DR-10.0.somatic.maf…

TCGA GRCh38

Access level ℹ	Open
Created datetime ℹ	2017-12-01T23:52:47.832Z
Data category ℹ	Simple Nucleotide Variation
Data format ℹ	MAF
Data type ℹ	Masked Somatic Mutation
Experimental strategy ℹ	WXS

Connections

Inbound: Case 183

01775b06-5836-469c-8537-120cb8cc94e9

02dbd5fa-e31f-4486-8df8-5b851f2e92bd

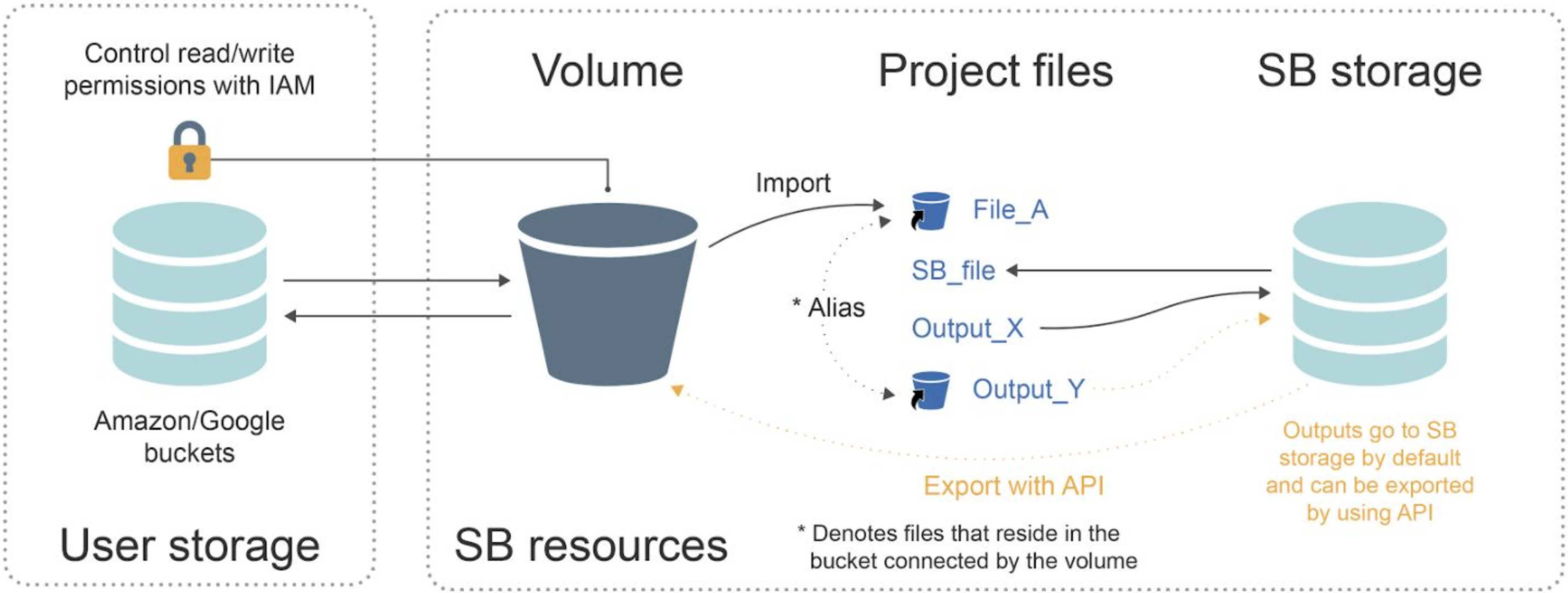
Outbound:

No outbound connections

Forum Terms Privacy Data Use

© 2021 Seven Bridges Genomics

Easily connect cloud volumes



Control read/write permissions with IAM

Amazon/Google buckets

User storage

Volume

Project files

SB storage

Import

File_A

SB_file

* Alias

Output_X

Output_Y

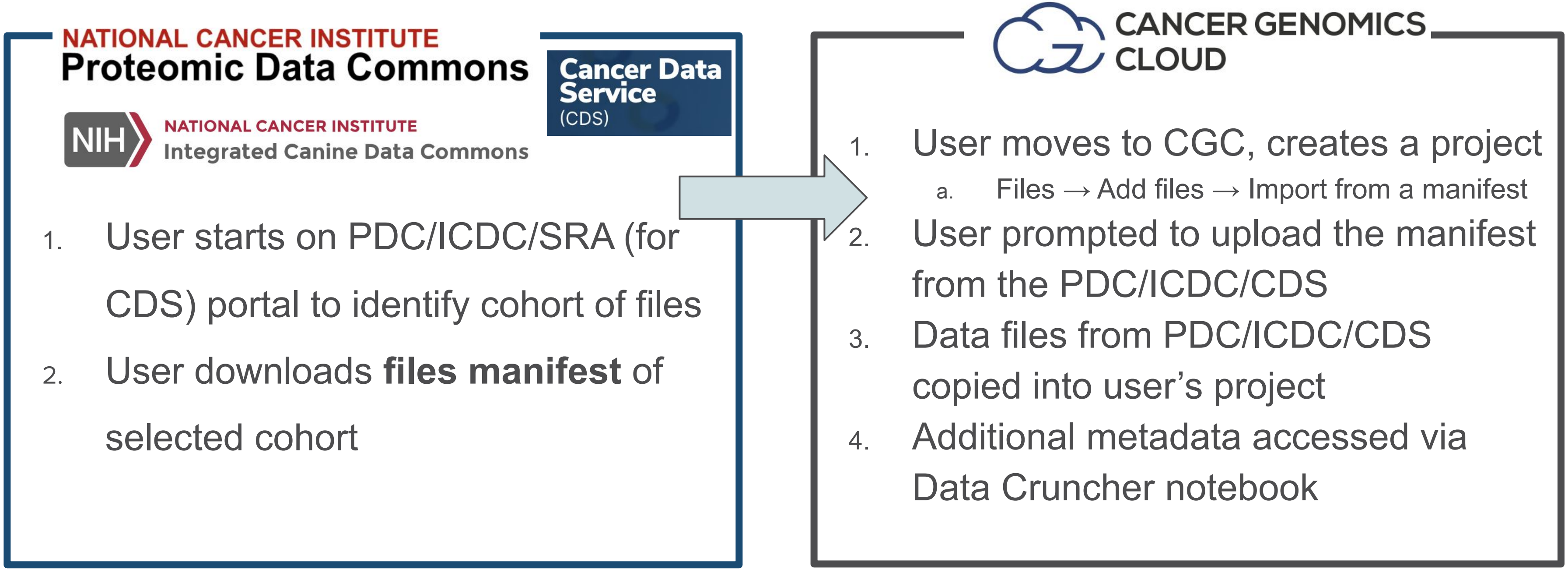
Export with API

Outputs go to SB storage by default and can be exported by using API

SB resources

* Denotes files that reside in the bucket connected by the volume

Enabling multi-omic research on the CGC through integrating with the PDC, ICDC, CDS



NATIONAL CANCER INSTITUTE
Proteomic Data Commons

Cancer Data Service
(CDS)

NIH
NATIONAL CANCER INSTITUTE
Integrated Canine Data Commons

1. User starts on PDC/ICDC/SRA (for CDS) portal to identify cohort of files
2. User downloads **files manifest** of selected cohort

CANCER GENOMICS CLOUD

1. User moves to CGC, creates a project
   a. Files → Add files → Import from a manifest
2. User prompted to upload the manifest from the PDC/ICDC/CDS
3. Data files from PDC/ICDC/CDS copied into user's project
4. Additional metadata accessed via Data Cruncher notebook

Links to doc pages to import data from: PDC, ICDC, CDS

User Flow

Create a Project

Organizational unit within the CGC

Find datasets of interest

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Bring/Build tools or workflows

Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Analyze

Specify how an analysis will be run

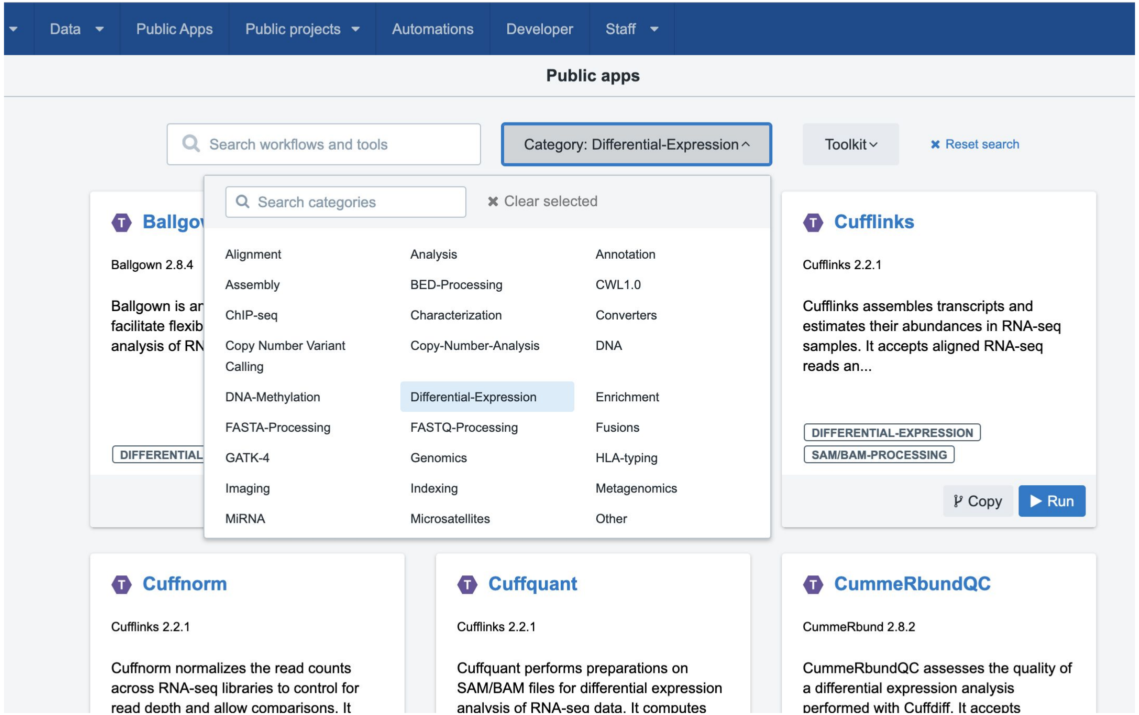
- Task page
- Notebooks in RStudio or JupyterLab

https://cgc.sbgenomics.com/public/apps

Find the tools you need in the Public Apps Gallery

A curated collection of ~**500** bioinformatics tools & workflows

- ○ Optimized for speed & cost in the cloud
- ○ Fully parameterized & customizable
- ○ Accessible via the GUI & API



Alignment, Analysis, Annotation, Assembly, BED-Processing, CWL1.0, ChIP-seq, Characterization, Converters, Copy Number Variant Calling, Copy-Number-Analysis, DNA, DNA-Methylation, Differential-Expression, Enrichment, FASTA-Processing, FASTQ-Processing, Fusions, GATK-4, Genomics, HLA-typing, Imaging, Indexing, Metagenomics, MiRNA, Microsatellites, Other

T Ballgo

Ballgown 2.8.4

Ballgown is ar facilitate flexib analysis of RN

DIFFERENTIAL

T Cufflinks

Cufflinks 2.2.1

Cufflinks assembles transcripts and estimates their abundances in RNA-seq samples. It accepts aligned RNA-seq reads an...

DIFFERENTIAL-EXPRESSION

SAM/BAM-PROCESSING

T Cuffnorm

Cufflinks 2.2.1

Cuffnorm normalizes the read counts across RNA-seq libraries to control for read depth and allow comparisons. It

T Cuffquant

Cufflinks 2.2.1

Cuffquant performs preparations on SAM/BAM files for differential expression analysis of RNA-seq data. It computes

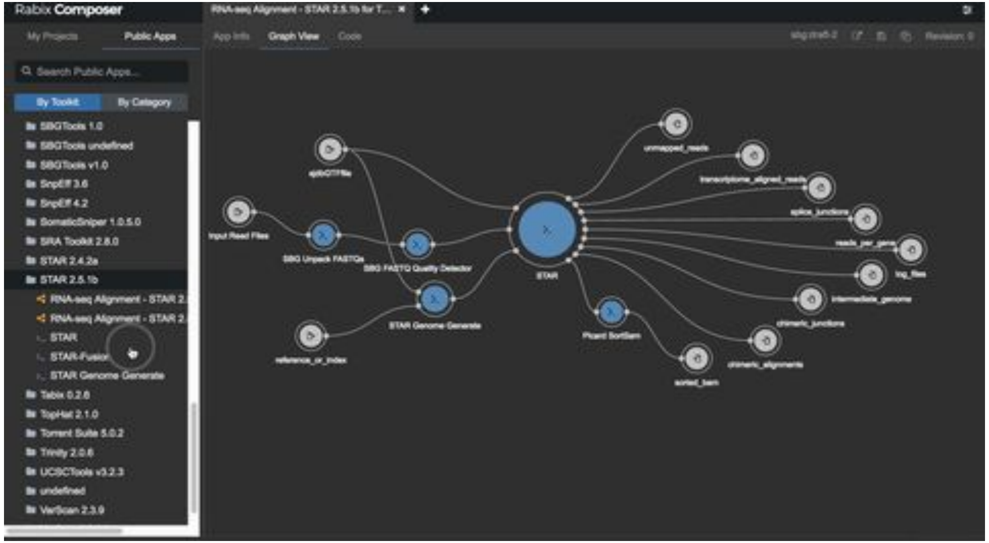
T CummeRbundQC

CummeRbund 2.8.2

CummeRbundQC assesses the quality of a differential expression analysis performed with Cuffdiff. It accepts

Bring Your Pipelines to the Platform with Web Composer

- An intuitive and flexible software development kit for developing and porting custom tools to the platform
- Conformance with community standards to ensure pipeline portability & reproducibility



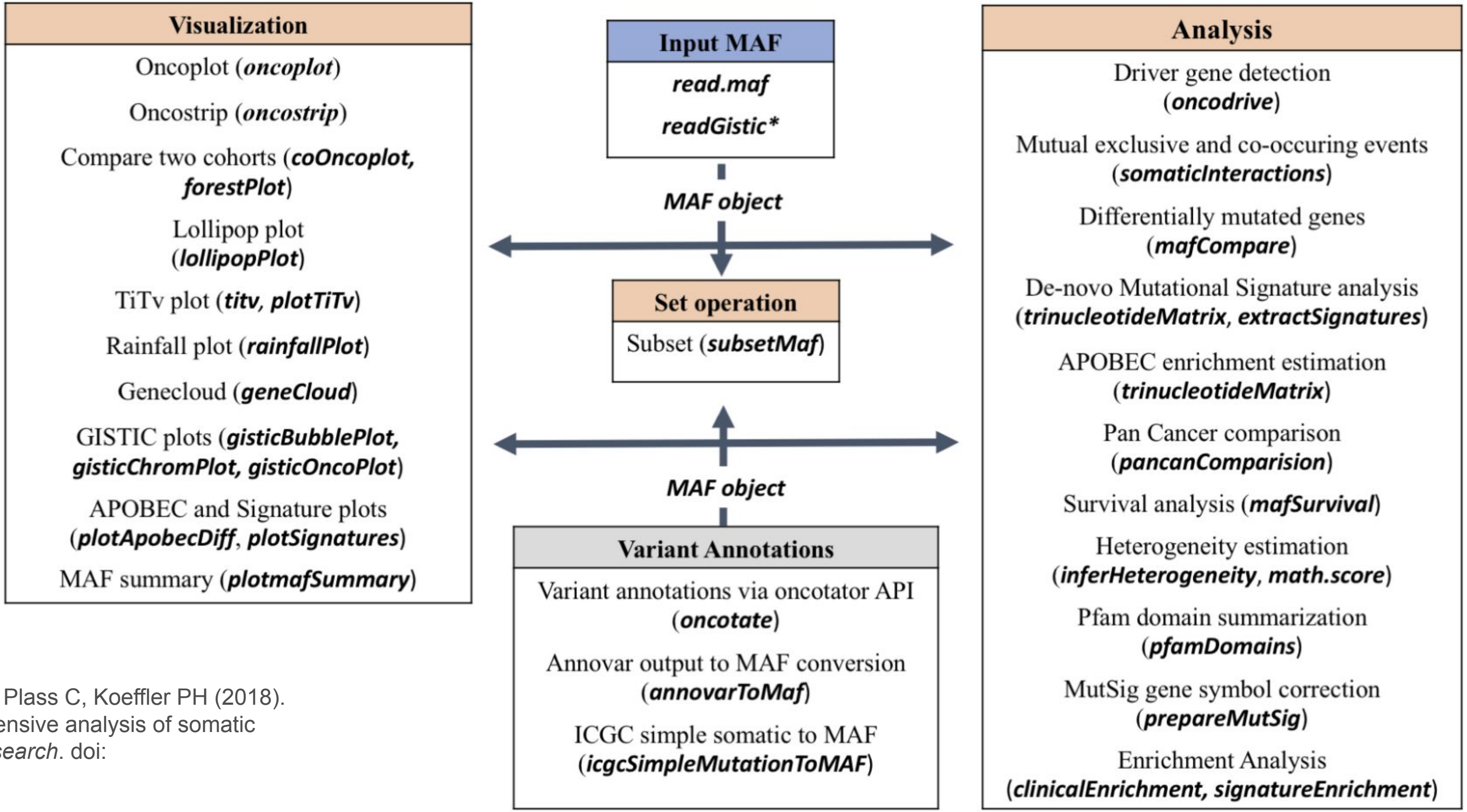
docker

COMMON WORKFLOW LANGUAGE

Rabix

[Reproducible Analysis for Bioinformatics] https://bioconductor.org/packages/release/bioc/html/maftools.html

Mutation Annotation Format (*maftools*)



Visualization

Oncoplot (***oncoplot***)

Oncostrip (***oncostrip***)

Compare two cohorts (***coOncoplot, forestPlot***)

Lollipop plot (***lollipopPlot***)

TiTv plot (***titv, plotTiTv***)

Rainfall plot (***rainfallPlot***)

Genecloud (***geneCloud***)

GISTIC plots (***gisticBubblePlot, gisticChromPlot, gisticOncoPlot***)

APOBEC and Signature plots (***plotApobecDiff*, *plotSignatures***)

MAF summary (***plotmafSummary***)

Input MAF

read.maf

readGistic*

MAF object

Set operation

Subset (***subsetMaf***)

MAF object

Variant Annotations

Variant annotations via oncotator API (***oncotate***)

Annovar output to MAF conversion (***annovarToMaf***)

ICGC simple somatic to MAF (***icgcSimpleMutationToMAF***)

Analysis

Driver gene detection (***oncodrive***)

Mutual exclusive and co-occuring events (***somaticInteractions***)

Differentially mutated genes (***mafCompare***)

De-novo Mutational Signature analysis (***trinucleotideMatrix*, *extractSignatures***)

APOBEC enrichment estimation (***trinucleotideMatrix***)

Pan Cancer comparison (***pancanComparision***)

Survival analysis (***mafSurvival***)

Heterogeneity estimation (***inferHeterogeneity*, *math.score***)

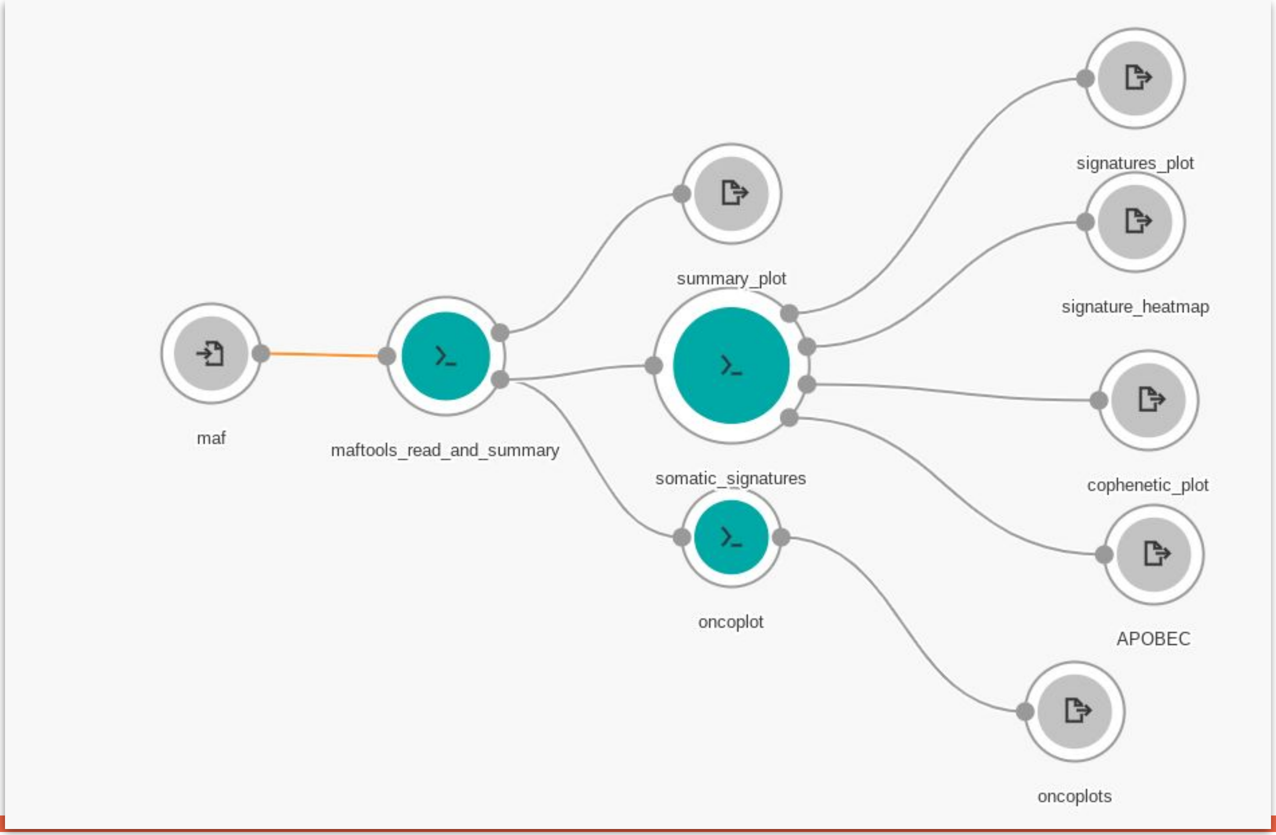
Pfam domain summarization (***pfamDomains***)

MutSig gene symbol correction (***prepareMutSig***)

Enrichment Analysis (***clinicalEnrichment, signatureEnrichment***)

Mayakonda A, Lin D, Assenov Y, Plass C, Koeffler PH (2018). “Maftools: efficient and comprehensive analysis of somatic variants in cancer.” *Genome Research*. doi: 10.1101/gr.239244.118.

Maftools Workflow



1. Dockerize individual tools
2. Wrap each tool in CWL
3. Connect tools into a workflow and set parameters

User Flow

Create a Project

Organizational unit within the CGC

Find datasets of interest

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Bring/Build tools or workflows

Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

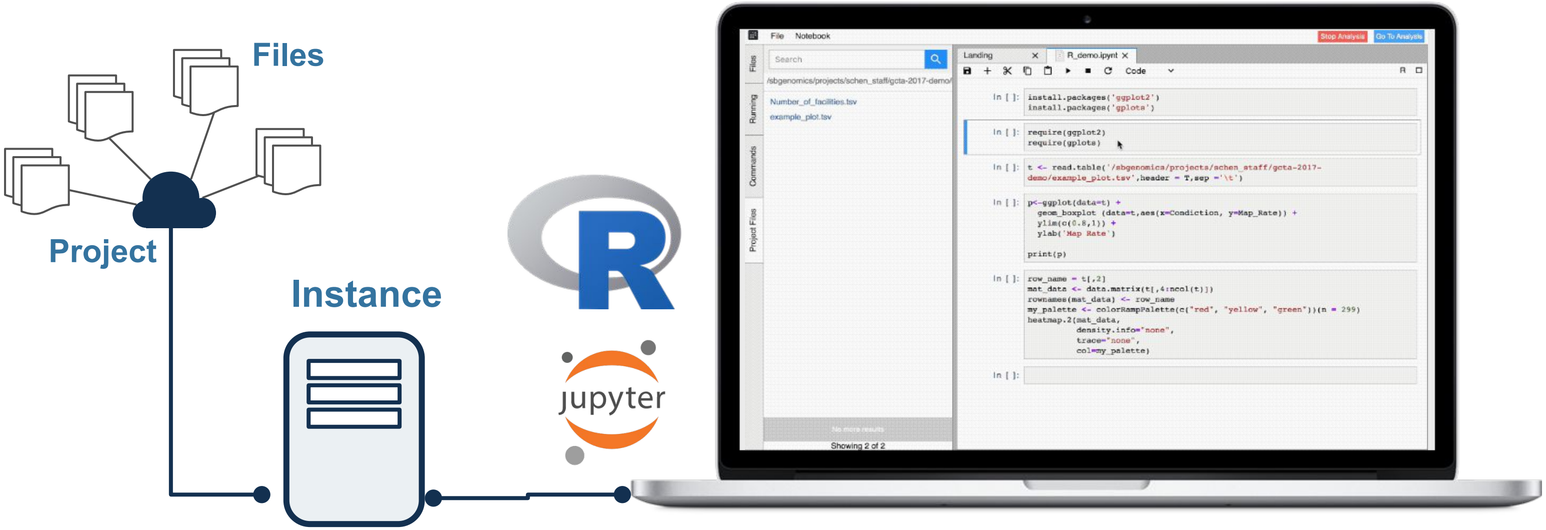
Analyze

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

Powerful, collaborative, & reproducible interactive analysis

Users create interactive analysis sessions within a project - all files are available and over 50 instances can be used (*c3.xlarge* to *x1.32xlarge* on AWS)



/sbgenomics/projects/schen_staff/gcta-2017-demo/

Number_of_facilities.tsv
example_plot.tsv

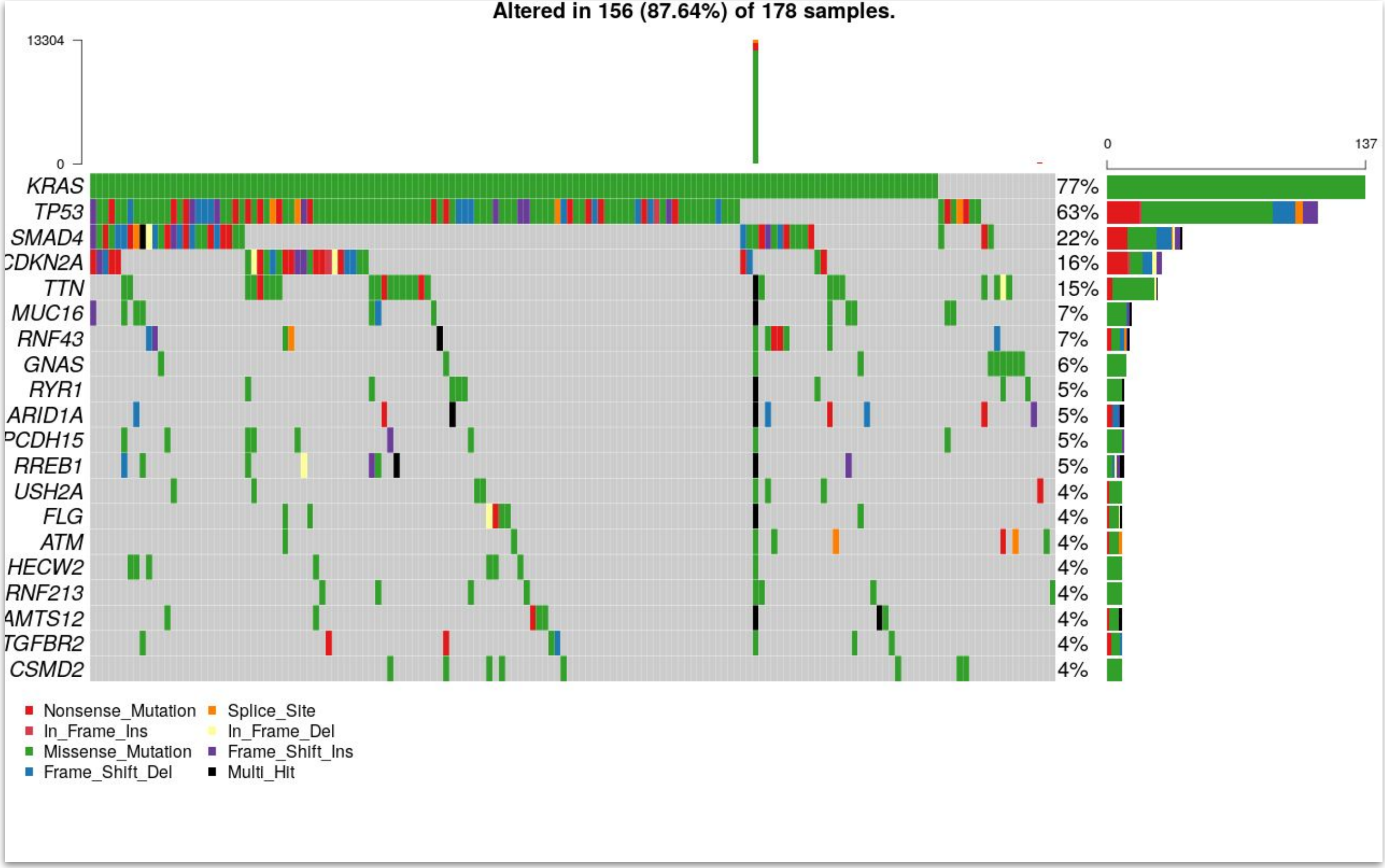
No more results
Showing 2 of 2

```
In [ ]: install.packages('ggplot2')
        install.packages('gplots')

In [ ]: require(ggplot2)
        require(gplots)

In [ ]: t <- read.table('/sbgenomics/projects/schen_staff/gcta-2017-
        demo/example_plot.tsv',header = T,sep ='\t')

In [ ]: p<-ggplot(data=t) +
          geom_boxplot (data=t,aes(x=Condiction, y=Map_Rate)) +
          ylim(c(0.8,1)) +
          ylab('Map Rate')

        print(p)

In [ ]: row_name = t[,2]
        mat_data <- data.matrix(t[,4:ncol(t)])
        rownames(mat_data) <- row_name
        my_palette <- colorRampPalette(c("red", "yellow", "green"))(n = 299)
        heatmap.2(mat_data,
                  density.info="none",
                  trace="none",
                  col=my_palette)

In [ ]:
```

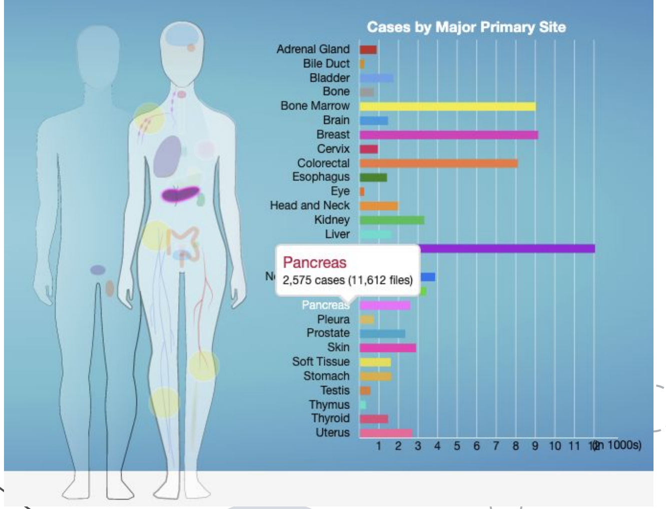
PAAD Oncoplot

Top 3 mutated genes

- KRAS
- TP53
- SMAD4



How do I identify the mutational burden (top mutated genes and somatic signature) in pancreatic adenocarcinoma cases?



I need access to different datasets, tools and workflows!

I need to share the results with my collaborators!

I am worried about security of analyzing my data in the cloud!

I am worried about cloud costs and billing!

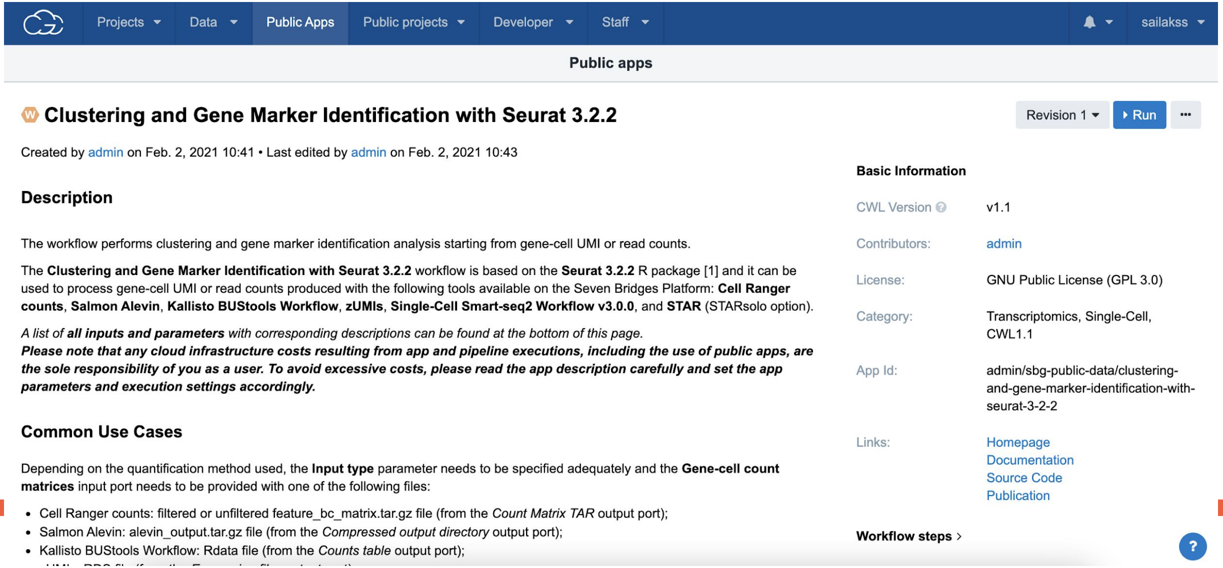
I need to re-run my analysis using a different set of parameters!

My data is only available on XYZ cloud!

Image courtesy: https://portal.gdc.cancer.gov/

Want to learn more?

- Learn how to perform cloud based loading of single cell data, quality control, normalization, PCA and clustering and biomarker identification.
- Using open data
- The workflow makes tables and an html report
- Also learn to use cloud based RStudio to dive deeper into the data



Clustering and Gene Marker Identification with Seurat 3.2.2

Created by admin on Feb. 2, 2021 10:41 • Last edited by admin on Feb. 2, 2021 10:43

Description

The workflow performs clustering and gene marker identification analysis starting from gene-cell UMI or read counts.

The **Clustering and Gene Marker Identification with Seurat 3.2.2** workflow is based on the **Seurat 3.2.2** R package [1] and it can be used to process gene-cell UMI or read counts produced with the following tools available on the Seven Bridges Platform: **Cell Ranger counts**, **Salmon Alevin**, **Kallisto BUStools Workflow**, **zUMIs**, **Single-Cell Smart-seq2 Workflow v3.0.0**, and **STAR** (STARsolo option).

*A list of **all inputs and parameters** with corresponding descriptions can be found at the bottom of this page.*
Please note that any cloud infrastructure costs resulting from app and pipeline executions, including the use of public apps, are the sole responsibility of you as a user. To avoid excessive costs, please read the app description carefully and set the app parameters and execution settings accordingly.

Common Use Cases

Depending on the quantification method used, the **Input type** parameter needs to be specified adequately and the **Gene-cell count matrices** input port needs to be provided with one of the following files:

- Cell Ranger counts: filtered or unfiltered feature_bc_matrix.tar.gz file (from the *Count Matrix TAR* output port);
- Salmon Alevin: alevin_output.tar.gz file (from the *Compressed output directory* output port);
- Kallisto BUStools Workflow: Rdata file (from the *Counts table* output port);

Basic Information

CWL Version	v1.1
Contributors:	admin
License:	GNU Public License (GPL 3.0)
Category:	Transcriptomics, Single-Cell, CWL1.1
App Id:	admin/sbg-public-data/clustering-and-gene-marker-identification-with-seurat-3-2-2
Links:	Homepage Documentation Source Code Publication

Workflow steps ›

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Hands-on Demo on 04/08 (Thursday)

Use Case 2: microRNA biogenesis in cancer

Using the CGC to understand microRNA biogenesis in cancer

Collaborative Project program to advance your research

- Submit a proposal for up to **$10,000** in cloud credits to cgc@sevenbridges.com
- Get additional access to our CGC team and bioinformatics support
- Projects have resulted in dozens of papers, many users submit multiple papers from one project
- We encourage applications from students and postdocs

Xavier Bofill de Ros - Research Fellow
Gu Lab, Center for Cancer Research NCI

NIH NATIONAL CANCER INSTITUTE
Center for Cancer Research

2017
2018
2019
2020

2017

- **I Collaborative application**
- Joint efforts on QuagmiR development
- TCGA data analysis with QuagmiR

2018

- QuagmiR publication
- QuagmiR tool available on CGC

Bioinformatics

2019

- **II Collaborative application**
- Subsequent research publication

Cell Reports

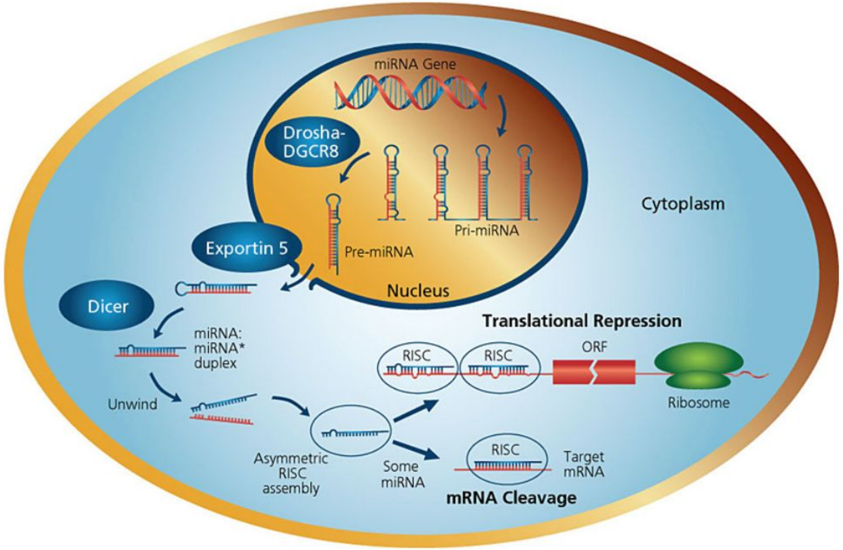
2020

- Third research publication
- Currently working with Case Explorer and other

nature COMMUNICATIONS

Case study - microRNAs & isomiRs

- microRNAs regulate gene expression
- Isoforms of miRNA (isomiRs) are correlated with cancer progression
- isomiRs very difficult to study because they are so heterogeneous



miRNA Pathway

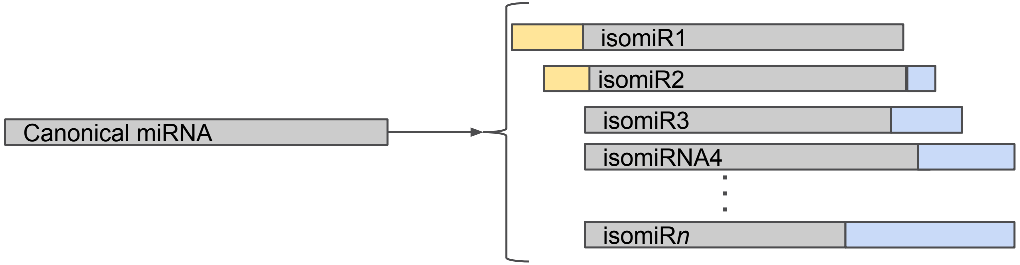
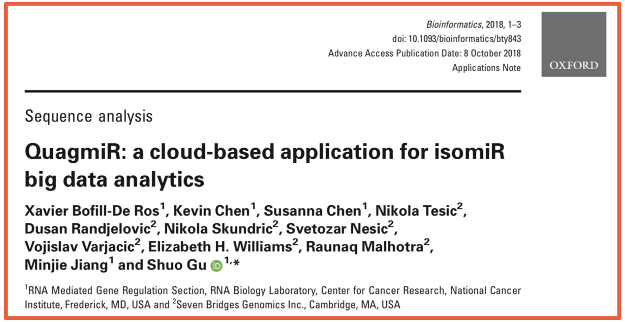


Image Courtesy: https://www.sigmaaldrich.com/life-science/functional-genomics-and-rnai/mirna/learning-center/mirna-introduction.html

Case study: QuagmiR



Bioinformatics, 2018, 1–3
doi: 10.1093/bioinformatics/bty843
Advance Access Publication Date: 8 October 2018
Applications Note

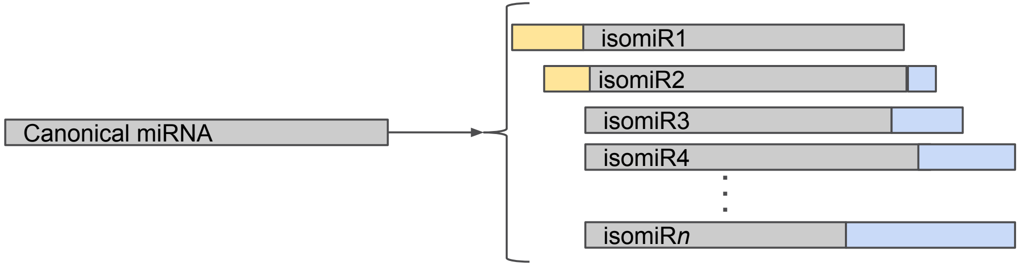
OXFORD

Sequence analysis

QuagmiR: a cloud-based application for isomiR big data analytics

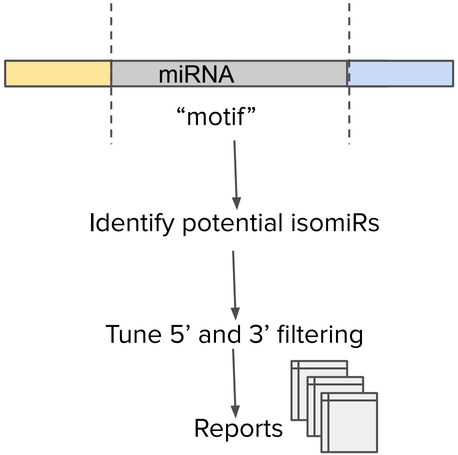
Xavier Bofill-De Ros[1], Kevin Chen[1], Susanna Chen[1], Nikola Tesic[2], Dusan Randjelovic[2], Nikola Skundric[2], Svetozar Nesic[2], Vojislav Varjacic[2], Elizabeth H. Williams[2], Raunaq Malhotra[2], Minjie Jiang[1] and Shuo Gu[1,*]

[1]RNA Mediated Gene Regulation Section, RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, MD, USA and [2]Seven Bridges Genomics Inc., Cambridge, MA, USA



- **QuagmiR**: a tool that pulls specific reads and aligns them against a consensus sequence in the middle of a miRNA, allowing mismatches on the ends to capture 3′ isomiRs
- **Initial idea:** reprocess all TCGA miRNAs with QuagmiR
- **The CGC enabled an efficient and highly scalable analysis, hence more research projects resulted from the initial one**
- Xavier used the fact that miRNA data tends to be smaller in size and leveraged CGC capabilities to easily analyze up to **70 samples per task across dozens of tasks**

QuagmiR tool



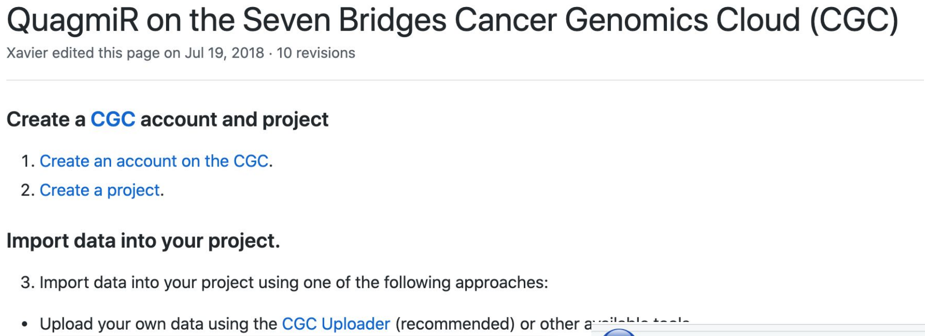
Identify potential isomiRs

Tune 5’ and 3’ filtering

Reports

GitHub Gu-Lab-RBL-NCI/QuagmiR

Quantify and visualize isomiR differences



QuagmiR on the Seven Bridges Cancer Genomics Cloud (CGC)

Xavier edited this page on Jul 19, 2018 · 10 revisions

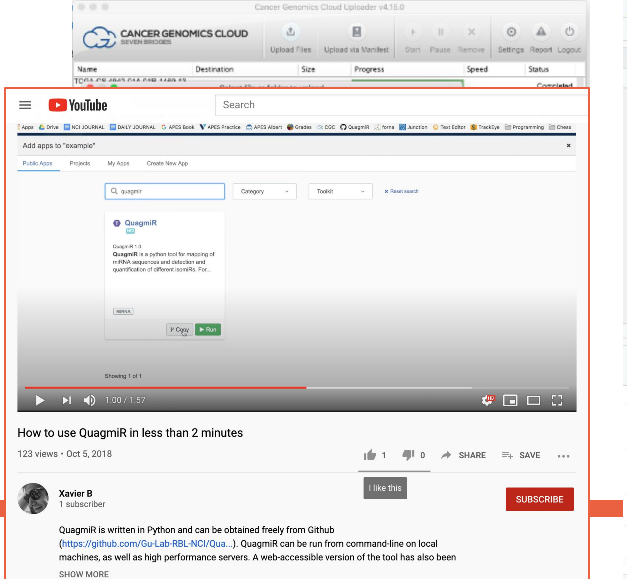
Create a CGC account and project

1. Create an account on the CGC.
2. Create a project.

Import data into your project.

3. Import data into your project using one of the following approaches:

- Upload your own data using the CGC Uploader (recommended) or other a



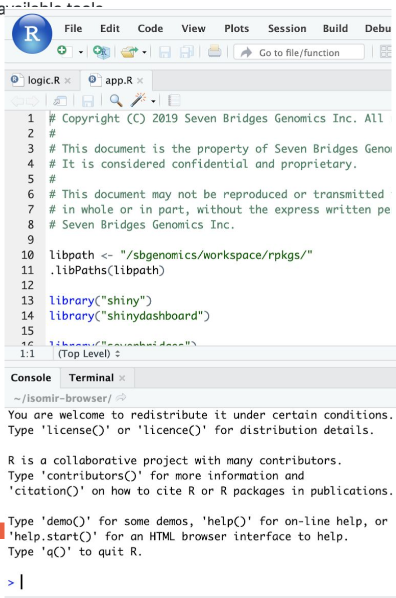
QuagmiR 1.0
QuagmiR is a python tool for mapping of miRNA sequences and detection and quantification of different isomiRs. For...

How to use QuagmiR in less than 2 minutes

123 views • Oct 5, 2018

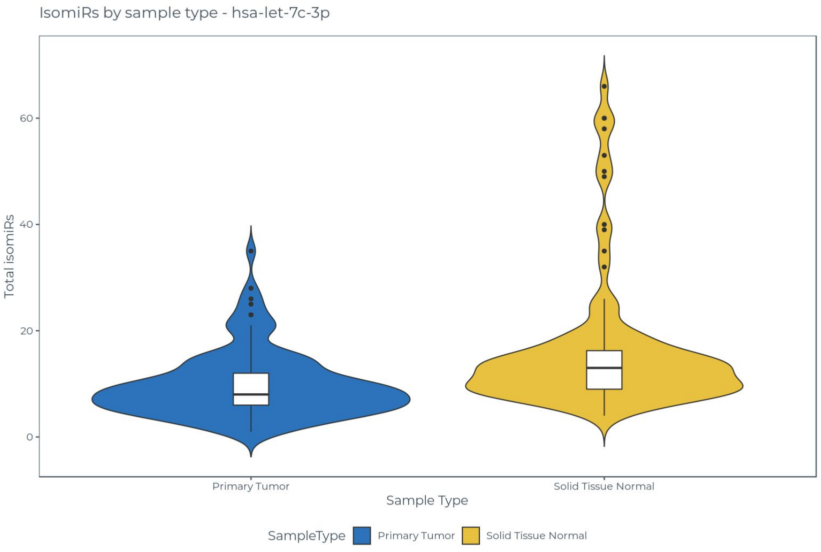
Xavier B
1 subscriber

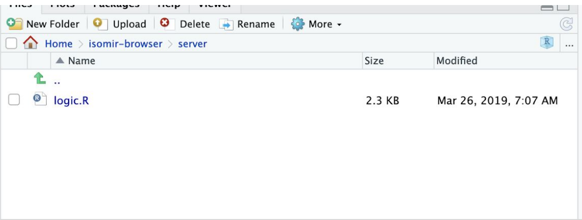
QuagmiR is written in Python and can be obtained freely from Github (https://github.com/Gu-Lab-RBL-NCI/Qua...). QuagmiR can be run from command-line on local machines, as well as high performance servers. A web-accessible version of the tool has also been



```
# Copyright (C) 2019 Seven Bridges Genomics Inc. All
#
# This document is the property of Seven Bridges Genon
# It is considered confidential and proprietary.
#
# This document may not be reproduced or transmitted
# in whole or in part, without the express written pe
# Seven Bridges Genomics Inc.

libpath <- "/sbgenomics/workspace/rpkgs/"
.libPaths(libpath)

library("shiny")
library("shinydashboard")

library("sevenbridges")
```

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```





```
Name          Size      Modified
..
logic.R       2.3 KB    Mar 26, 2019, 7:07 AM
```

Check out CGC March Webinar recording at https://www.cancergenomicscloud.org/webinars

Support and Resources

CGC Monthly Webinar Series

Learn about CGC platform features that you can use in your projects.

Variety of research and technical topics in the field of cancer research using the CGC

Resources: Upcoming webinar info, slides and recordings are available at: https://www.cancergenomicscloud.org/webinars

Save the date/time: 4th Wednesday of each month at 2pm ET

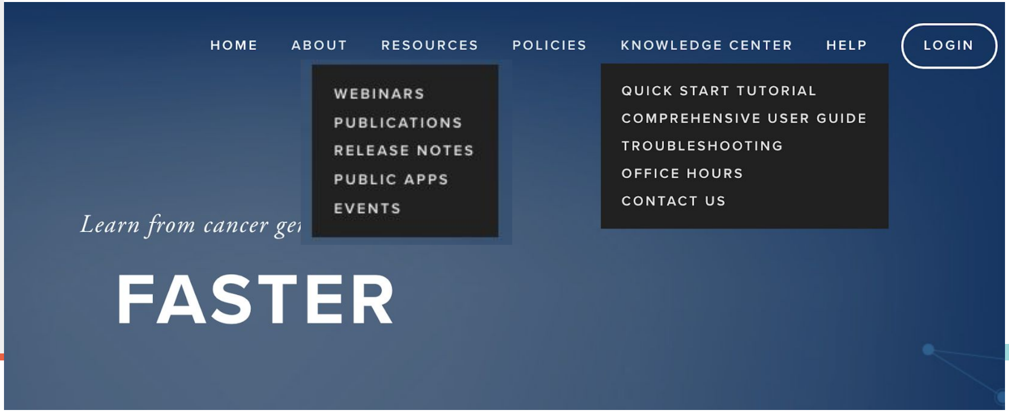
CGC Knowledge Center

https://docs.cancergenomicscloud.org/

Contact CGC Support: cgc@sevenbridges.com

Office Hours: Every week on Thursdays

https://www.cancergenomicscloud.org/officehours



In Summary

Data Access

Immediately access petabytes of ***Open and Controlled*** TCGA, CPTAC, TCIA, and other omics datasets

Bring your own private cohorts alongside public data.

Tools and Workflows

Standard bioinformatics pipelines

Bring your own analysis tools directly to the platform

Connect multiple tools together using our interactive custom workflow builder

Collaborate on the cloud

Collaborate with other researchers around the world in a secure workspace

Access to high-throughput, cost-effective cloud computing resources and storage on demand and at cost.

6000 users

>80 countries

1,600,000+ computational tasks

1400+ years of total compute time

66,800+ workflows

500+ public apps

Interactive Analysis

The ability to perform custom, interactive analysis and visualization on the platform using Python, RStudio.

Support & Resources

Access comprehensive online documentation and training resources; Technical support from a team of >200 expert scientists, bioinformaticians, and engineers.

Questions?