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

May 14, 2021

Sai Lakshmi Subramanian
Program Manager, Seven Bridges
sai.subramanian@sbgenomics.com

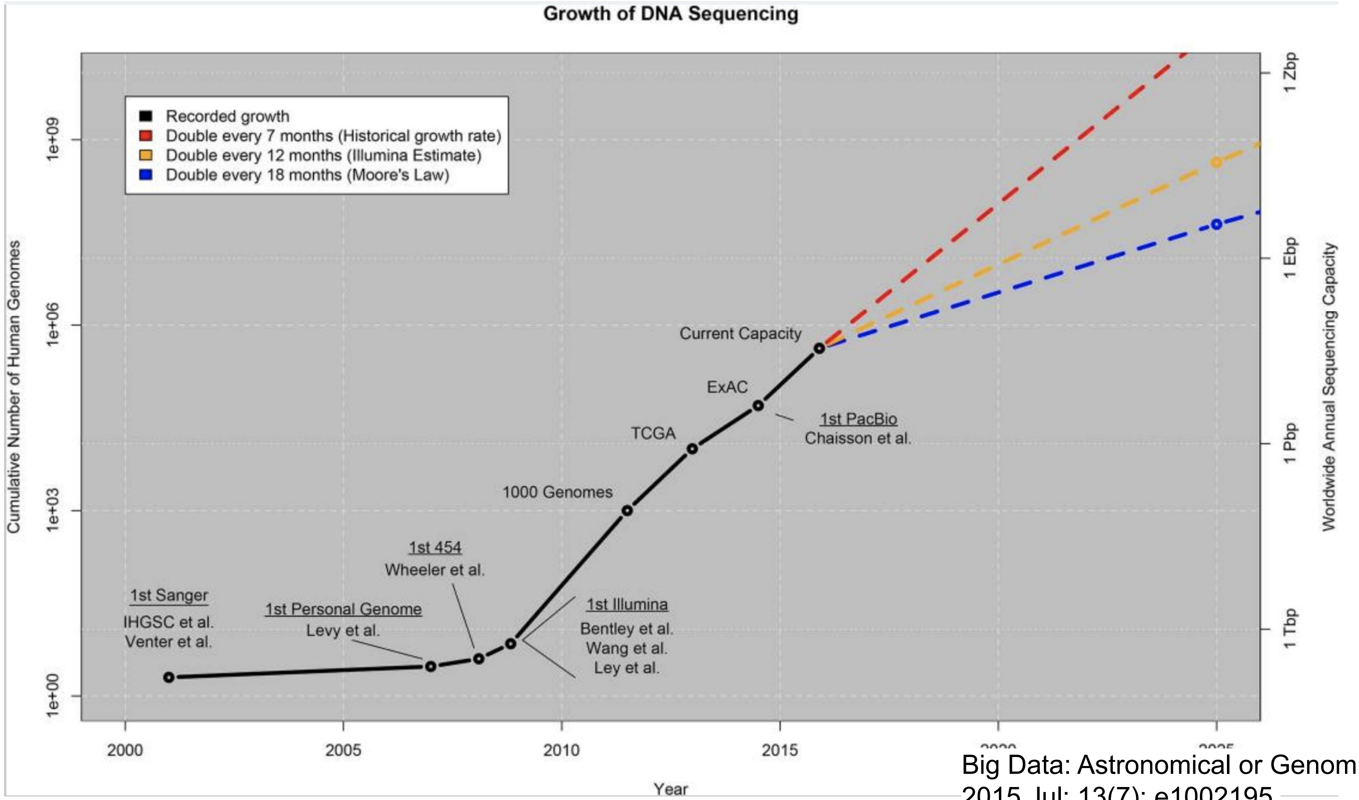
CANCER GENOMICS CLOUD
SEVEN BRIDGES

Agenda

- Background
- Access multi-omics datasets in the CGC
- Use Case: Exploring the human proteome - Analysis of CPTAC datasets
  - Features to control cloud costs
- Ongoing Projects - New workflows for novel science
- Questions/Discussion

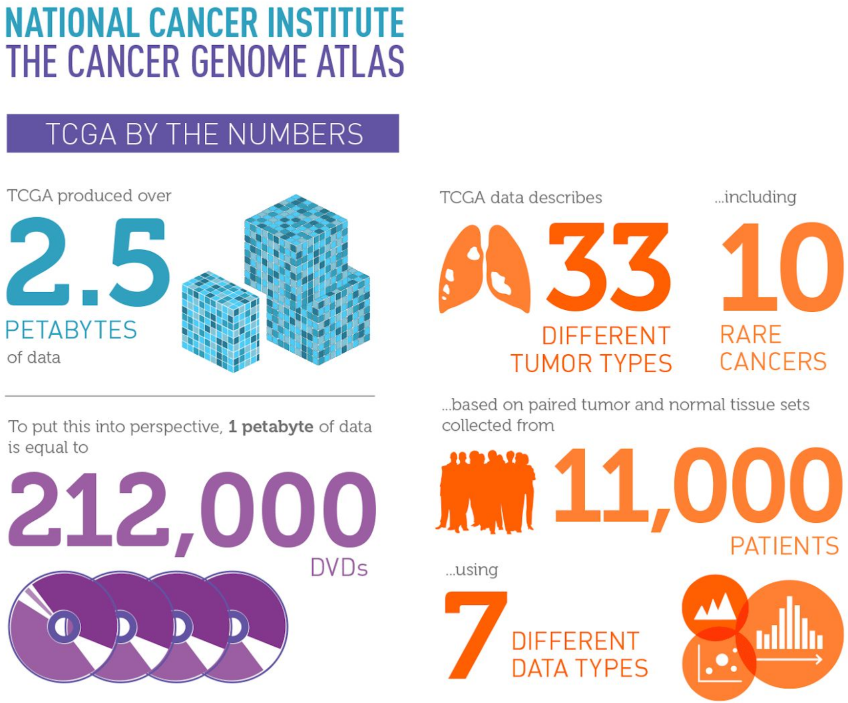
Background

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

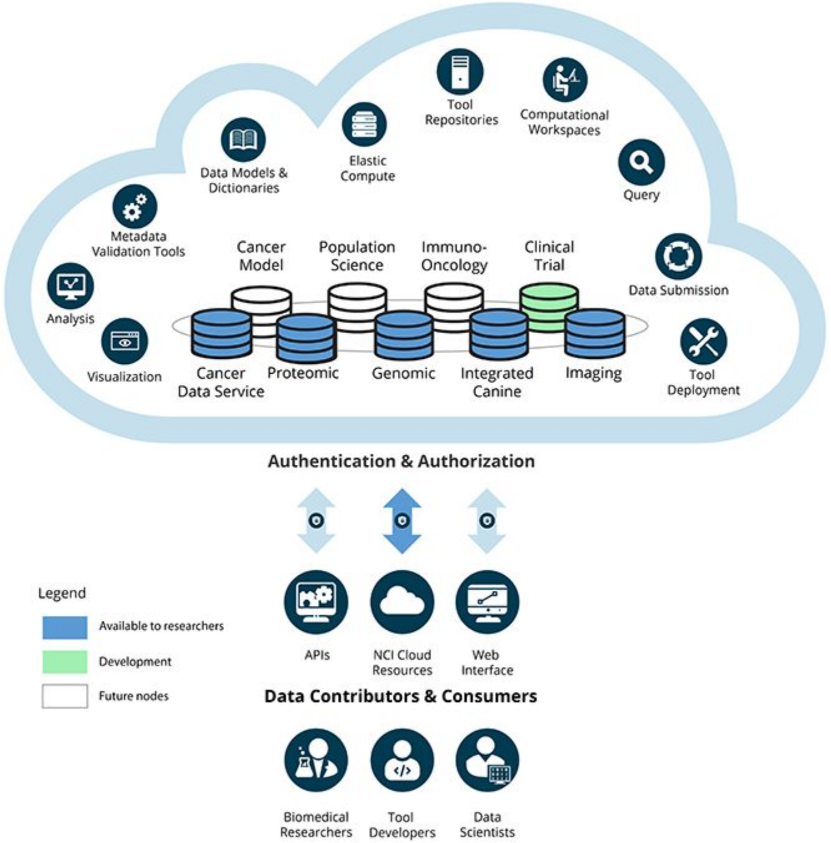
The Seven Bridges Cancer Genomics Cloud (CGC)

CANCER GENOMICS CLOUD
SEVEN BRIDGES

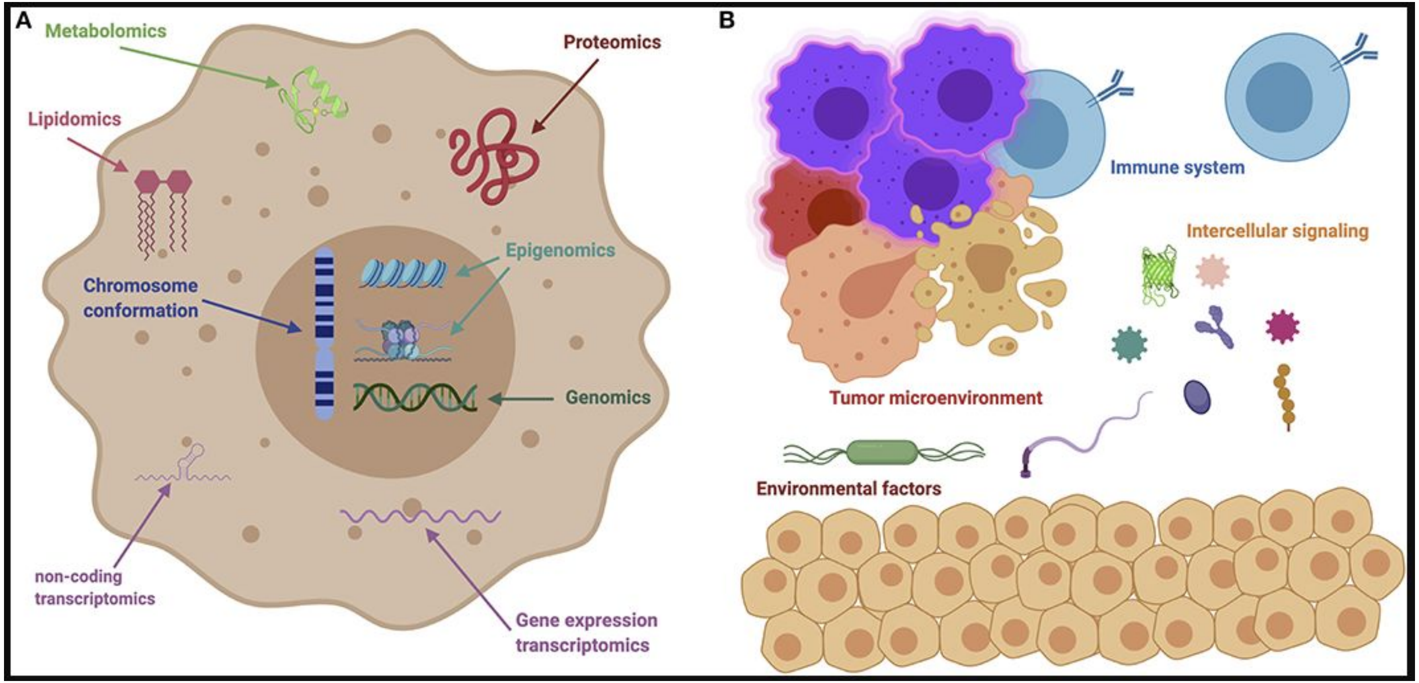
A Cloud Resource within the NCI Cancer Research Data Commons for secure storage, sharing & analysis of petabytes of public, multi-omic cancer datasets

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

NCI Cancer Research Data Commons (CRDC)



Multi-omic data is critical for cancer research

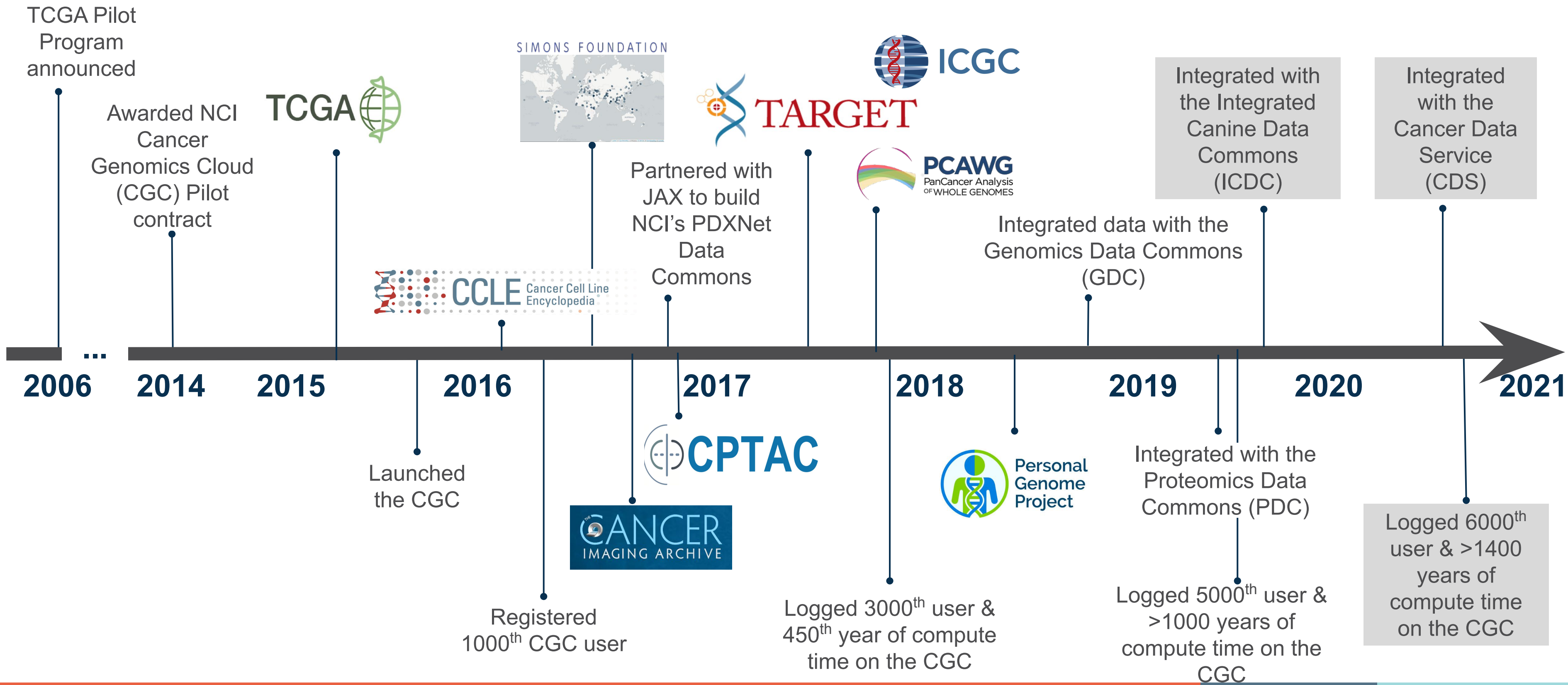


Cancer is a complex disease!

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

Growth of the Cancer Genomics Cloud Ecosystem



2006 … 2014 2015 2016 2017 2018 2019 2020 2021

TCGA Pilot Program announced

Awarded NCI Cancer Genomics Cloud (CGC) Pilot contract

TCGA

Launched the CGC

CCLE Cancer Cell Line Encyclopedia

Registered 1000th CGC user

SIMONS FOUNDATION

THE CANCER IMAGING ARCHIVE

Partnered with JAX to build NCI's PDXNet Data Commons

CPTAC

TARGET

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

PCAWG PanCancer Analysis OF WHOLE GENOMES

ICGC

Personal Genome Project

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

Cloud-based Environment for Collaborative Research and Bioinformatics Data Analysis

- 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

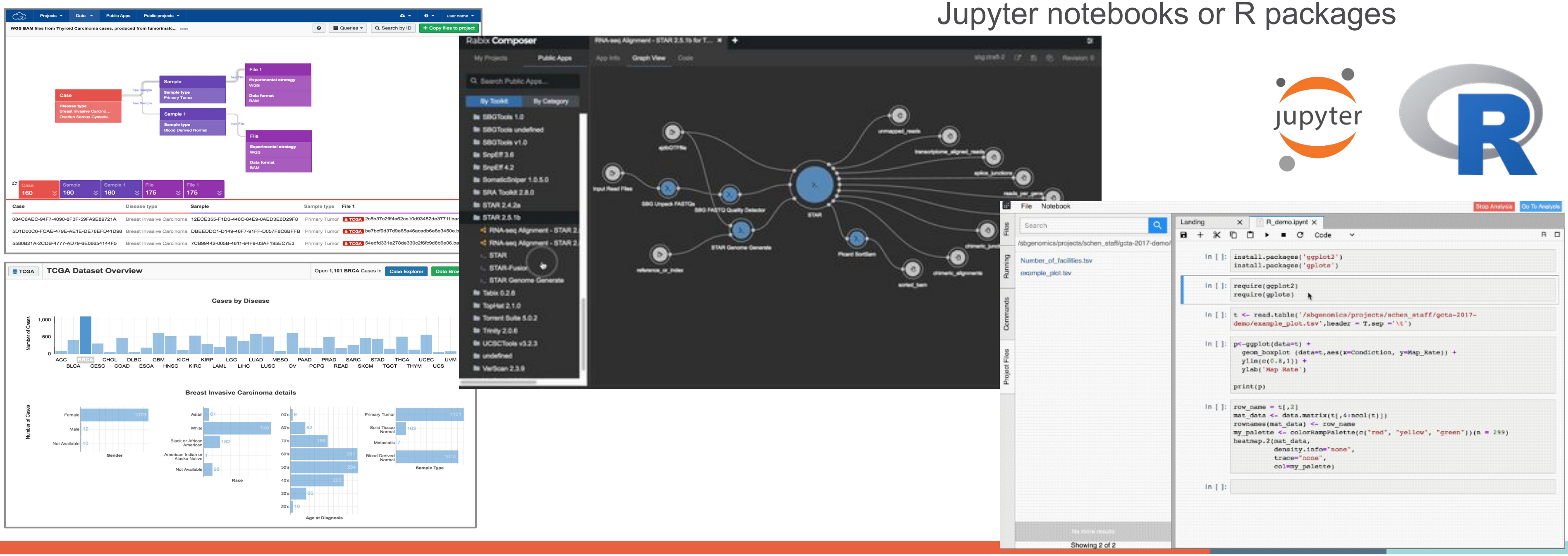
Accelerating cancer research

- 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

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 [ ]:
```

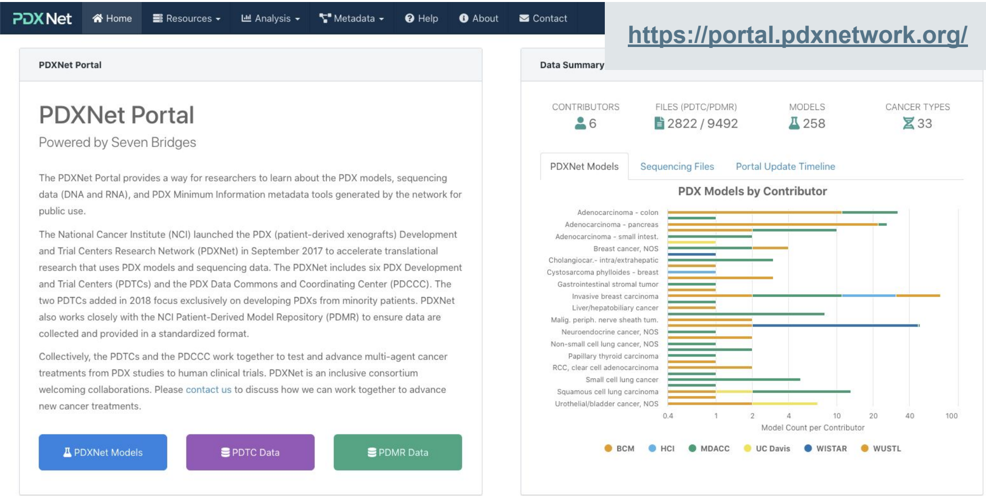
Empowering a coordinating center on the CGC

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



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.

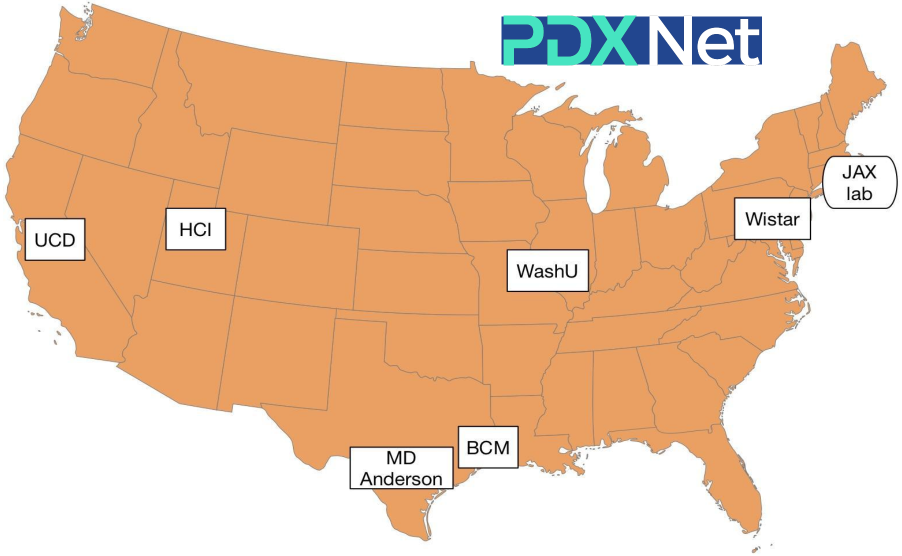
Collectively, the PDTCs and the PDCCC work together to test and advance multi-agent cancer treatments from PDX studies to human clinical trials. PDXNet is an inclusive consortium welcoming collaborations. Please contact us to discuss how we can work together to advance new cancer treatments.

Data Summary

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

PDXNet Models
Sequencing Files
Portal Update Timeline
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

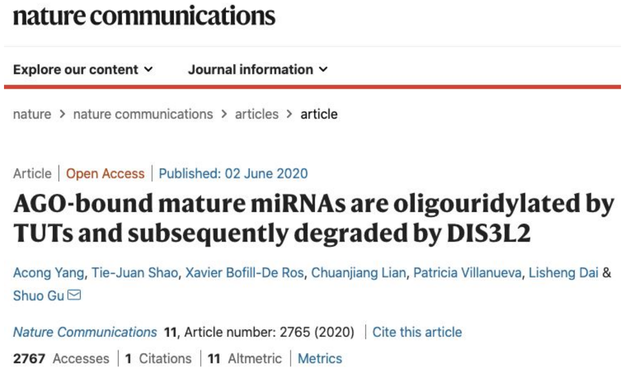
https://portal.pdxnetwork.org/



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

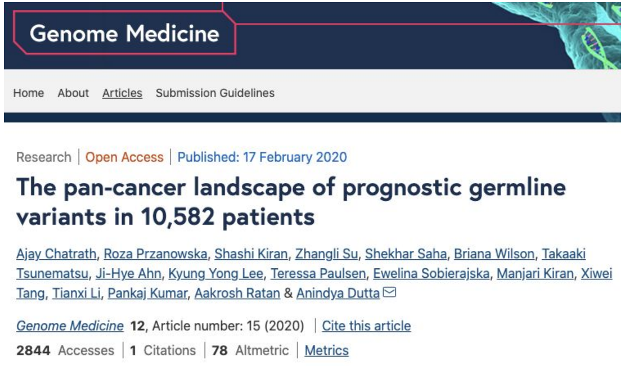
Article | Open Access | Published: 02 June 2020

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

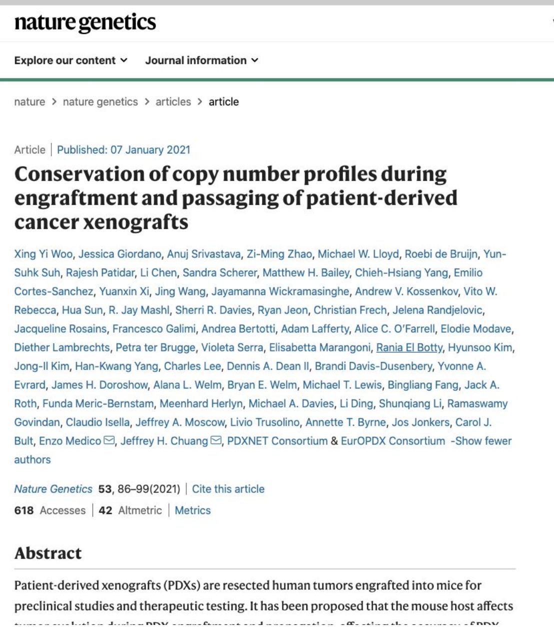
Research | Open Access | Published: 17 February 2020

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

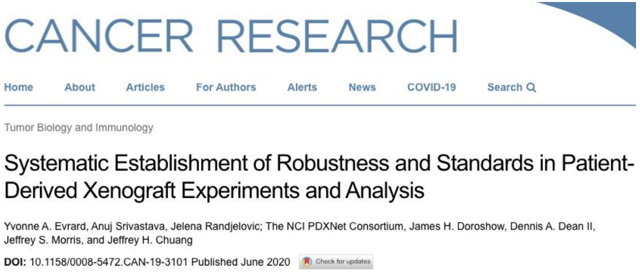
Xing Yi Woo, Jessica Giordano, Anuj Srivastava, Zi-Ming Zhao, Michael W. Lloyd, Roebi de Bruijn, Yun-Suhk Suh, Rajesh Patidar, Li Chen, Sandra Scherer, Matthew H. Bailey, Chieh-Hsiang Yang, Emilio Cortes-Sanchez, Yuanxin Xi, Jing Wang, Jayamanna Wickramasinghe, Andrew V. Kossenkov, Vito W. Rebecca, Hua 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



CANCER RESEARCH

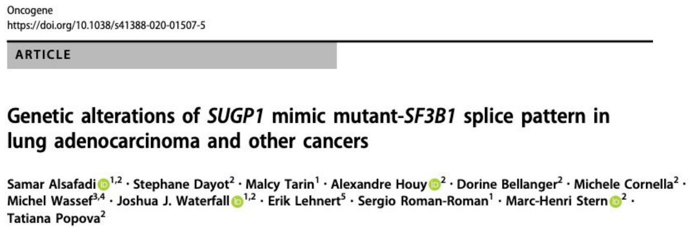
Home About Articles For Authors Alerts News COVID-19 Search

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]

Find a growing list of publications at: https://www.cancergenomicscloud.org/publications

Participating in open standards groups helps make us more FAIR

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

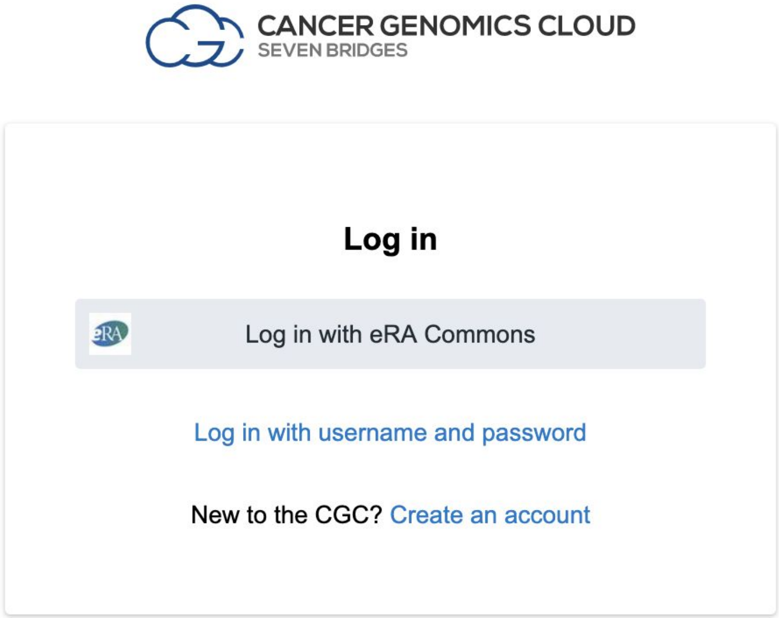
BioCompute
Objects

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



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
THE 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!

REPOSITORIES

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

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)

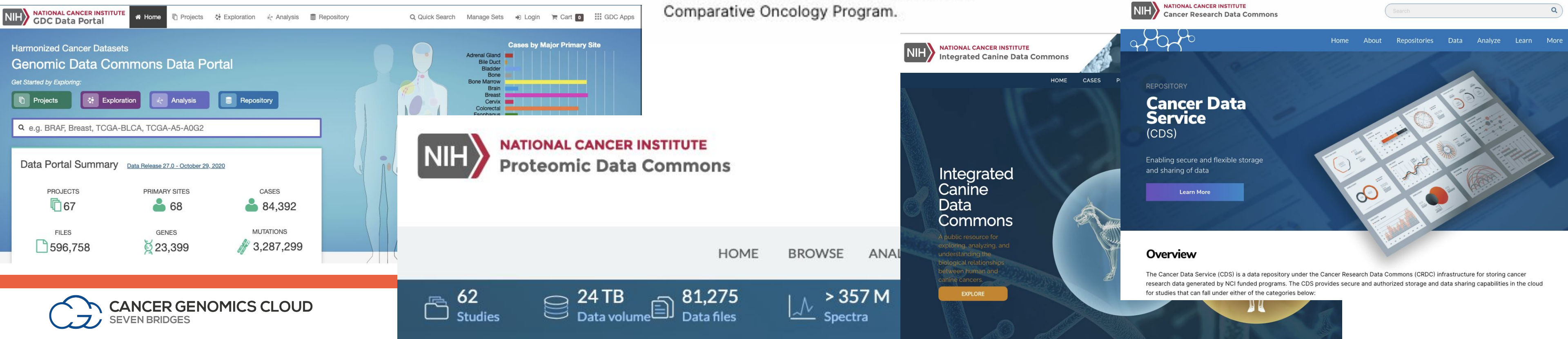
Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies.

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	24 TB	81,275	> 357 M
Studies	Data volume	Data files	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: Exploring the human proteome - Analysis of CPTAC datasets

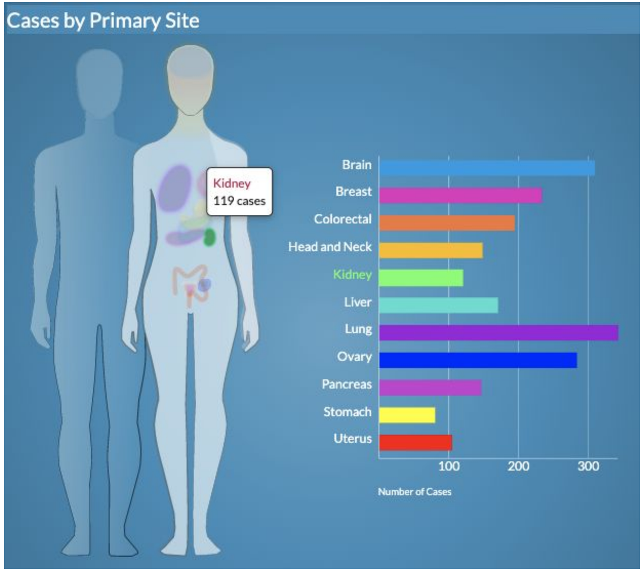
Proteogenome characterization of ccRCC

How do I analyze the spectra data of the proteome in clear cell renal cell carcinoma cases?

Kidney cancer is among the 10 most common cancers in both men and women

~73,000 new cases with >14,000 deaths in 2020

https://seer.cancer.gov/statfacts/html/kidrp.html



- Proteome and phosphoproteome data from the ccRCC tumors is available in PDC along with peptide spectrum analyses (PSMs) and protein summary reports from the CPTAC common data analysis pipeline (CDAP).
- Using the CGC, process high-throughput data-dependent acquisition (DDA) tandem mass spectrometry data acquired from peptides labeled with TMT tags.
- The multiplexed labeling, as applied in the CPTAC program, allows for differential quantitation across multiple tumor/normal samples.

Image Courtesy: https://pdc.cancer.gov/pdc/

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

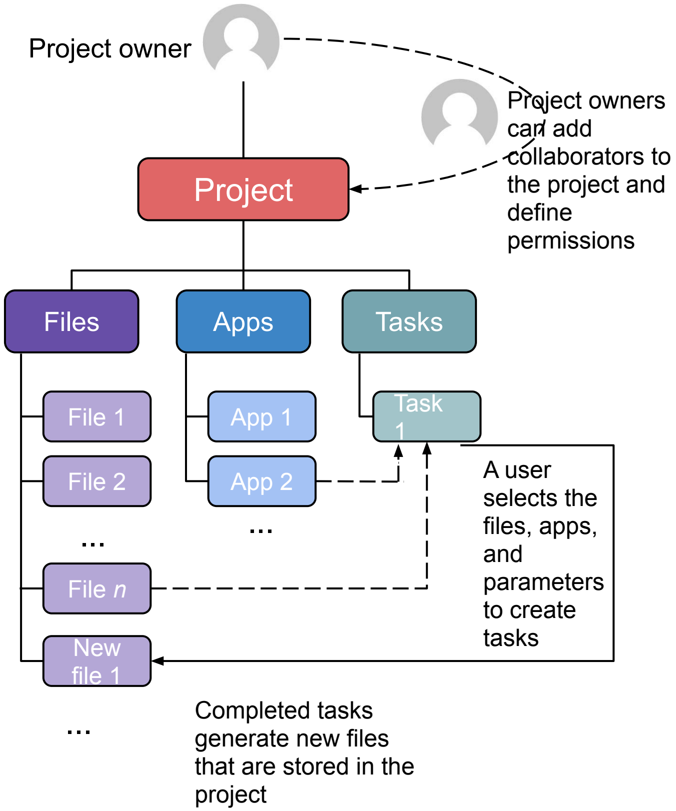
- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Analyze

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

Projects organize files, methods, and results



Project owner

Project owners can add collaborators to the project and define permissions

Project

Files

Apps

Tasks

File 1

File 2

...

File *n*

New file 1

...

App 1

App 2

...

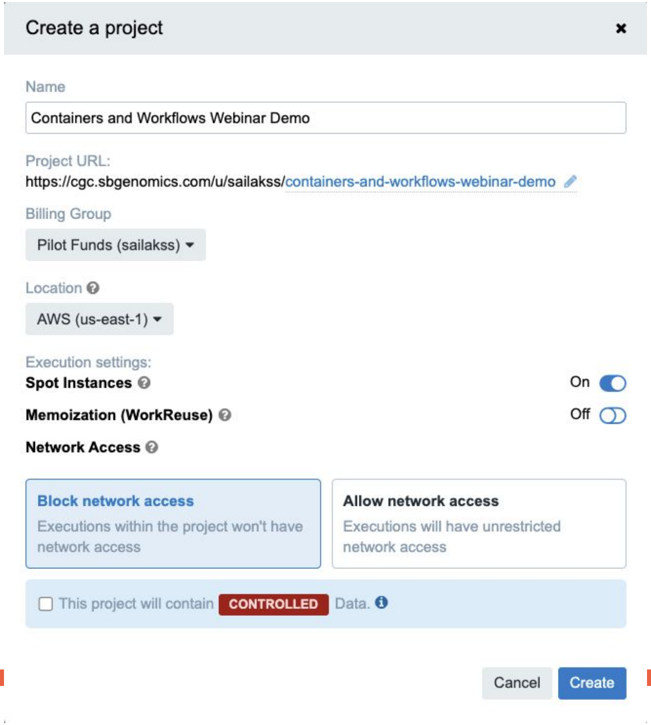
Task 1

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



Create a project ×

Name

Containers and Workflows Webinar Demo

Project URL:

https://cgc.sbgenomics.com/u/sailakss/containers-and-workflows-webinar-demo

Billing Group

Pilot Funds (sailakss) ▾

Location

AWS (us-east-1) ▾

Execution settings:

Spot Instances On

Memoization (WorkReuse) Off

Network Access

Block network access
Executions within the project won't have network access

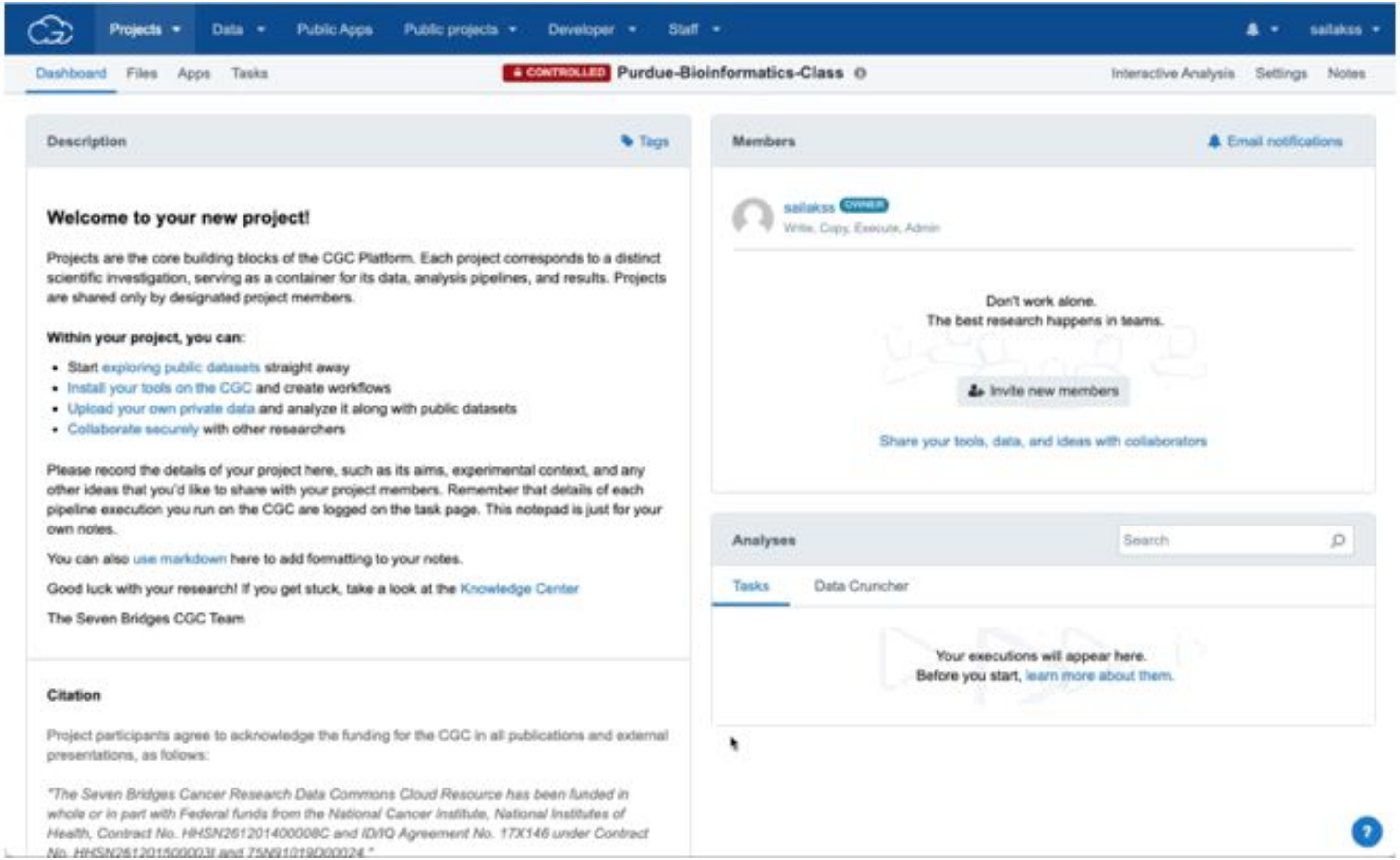
Allow network access
Executions will have unrestricted network access

□ 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



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.

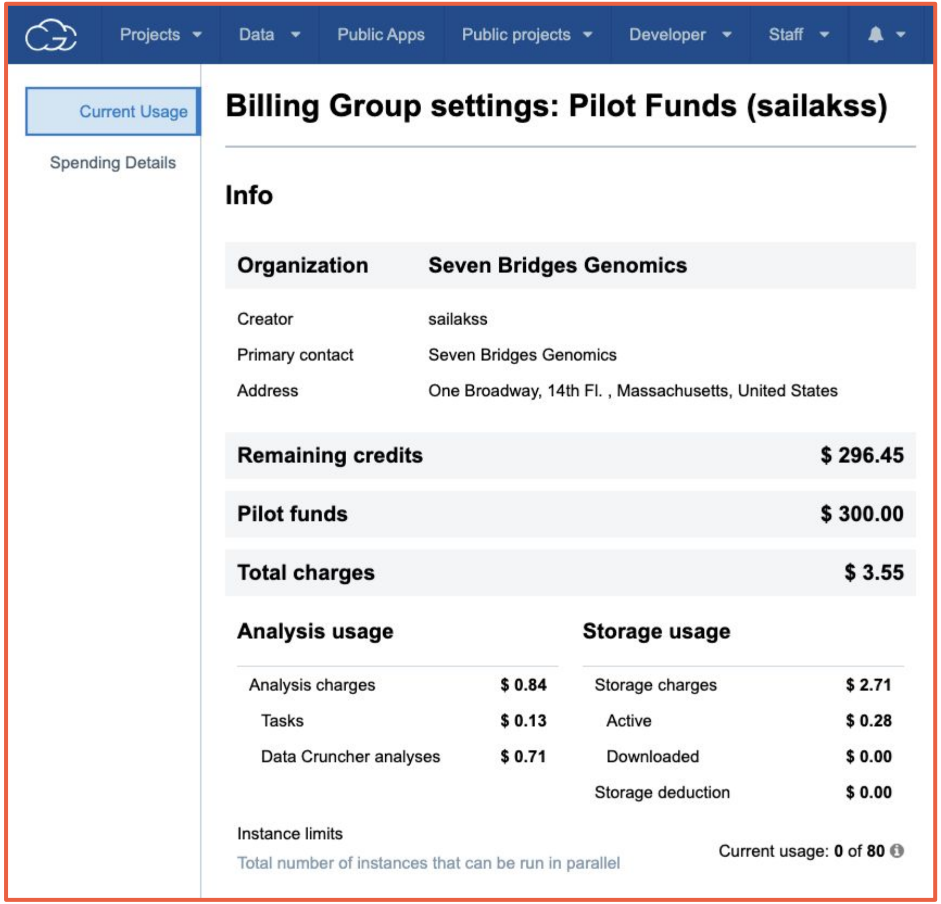
Billing groups

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	**$ 296.45**
Pilot funds	**$ 300.00**
Total charges	**$ 3.55**

Analysis usage

Analysis charges	$ 0.84
Tasks	$ 0.13
Data Cruncher analyses	$ 0.71

Storage usage

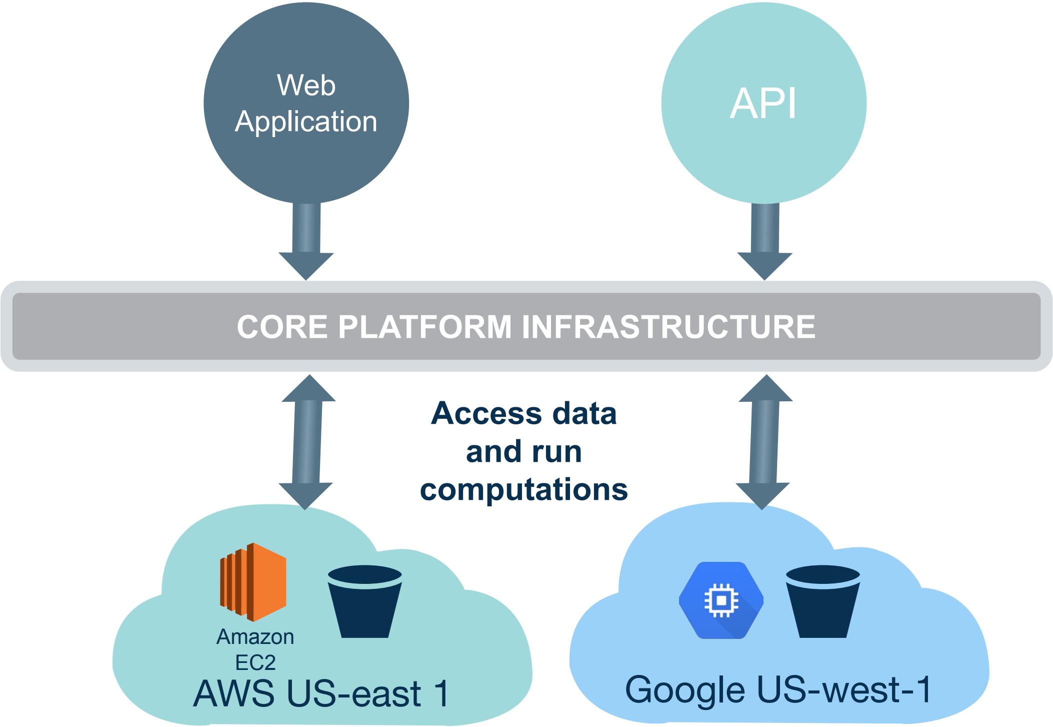
Storage charges	$ 2.71
Active	$ 0.28
Downloaded	$ 0.00
Storage deduction	$ 0.00

Instance limits

Total number of instances that can be run in parallel

Current usage: **0** of **80**

Multi-cloud implementation on the CGC



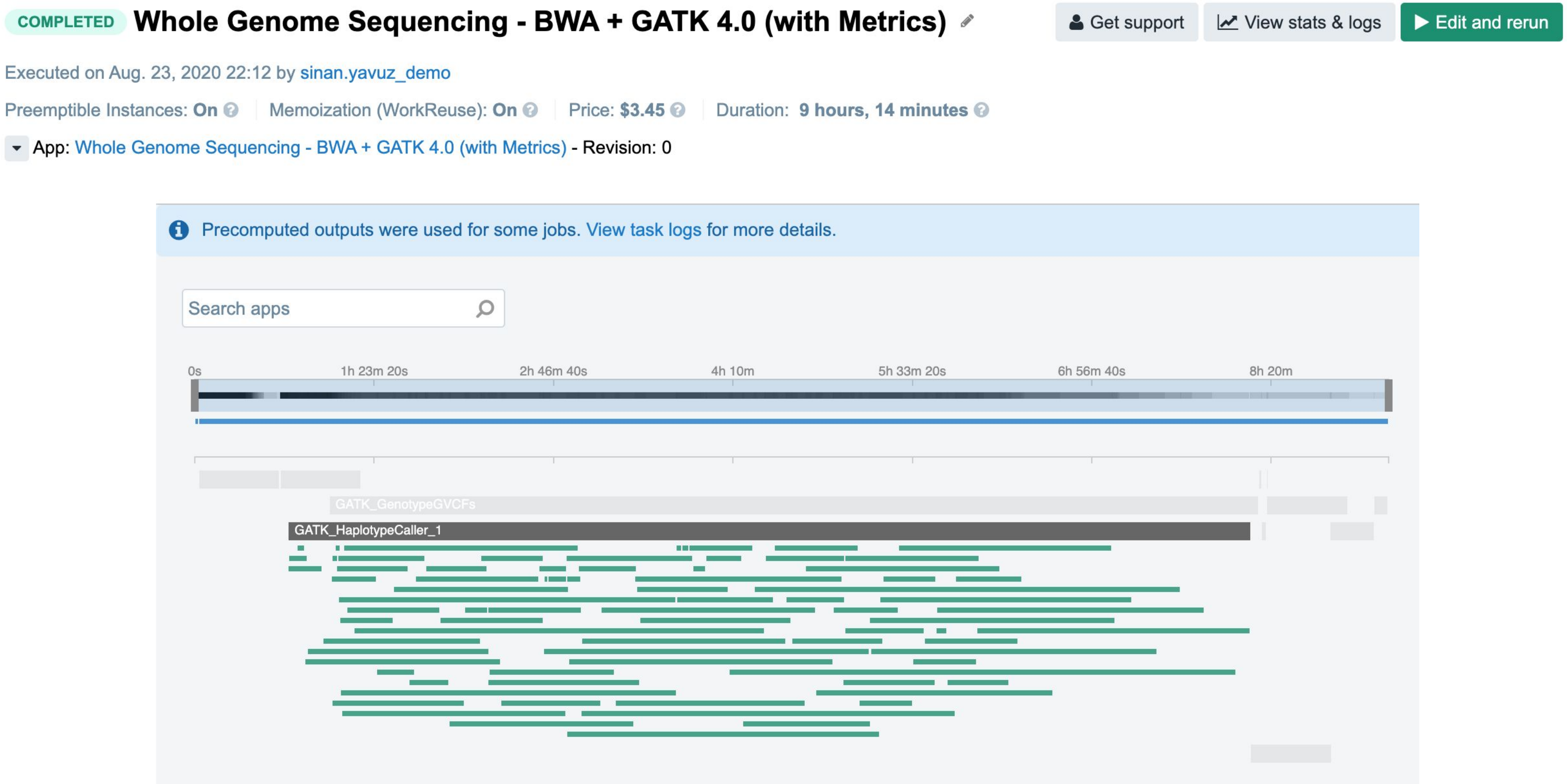
Web Application

API

CORE PLATFORM INFRASTRUCTURE

Access data and run computations

Memoization allows use of previously computed results



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

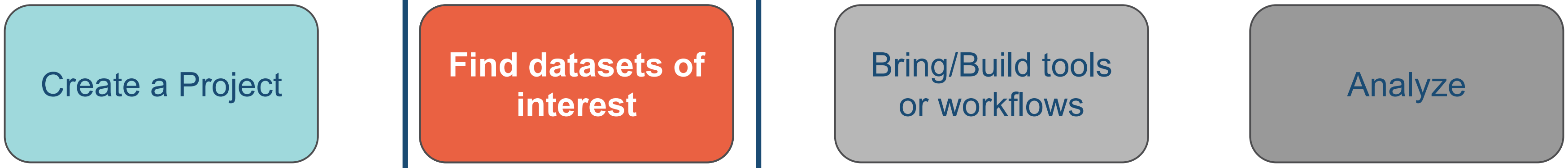
Executed on Aug. 23, 2020 22:12 by sinan.yavuz_demo

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

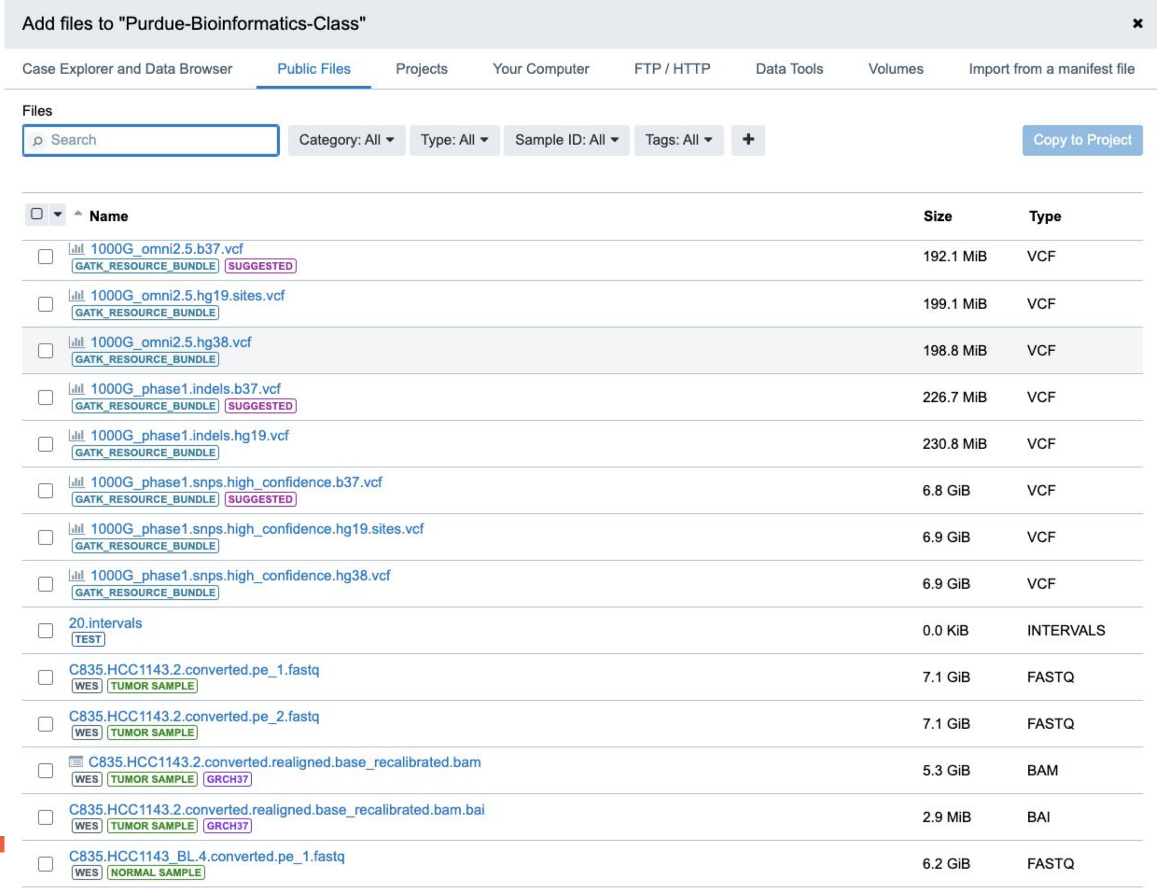
Tools, workflows, and software packages

- 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

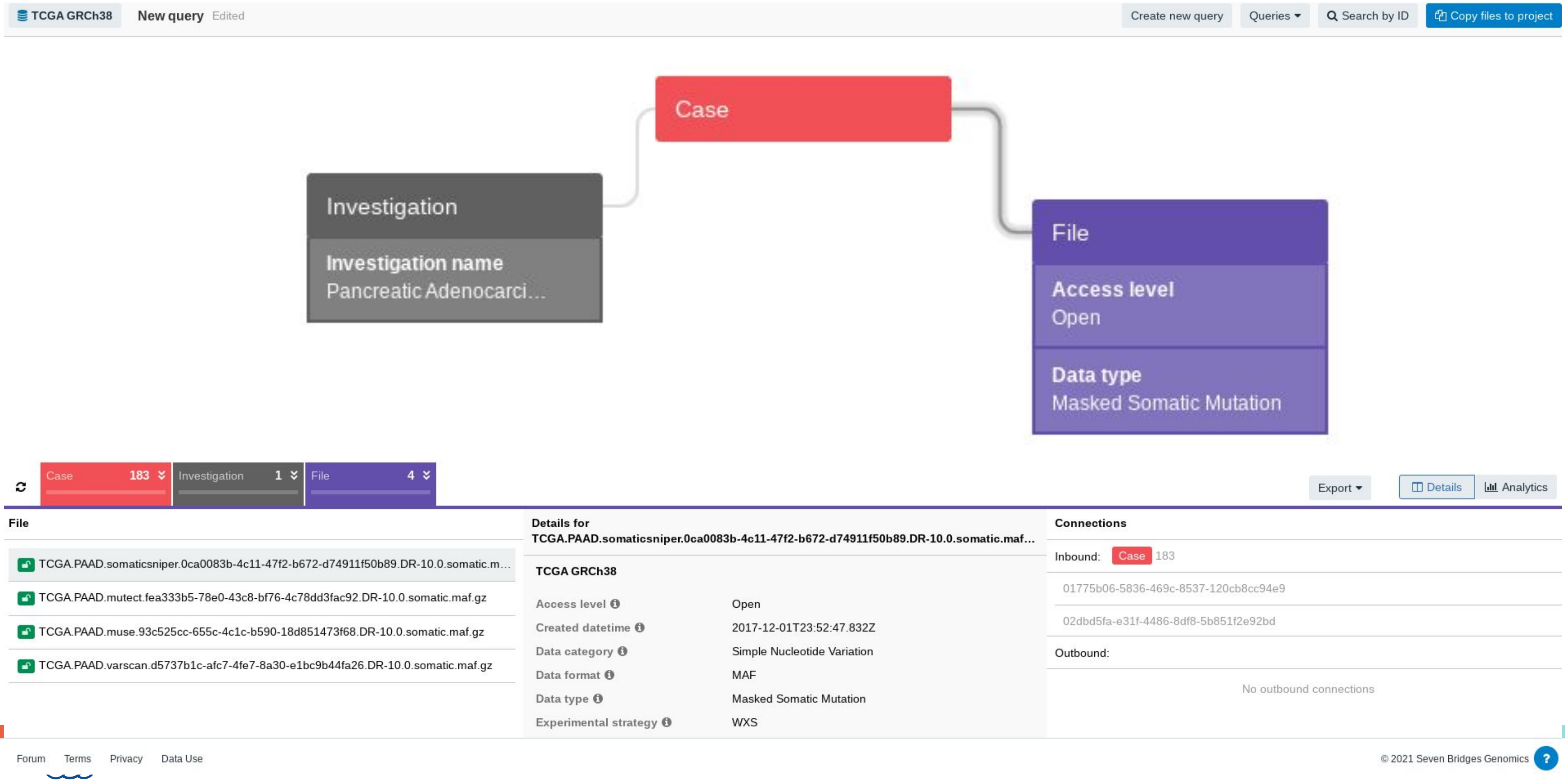


Add files to "Purdue-Bioinformatics-Class" ×

□	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



Investigation

Investigation name
Pancreatic Adenocarci...

File

Access level
Open

Data type
Masked Somatic Mutation

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

Connections

Inbound: Case 183

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

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

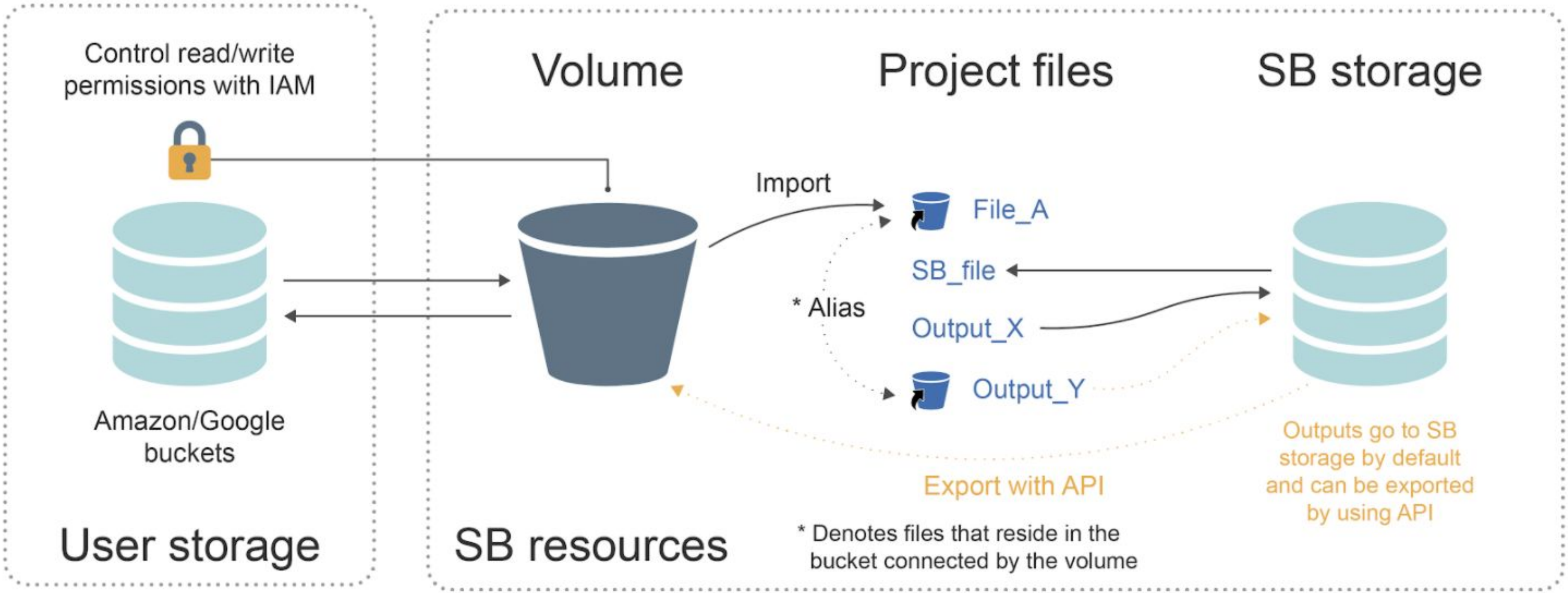
Outbound:

No outbound connections

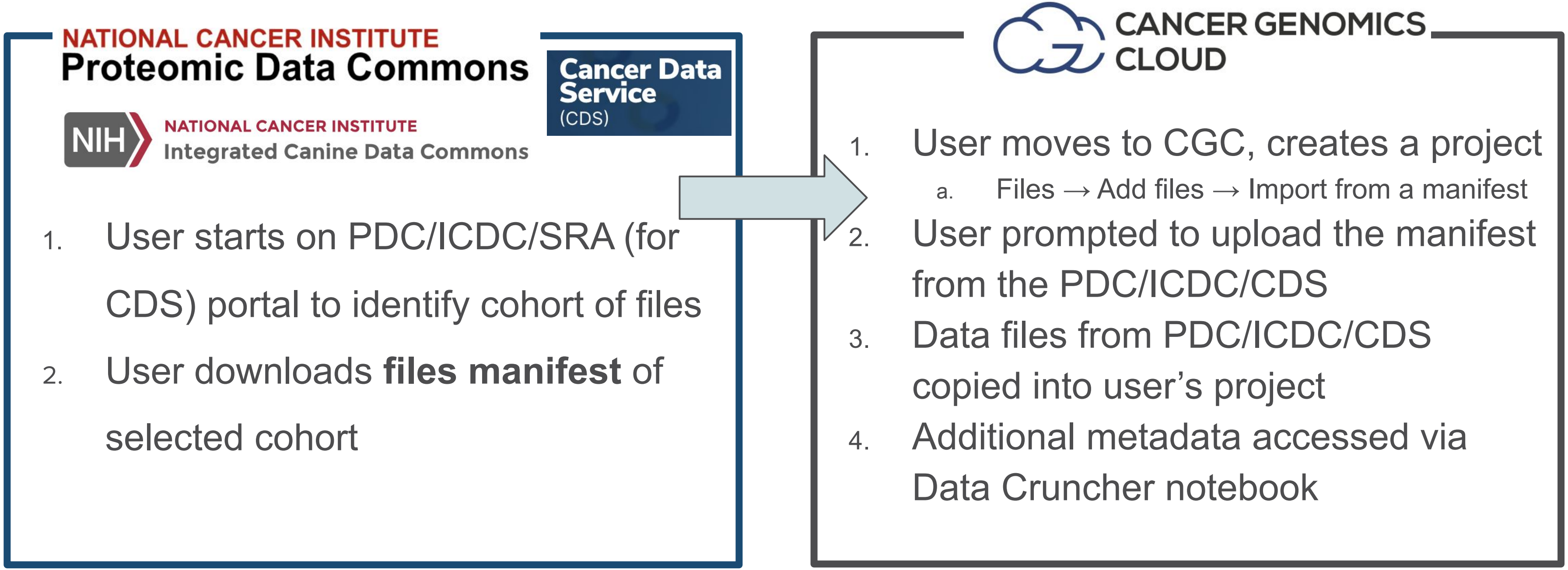
Forum Terms Privacy Data Use

© 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

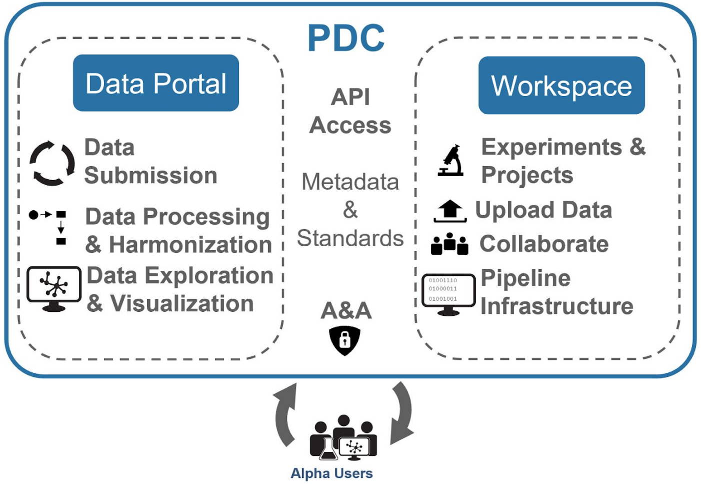
CANCER GENOMICS CLOUD

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

Integration with the Proteomic Data Commons (PDC)

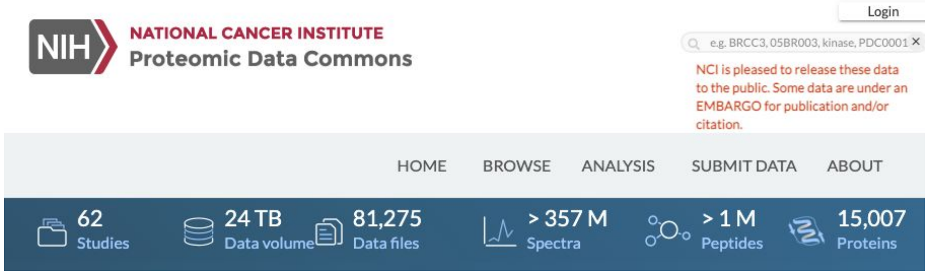
Proteome Data Commons – democratize access to cancer-related proteomic datasets



Courtesy of Izumi HInkson

CGC integration with PDC enables both tool developers and researchers to take full advantage of rich data resources

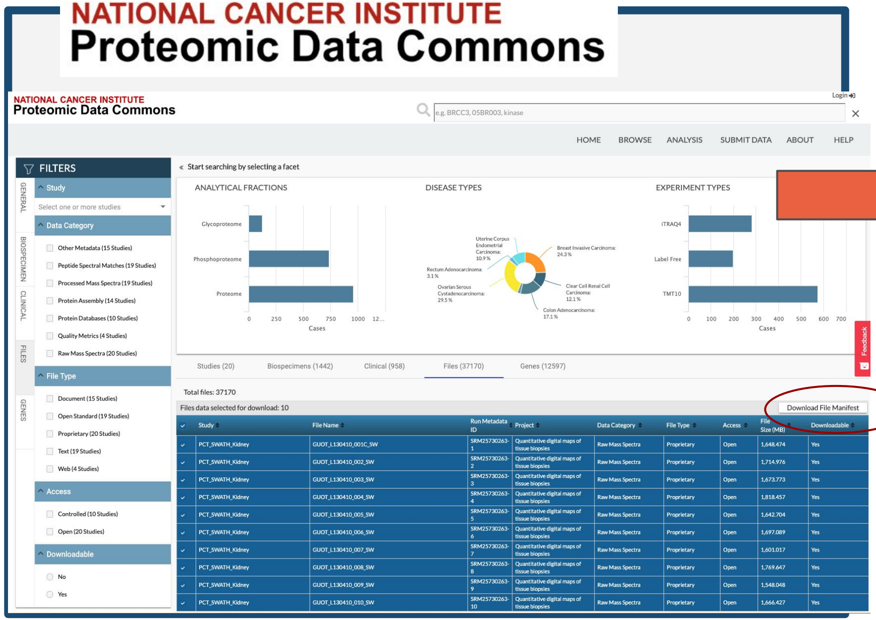
Proteomics tools can run quickly and efficiently in the cloud, lowering barriers for usage



NCI is pleased to release these data to the public. Some data are under an EMBARGO for publication and/or citation.

62 Studies | 24 TB Data volume | 81,275 Data files | > 357 M Spectra | > 1 M Peptides | 15,007 Proteins

How to import data from PDC



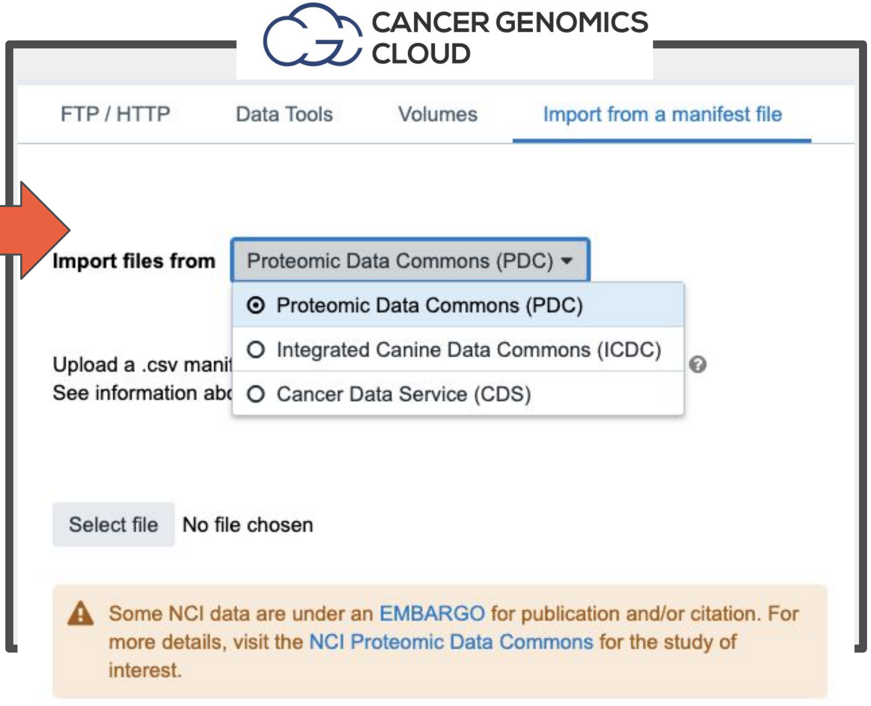
NATIONAL CANCER INSTITUTE
Proteomic Data Commons

NATIONAL CANCER INSTITUTE
Proteomic Data Commons
Login
e.g. BRCC3, 05BR003, kinase
HOME
BROWSE
ANALYSIS
SUBMIT DATA
ABOUT
HELP
FILTERS
GENERAL
BIOSPECIMEN
CLINICAL
FILES
GENES
Study
Select one or more studies
Data Category
□ Other Metadata (15 Studies)
□ Peptide Spectral Matches (19 Studies)
□ Processed Mass Spectra (19 Studies)
□ Protein Assembly (14 Studies)
□ Protein Databases (10 Studies)
□ Quality Metrics (4 Studies)
□ Raw Mass Spectra (20 Studies)
File Type
□ Document (15 Studies)
□ Open Standard (19 Studies)
□ Proprietary (20 Studies)
□ Text (19 Studies)
□ Web (4 Studies)
Access
□ Controlled (10 Studies)
□ Open (20 Studies)
Downloadable
○ No
○ Yes
« Start searching by selecting a facet
ANALYTICAL FRACTIONS
Glycoproteome
Phosphoproteome
Proteome
0
250
500
750
1000
12...
Cases
DISEASE TYPES
Uterine Corpus Endometrial Carcinoma: 10.9 %
Breast Invasive Carcinoma: 24.3 %
Rectum Adenocarcinoma: 3.1 %
Ovarian Serous Cystadenocarcinoma: 29.5 %
Clear Cell Renal Cell Carcinoma: 12.1 %
Colon Adenocarcinoma: 17.1 %
EXPERIMENT TYPES
iTRAQ4
Label Free
TMT10
0
100
200
300
400
500
600
700
Cases
Feedback
Studies (20)
Biospecimens (1442)
Clinical (958)
Files (37170)
Genes (12597)
Download File Manifest

Total files: 37170

Files data selected for download: 10

✓	Study	File Name	Run Metadata ID	Project	Data Category	File Type	Access	File Size (MB)	Downloadable
✓	PCT_SWATH_Kidney	GUOT_L130410_001C_SW	SRM25730263-1	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,648.474	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_002_SW	SRM25730263-2	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,714.976	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_003_SW	SRM25730263-3	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,673.773	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_004_SW	SRM25730263-4	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,818.457	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_005_SW	SRM25730263-5	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,642.704	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_006_SW	SRM25730263-6	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,697.089	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_007_SW	SRM25730263-7	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,601.017	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_008_SW	SRM25730263-8	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,769.647	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_009_SW	SRM25730263-9	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,548.048	Yes
✓	PCT_SWATH_Kidney	GUOT_L130410_010_SW	SRM25730263-10	Quantitative digital maps of tissue biopsies	Raw Mass Spectra	Proprietary	Open	1,666.427	Yes



CANCER GENOMICS CLOUD

Import files from: Proteomic Data Commons (PDC)

- ● Proteomic Data Commons (PDC)
- ○ Integrated Canine Data Commons (ICDC)
- ○ Cancer Data Service (CDS)

Upload a .csv manif

See information abo

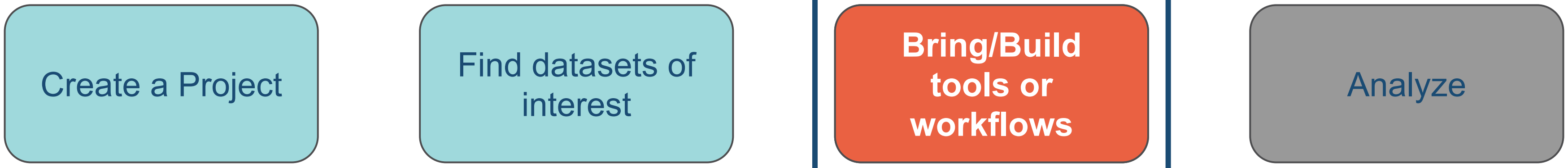
Select file No file chosen

Some NCI data are under an EMBARGO for publication and/or citation. For more details, visit the NCI Proteomic Data Commons for the study of interest.

Researchers can also bring private data to co-analyze with public resources.

Import

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

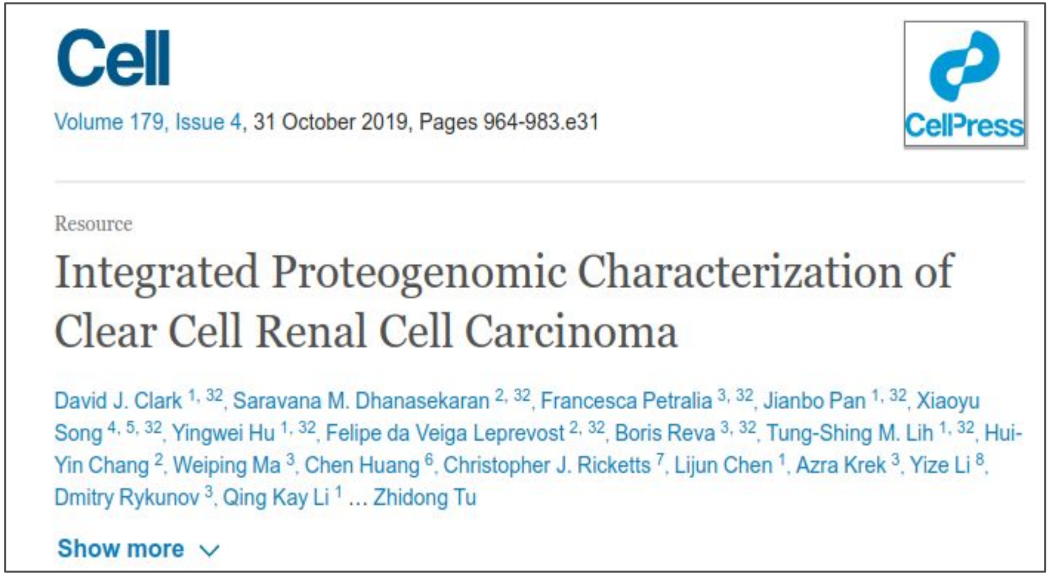
Tools, workflows, and software packages

- 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

FragPipe: A complete proteomics pipeline



Cell

Volume 179, Issue 4, 31 October 2019, Pages 964-983.e31

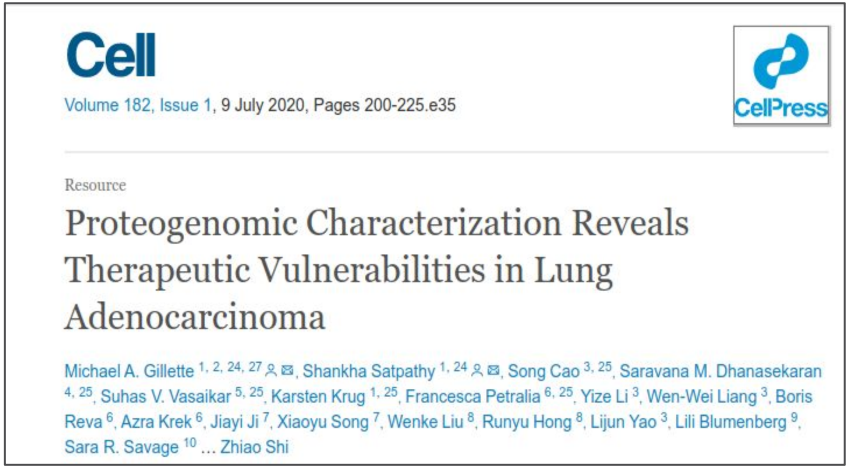
CellPress

Resource

Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma

David J. Clark [1, 32], Saravana M. Dhanasekaran [2, 32], Francesca Petralia [3, 32], Jianbo Pan [1, 32], Xiaoyu Song [4, 5, 32], Yingwei Hu [1, 32], Felipe da Veiga Leprevost [2, 32], Boris Reva [3, 32], Tung-Shing M. Lih [1, 32], Hui-Yin Chang [2], Weiping Ma [3], Chen Huang [6], Christopher J. Ricketts [7], Lijun Chen [1], Azra Krek [3], Yize Li [8], Dmitry Rykunov [3], Qing Kay Li [1] ... Zhidong Tu

Show more ∨



Cell

Volume 182, Issue 1, 9 July 2020, Pages 200-225.e35

CellPress

Resource

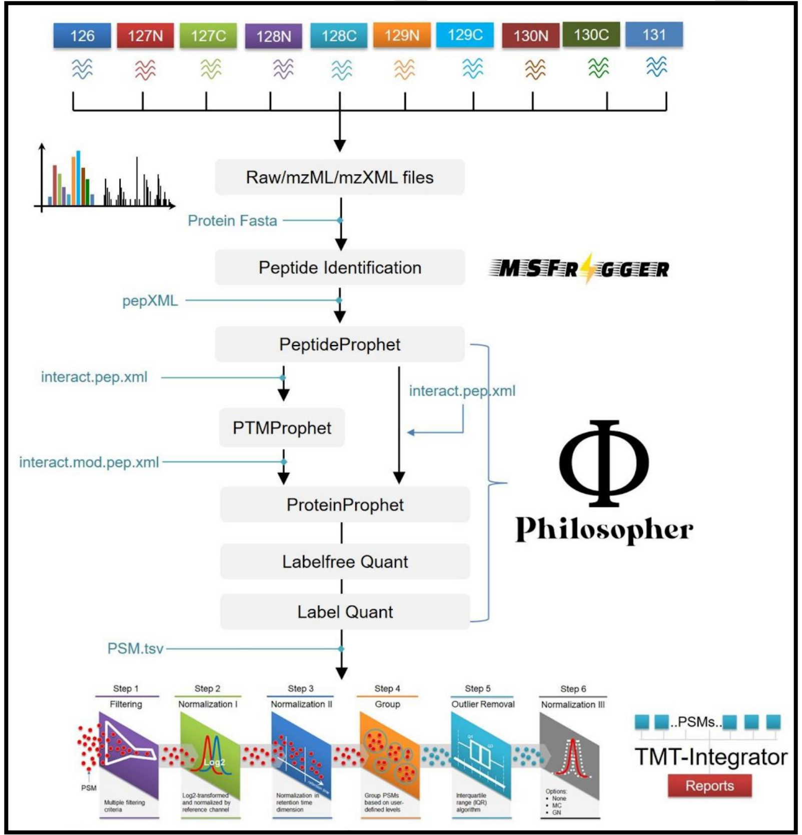
Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma

Michael A. Gillette [1, 2, 24, 27] ✉, Shankha Satpathy [1, 24] ✉, Song Cao [3, 25], Saravana M. Dhanasekaran [4, 25], Suhas V. Vasaikar [5, 25], Karsten Krug [1, 25], Francesca Petralia [6, 25], Yize Li [3], Wen-Wei Liang [3], Boris Reva [6], Azra Krek [6], Jiayi Ji [7], Xiaoyu Song [7], Wenke Liu [8], Runyu Hong [8], Lijun Yao [3], Lili Blumenberg [9], Sara R. Savage [10] ... Zhiao Shi

FragPipe Proteomics Pipeline

- Developed by Nesvizhskii lab at the University of Michigan
- **FragPipe**: a complete proteomics pipeline for comprehensive analysis of proteomics data which is powered by
  - **MSFragger**, an ultrafast peptide identification tool for mass spec-based proteomics
  - **Philosopher** toolkit, for post-processing MSFragger results
  - **TMT-Integrator**, a tools for integrating channel abundances from multiple TMT or iTRAQ-labeled samples and generating reports

https://github.com/Nesvilab/FragPipe



126 127N 127C 128N 128C 129N 129C 130N 130C 131

Raw/mzML/mzXML files

Protein Fasta

Peptide Identification

MSFRAGGER

pepXML

PeptideProphet

interact.pep.xml

interact.pep.xml

PTMProphet

interact.mod.pep.xml

ProteinProphet

Labelfree Quant

Label Quant

Φ Philosopher

PSM.tsv

Step 1 Filtering

PSM

Multiple filtering criteria

Step 2 Normalization I

Log2

Log2-transformed and normalized by reference channel

Step 3 Normalization II

retention time

Normalization in retention time dimension

Step 4 Group

Group PSMs based on user-defined levels

Step 5 Outlier Removal

Q1 Q3

Interquartile range (IQR) algorithm

Step 6 Normalization III

Options:
- None
- MC
- GN

..PSMs..

TMT-Integrator

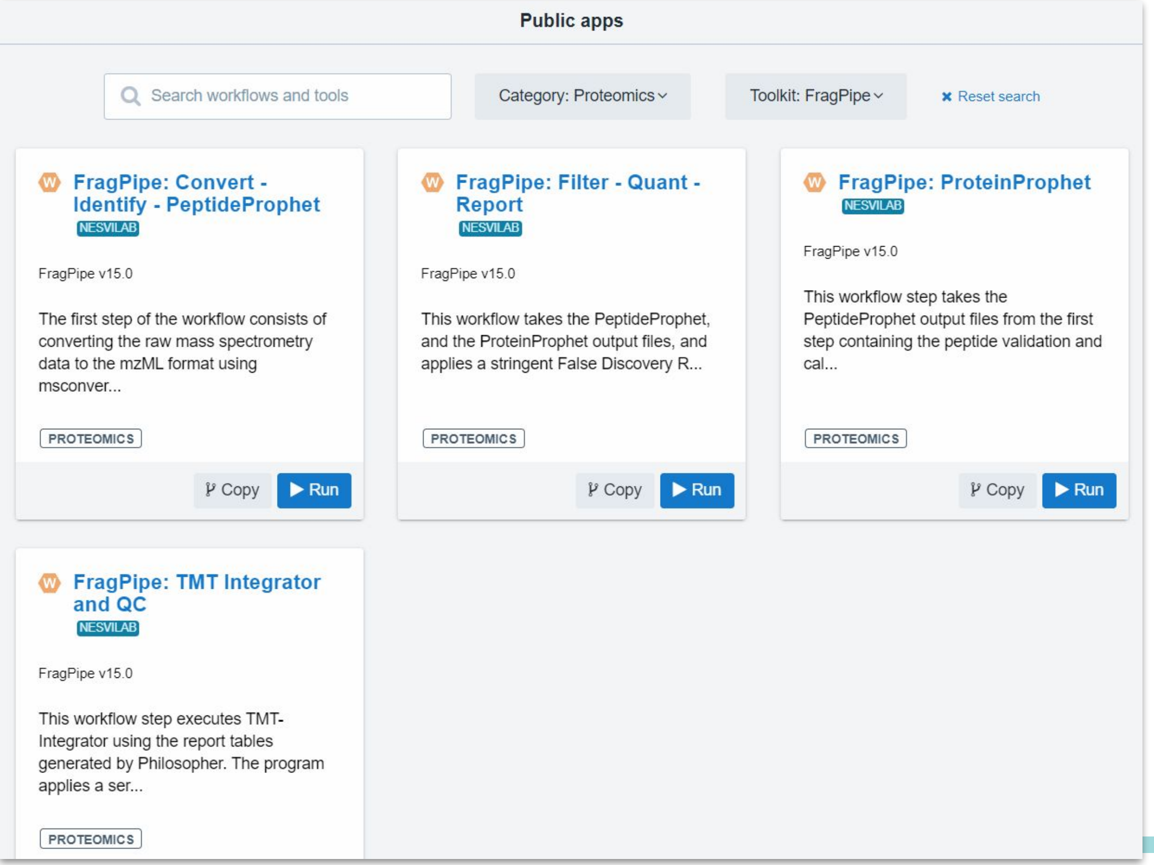
Reports

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

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



Public apps

FragPipe: Convert - Identify - PeptideProphet
NESVILAB

FragPipe v15.0

The first step of the workflow consists of converting the raw mass spectrometry data to the mzML format using msconver...

PROTEOMICS

FragPipe: Filter - Quant - Report
NESVILAB

FragPipe v15.0

This workflow takes the PeptideProphet, and the ProteinProphet output files, and applies a stringent False Discovery R...

PROTEOMICS

FragPipe: ProteinProphet
NESVILAB

FragPipe v15.0

This workflow step takes the PeptideProphet output files from the first step containing the peptide validation and cal...

PROTEOMICS

FragPipe: TMT Integrator and QC
NESVILAB

FragPipe v15.0

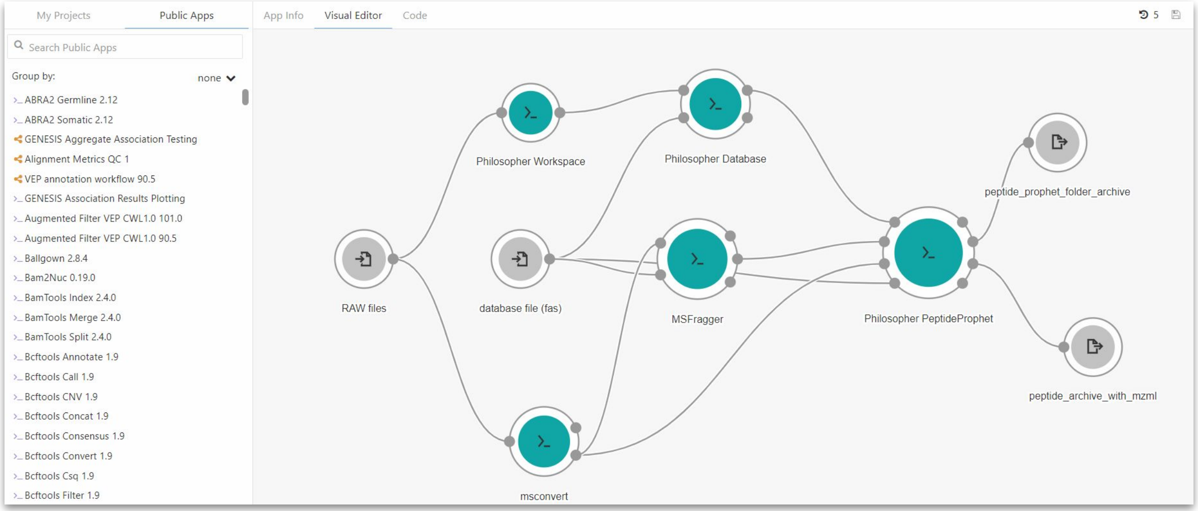
This workflow step executes TMT-Integrator using the report tables generated by Philosopher. The program applies a ser...

PROTEOMICS

Bring Your Own Tools & Tailor new Pipelines in 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



My Projects | Public Apps

App Info | Visual Editor | Code

Search Public Apps

Group by: none

- ABRA2 Germline 2.12
- ABRA2 Somatic 2.12
- GENESIS Aggregate Association Testing
- Alignment Metrics QC 1
- VEP annotation workflow 90.5
- GENESIS Association Results Plotting
- Augmented Filter VEP CWL1.0 101.0
- Augmented Filter VEP CWL1.0 90.5
- Ballgown 2.8.4
- Bam2Nuc 0.19.0
- BamTools Index 2.4.0
- BamTools Merge 2.4.0
- BamTools Split 2.4.0
- Bcftools Annotate 1.9
- Bcftools Call 1.9
- Bcftools CNV 1.9
- Bcftools Concat 1.9
- Bcftools Consensus 1.9
- Bcftools Convert 1.9
- Bcftools Csq 1.9
- Bcftools Filter 1.9

docker

COMMON WORKFLOW LANGUAGE

Rabix

[Reproducible Analysis for Bioinformatics]

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

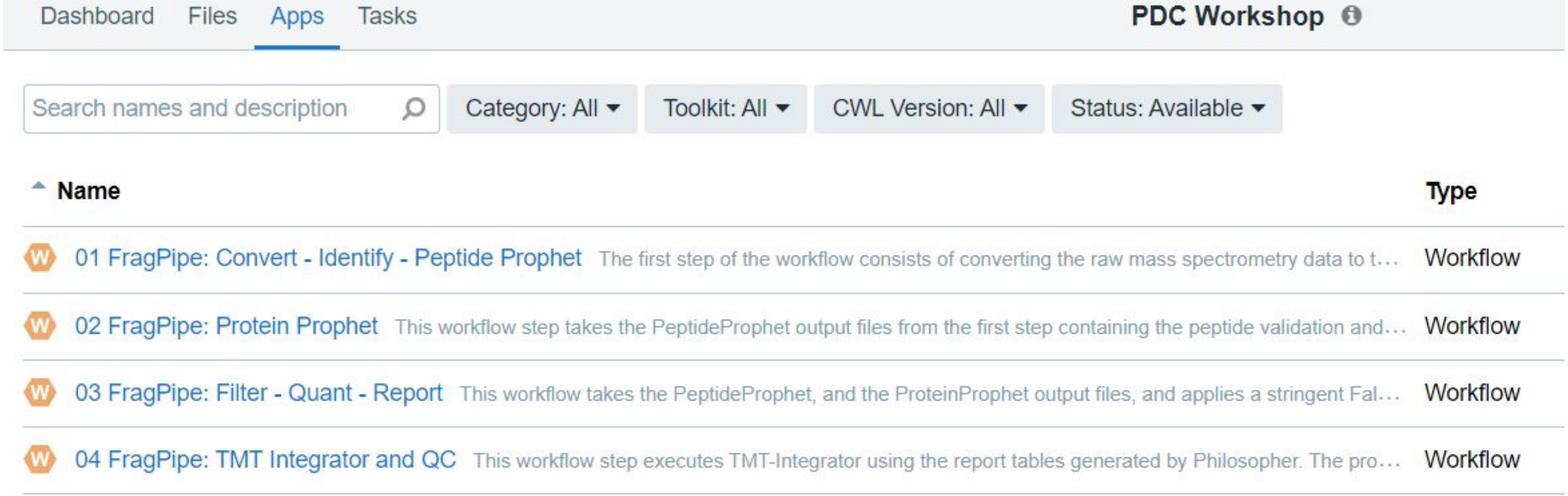
- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Analyze

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

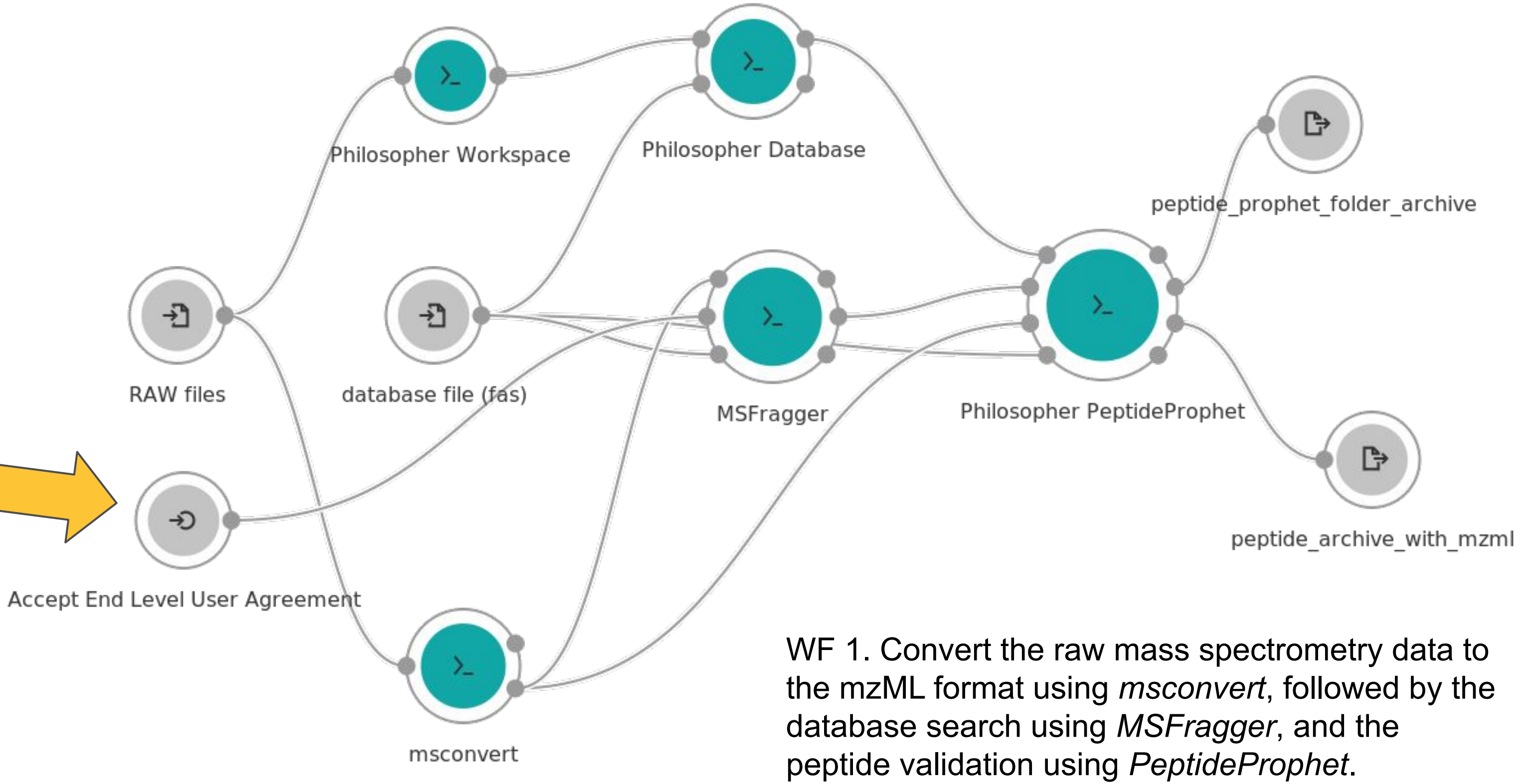
FragPipe Workflows



Name		Type
W 01 FragPipe: Convert - Identify - Peptide Prophet	The first step of the workflow consists of converting the raw mass spectrometry data to t...	Workflow
W 02 FragPipe: Protein Prophet	This workflow step takes the PeptideProphet output files from the first step containing the peptide validation and...	Workflow
W 03 FragPipe: Filter - Quant - Report	This workflow takes the PeptideProphet, and the ProteinProphet output files, and applies a stringent Fal...	Workflow
W 04 FragPipe: TMT Integrator and QC	This workflow step executes TMT-Integrator using the report tables generated by Philosopher. The pro...	Workflow

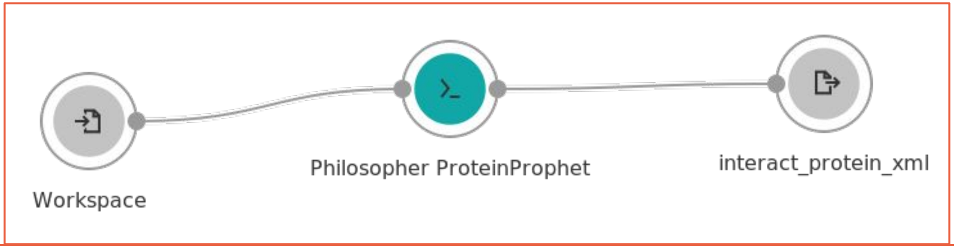
FragPipe - 4 Workflows

1


Philosopher Workspace
Philosopher Database
peptide_prophet_folder_archive
RAW files
database file (fas)
MSFragger
Philosopher PeptideProphet
peptide_archive_with_mzml
Accept End Level User Agreement
msconvert

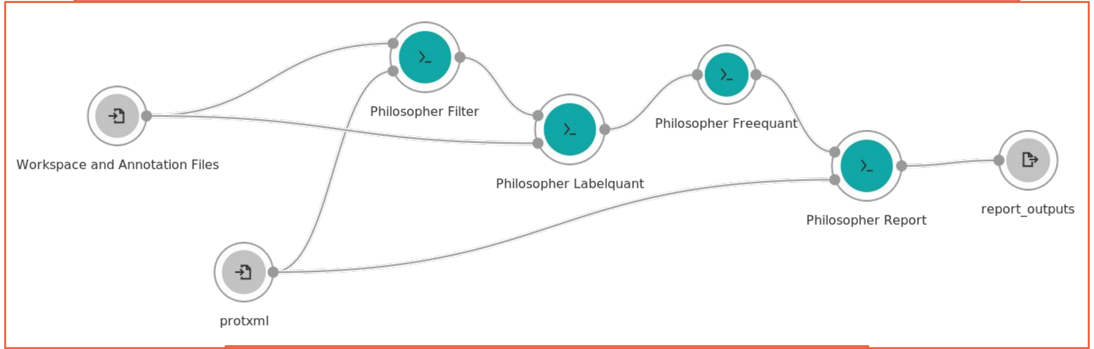
WF 1. Convert the raw mass spectrometry data to the mzML format using *msconvert*, followed by the database search using *MSFragger*, and the peptide validation using *PeptideProphet*.

2



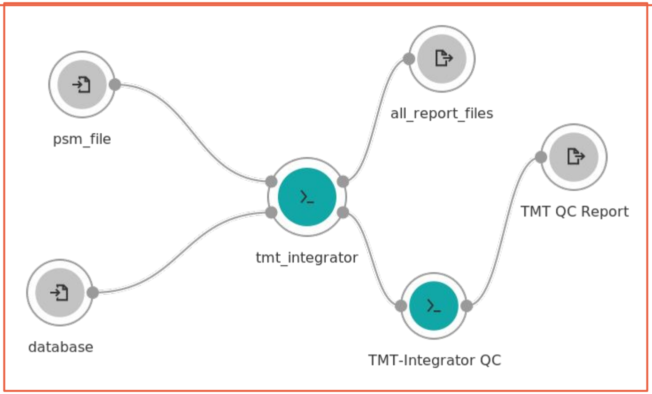
WF 2. Takes the PeptideProphet output files from the first step containing the peptide validation and calculates the protein inference using ProteinProphet.

3



WF 3. takes the PeptideProphet, and the ProteinProphet output files, and applies a stringent False Discovery Rate (FDR) filtering. Peptide and proteins are filtered individually at 1% FDR. The high-quality PSMs, peptides, and proteins are then quantified using a label-free algorithm that uses the apex peak intensity as a measurement. Finally, the isobaric tags are quantified and annotated with the correct sample labels.

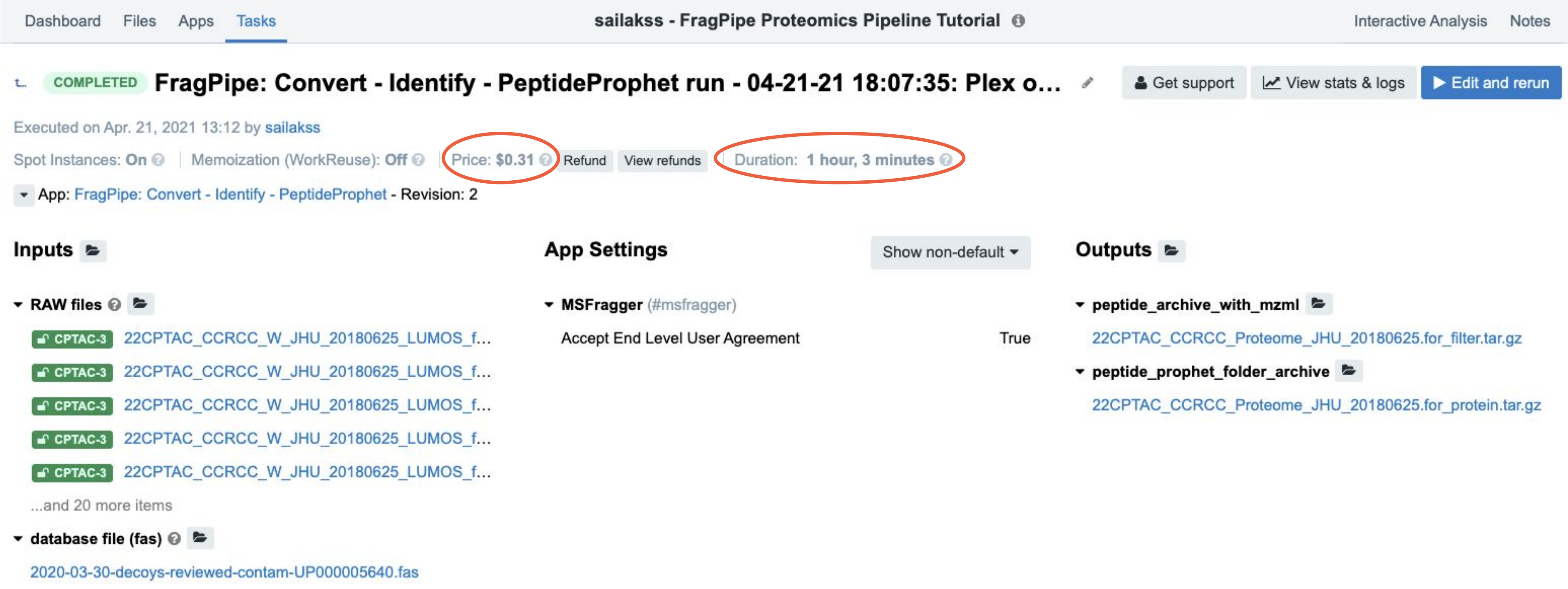
4



WF 4. Executes TMT-Integrator using the report tables generated by Philosopher. The program applies a series of statistical filters, and high-quality thresholds to filter the data. Summary report tables are created containing peptides, proteins, genes, and phosphosites (only for phospho-enriched data sets).

Features to control cloud costs

Monitor analysis at each step in **macro** or micro levels



sailakss - FragPipe Proteomics Pipeline Tutorial

COMPLETED FragPipe: Convert - Identify - PeptideProphet run - 04-21-21 18:07:35: Plex o...

Executed on Apr. 21, 2021 13:12 by sailakss

Spot Instances: On | Memoization (WorkReuse): Off | Price: $0.31 | Duration: 1 hour, 3 minutes

App: FragPipe: Convert - Identify - PeptideProphet - Revision: 2

Inputs

RAW files

- CPTAC-3 22CPTAC_CCRCC_W_JHU_20180625_LUMOS_f...
- CPTAC-3 22CPTAC_CCRCC_W_JHU_20180625_LUMOS_f...
- CPTAC-3 22CPTAC_CCRCC_W_JHU_20180625_LUMOS_f...
- CPTAC-3 22CPTAC_CCRCC_W_JHU_20180625_LUMOS_f...
- CPTAC-3 22CPTAC_CCRCC_W_JHU_20180625_LUMOS_f...

...and 20 more items

database file (fas)

2020-03-30-decoys-reviewed-contam-UP000005640.fas

App Settings

MSFragger (#msfragger)

Accept End Level User Agreement True

Outputs

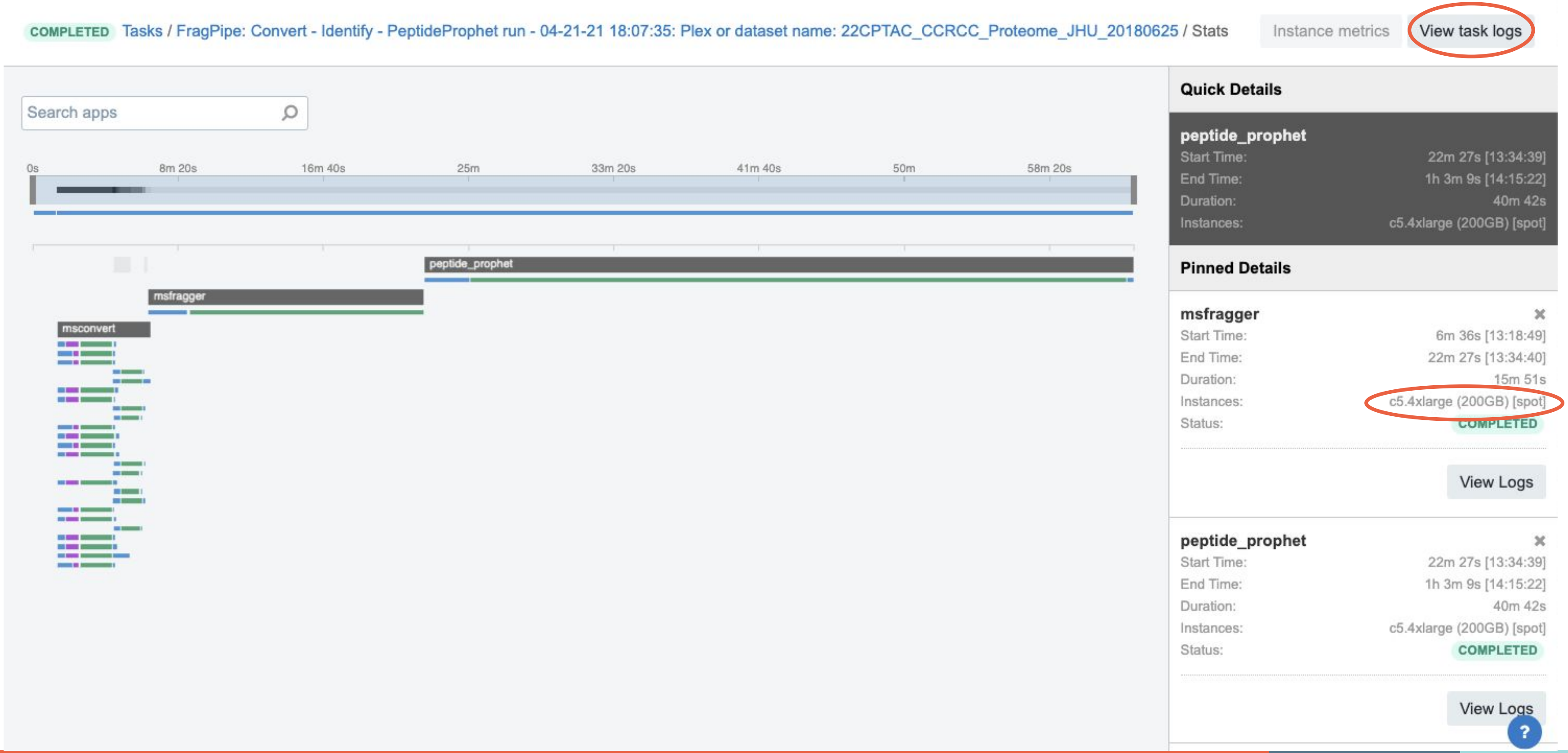
peptide_archive_with_mzml

22CPTAC_CCRCC_Proteome_JHU_20180625.for_filter.tar.gz

peptide_prophet_folder_archive

22CPTAC_CCRCC_Proteome_JHU_20180625.for_protein.tar.gz

Monitor analysis at each step in macro or **micro** levels



COMPLETED Tasks / FragPipe: Convert - Identify - PeptideProphet run - 04-21-21 18:07:35: Plex or dataset name: 22CPTAC_CCRCC_Proteome_JHU_20180625 / Stats

Quick Details

peptide_prophet

Start Time:	22m 27s [13:34:39]
End Time:	1h 3m 9s [14:15:22]
Duration:	40m 42s
Instances:	c5.4xlarge (200GB) [spot]

Pinned Details

msfragger

Start Time:	6m 36s [13:18:49]
End Time:	22m 27s [13:34:40]
Duration:	15m 51s
Instances:	c5.4xlarge (200GB) [spot]
Status:	COMPLETED

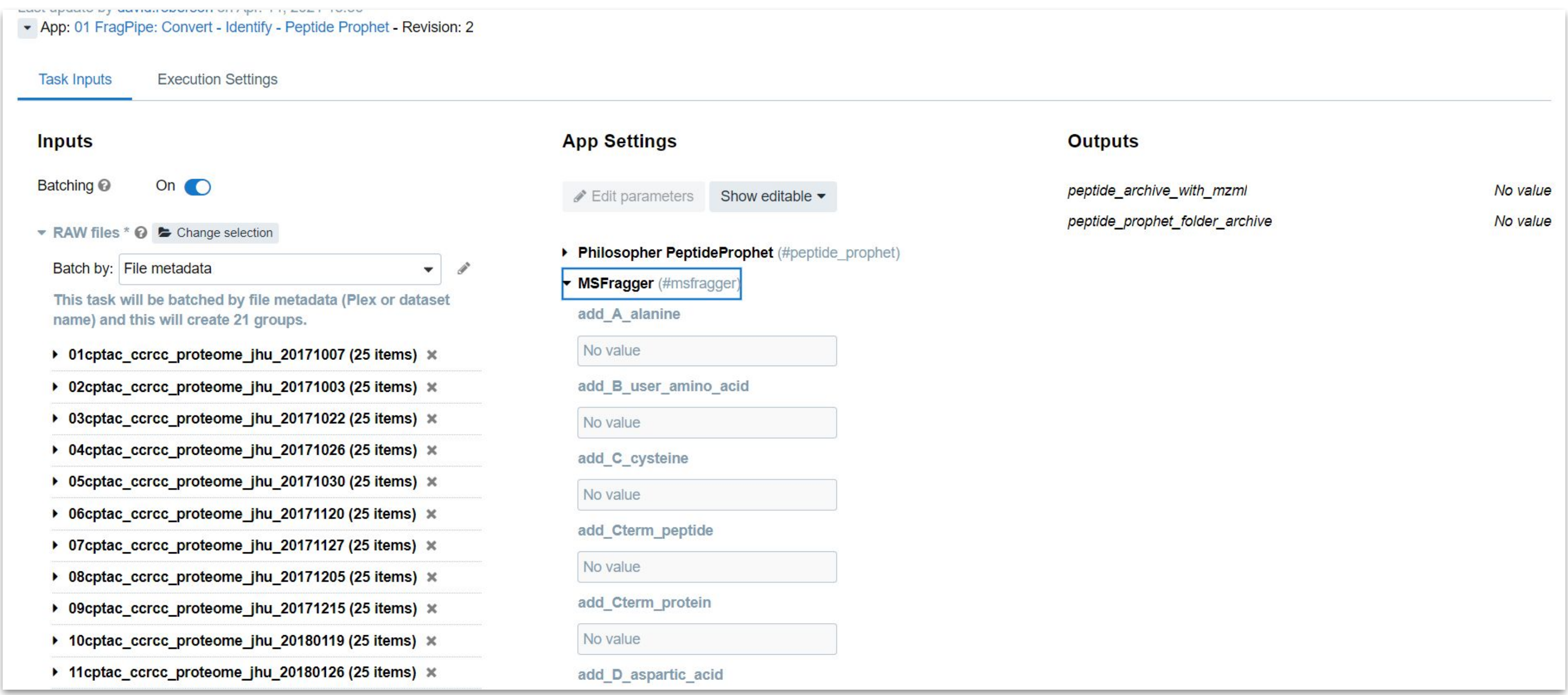
View Logs

peptide_prophet

Start Time:	22m 27s [13:34:39]
End Time:	1h 3m 9s [14:15:22]
Duration:	40m 42s
Instances:	c5.4xlarge (200GB) [spot]
Status:	COMPLETED

View Logs

Scale Up with batching by PDC file metadata



Last update by david.roberson on Apr. 11, 2021 10:30

App: 01 FragPipe: Convert - Identify - Peptide Prophet - Revision: 2

Inputs

Batching: On

RAW files *

Batch by: File metadata

This task will be batched by file metadata (Plex or dataset name) and this will create 21 groups.

- 01cptac_ccrcc_proteome_jhu_20171007 (25 items)
- 02cptac_ccrcc_proteome_jhu_20171003 (25 items)
- 03cptac_ccrcc_proteome_jhu_20171022 (25 items)
- 04cptac_ccrcc_proteome_jhu_20171026 (25 items)
- 05cptac_ccrcc_proteome_jhu_20171030 (25 items)
- 06cptac_ccrcc_proteome_jhu_20171120 (25 items)
- 07cptac_ccrcc_proteome_jhu_20171127 (25 items)
- 08cptac_ccrcc_proteome_jhu_20171205 (25 items)
- 09cptac_ccrcc_proteome_jhu_20171215 (25 items)
- 10cptac_ccrcc_proteome_jhu_20180119 (25 items)
- 11cptac_ccrcc_proteome_jhu_20180126 (25 items)

App Settings

- Philosopher PeptideProphet (#peptide_prophet)
- MSFragger (#msfragger)

add_A_alanine: No value

add_B_user_amino_acid: No value

add_C_cysteine: No value

add_Cterm_peptide: No value

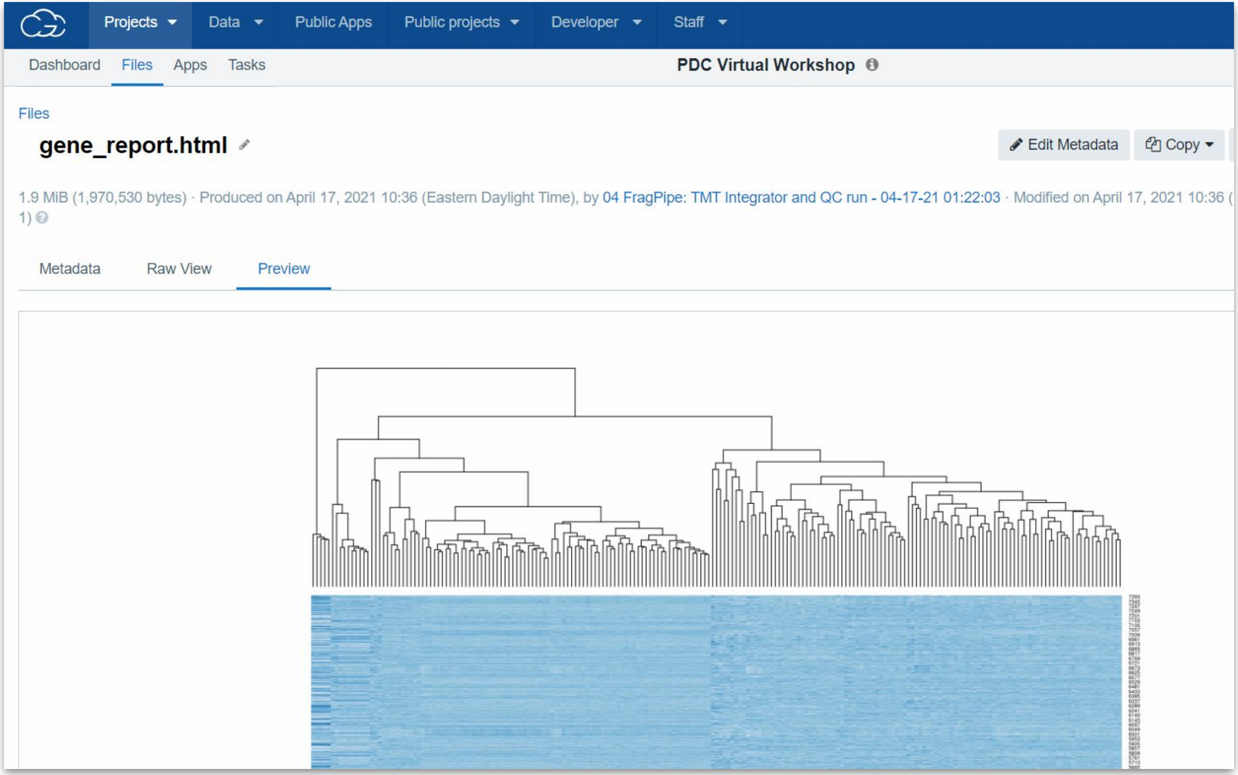
add_Cterm_protein: No value

add_D_aspartic_acid

Outputs

peptide_archive_with_mzml	*No value*
peptide_prophet_folder_archive	*No value*

Review Outputs without downloading



gene_report.html

1.9 MiB (1,970,530 bytes) · Produced on April 17, 2021 10:36 (Eastern Daylight Time), by 04 FragPipe: TMT Integrator and QC run - 04-17-21 01:22:03 · Modified on April 17, 2021 10:36 (
1)

Metrics of FragPipe Cloud vs. Local

Analysis of ccRCC whole cell lysate samples @Michigan on local server (non-cloud)

Computation:

- local server
- 56 cores (Xeon(R) CPU, 2.60GHz)
- 500GB RAM

Time for analysis:

27 hours, and a total of 16 GB RAM

Analysis of ccRCC samples on the CGC

Time for accessing the data:

Less than a minute; no need for downloads

Time for analysis on CGC:

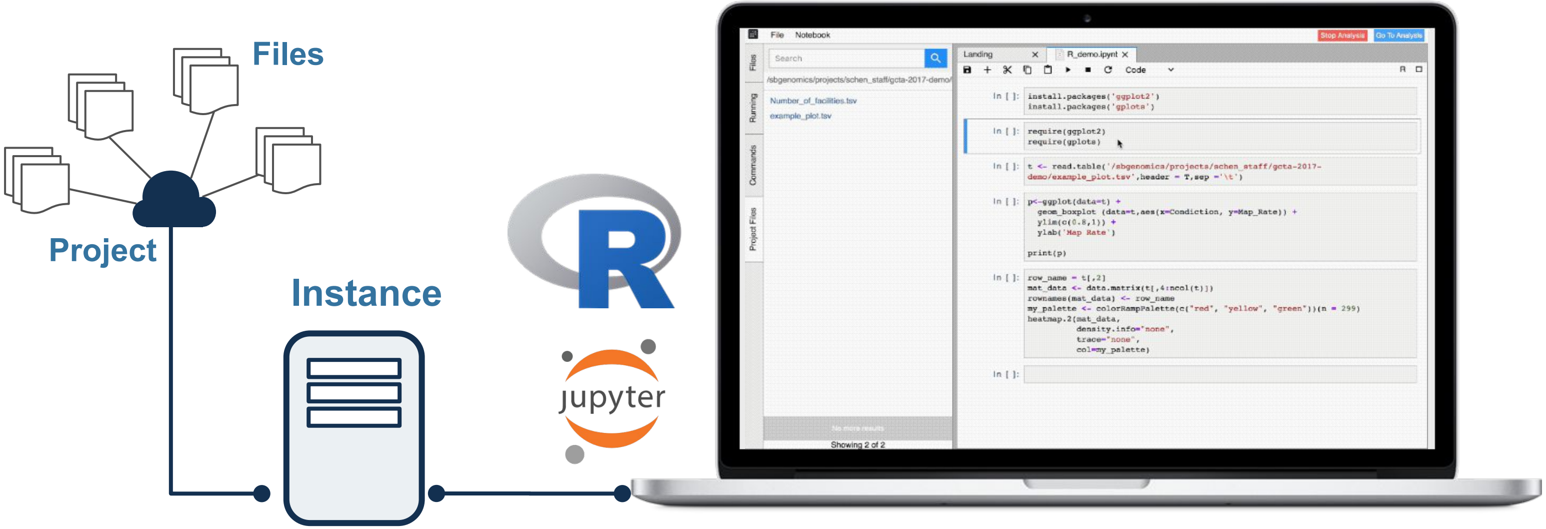
4.5 hours total. Up to 23 aws nodes in parallel @ 8 vCPU 16GB RAM, 48 aws nodes cumulative

Cost of analysis on CGC:

$0.49 per TMT plex, ~$12 for total analysis

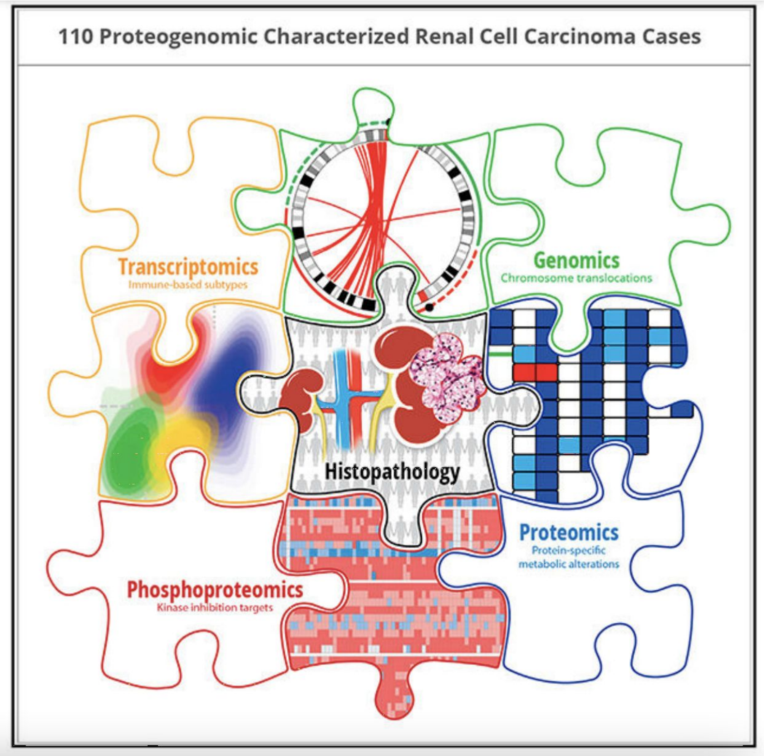
Powerful, collaborative, & reproducible interactive analysis

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)



```
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 [ ]:
```

Multi-omic data is critical for cancer research


110 Proteogenomic Characterized Renal Cell Carcinoma Cases
Transcriptomics
Immune-based subtypes
Genomics
Chromosome translocations
Histopathology
Proteomics
Protein-specific metabolic alterations
Phosphoproteomics
Kinase inhibition targets

Cancer is a complex disease!

Many research questions go beyond genomics data!

Different modalities should be examined to comprehensively understanding the full picture of a research question

As a researcher, it is essential to focus on the data. The CGC provides all tools for a multi omics analysis, so you can spend more time interpreting the results, and not configuring programs.

Clark et al., Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma, Cell. 2019 Oct 31; 179(4): 964–983.e31. DOI:https://doi.org/10.1016/j.cell.2019.10.007

Ongoing Projects

Adding new resources and capabilities to CGC
to support broader areas of cancer research

Epigenetic mechanisms for transcription regulation in cancer

Overall aim:

- Enable secondary analysis of epigenomics data on CGC, focusing on ATACseq and ChIPseq experiments
- Develop multi-omics analysis for data coming from different omics experiments (RNAseq, ATACseq, ChIPseq, WGBS, RRBS, proteomics)
- Analyze data from several publicly available repositories in order to characterize epigenetic markers in cancer.

Deliverables (publicly available on CGC soon):

- Workflows for ATACseq and ChIPseq analysis, based on ENCODE's specification
- Data Cruncher Interactive Analysis combining ATACseq, ChIPseq, WBGS, RNAseq and proteomics data
- Workflow for multi-omics analysis as a one-click solution

Multi-omics and imaging machine learning analysis

Overall aim:

Create a predictive model from both genomics and image features combined with the available clinical data in order to **predict therapy response among patients with cancer**.

- Build machine learning tools for processing multi-omics and imaging data from dbGAP/TCGA dataset.
- Use existing deep learning algorithms and libraries and adapt them for execution on CGC platform.

Deliverables (publicly available on CGC soon):

- Data Cruncher Interactive Analysis
- Tools and workflows for deep learning models adapted for imaging data, utilizing GPU instances and Tensorflow Python deep learning library.

Support and Resources

CGC Monthly Webinar Series

- Learn about CGC platform features that you can use in your projects.
- 4th Wednesday of each month at 2pm ET
- Upcoming webinar info, slides and recordings are available at: https://www.cancergenomicscloud.org/webinars

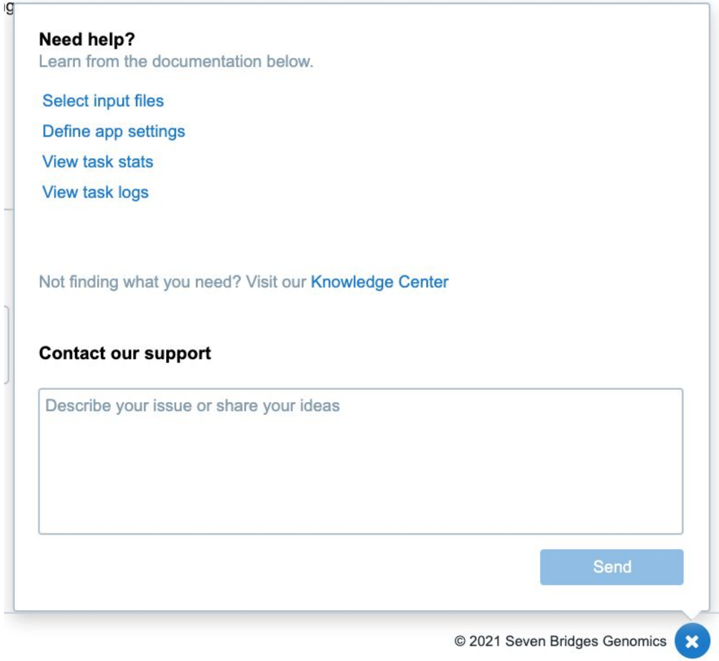
CGC Knowledge Center

https://docs.cancergenomicscloud.org/

Contact CGC Support: cgc@sevenbridges.com

Office Hours: Every week on Thursdays

https://www.cancergenomicscloud.org/officehours



Need help?
Learn from the documentation below.

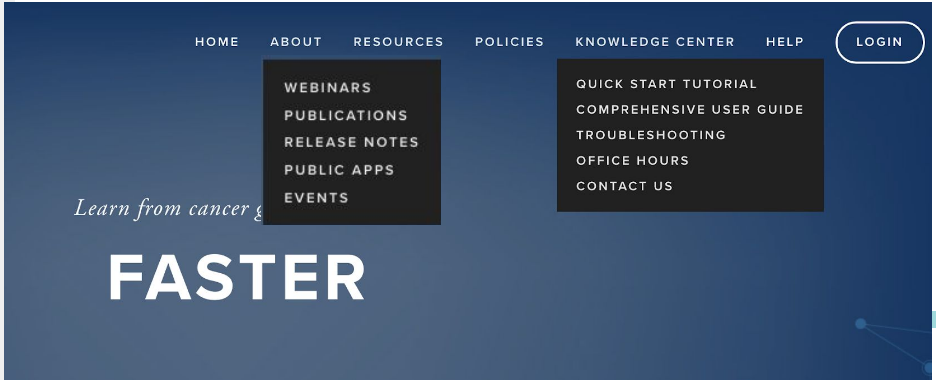
- Select input files
- Define app settings
- View task stats
- View task logs

Not finding what you need? Visit our Knowledge Center

Contact our support

Describe your issue or share your ideas

© 2021 Seven Bridges Genomics



In Summary

Data Access

Immediately access petabytes of ***Open and Controlled*** TCGA, CPTAC, 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.

CGC Webinar Series

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Towards best practice in cancer mutation detection with whole-genome and whole-exome sequencing

May 26, 2021
2 pm EDT/11 am PDT/7 pm GMT

Dr. Wenming Xiao
Lead Bioinformatics Scientist
Office of New Drug, Office of Oncological Diseases, Center for
Drug Evaluation and Research,
Food and Drug Administration

Acknowledgements

CGC Team

Manisha Ray

Jelena Radenkovic

Milos Stanojevic

Milos Trboljevac

Marko Tosic

Ana Stelkic

Dave Roberson

Sai Lakshmi Subramanian

Jack DiGiovanna

Brandi Davis-Dusenbery

& The Global Seven Bridges Team

UMichigan Team

Alexey I. Nesvizhskii

Felipe da Veiga Leprevost

Hui-Yin Chang

Guo Ci Teo

Fengchao Yu

PDC Team

Paul Rudnick

Rajesh Thangudu

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

NIH NATIONAL CANCER INSTITUTE
Proteomic Data Commons

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.

Questions?