

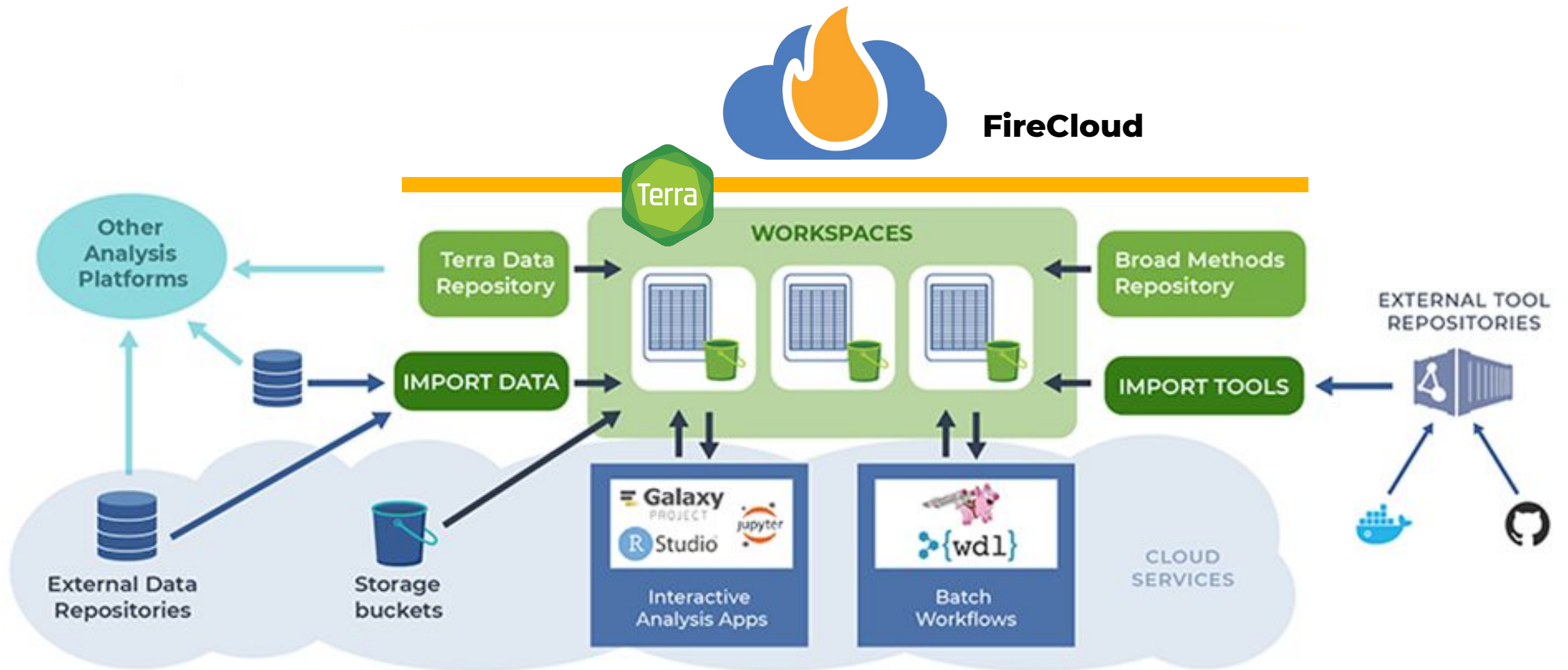


FireCloud

Access data, run analysis tools,
and collaborate on the cloud

<http://firecloud.terra.bio>

June 2021



FireCloud is powered by the Terra platform, a hub in the cloud data ecosystem

Welcome to FireCloud

FireCloud is a NCI Cloud Resource project powered by Terra for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**.

Find how-to's, documentation, video tutorials, and discussion forums [↗](#)

Already a FireCloud user? Learn what's new. [↗](#)

Learn more about the Cancer Research Data Commons and other NCI Cloud Resources [↗](#)

View Workspaces

Workspaces connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results easily and securely.



View Examples

Browse our gallery of showcase Workspaces to see how science gets done.



Browse Data

Access data from a rich ecosystem of data portals.



This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17X053 under Contract No. HHSN261200800001E



A rich catalog of data hosted by various organizations

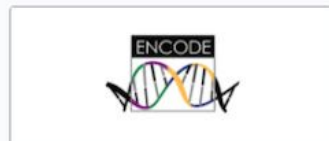


CMG presented by NHGRI AnVIL

The National Human Genome Research Institute funded the Centers for Mendelian Genomics (CMG) with the charge to discover as many genes underlying human Mendelian disorders as possible.

Participants: > 5,000

[BROWSE DATA](#)



ENCODE Project

The **E**ncyclopedia **O**f **D**NA **E**lements (ENCODE) project aims to delineate all functional elements encoded in the human genome. To this end, ENCODE has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification.

Donors: > 650 ; Files: > 158,000

[BROWSE DATA](#)



Broad Dataset Workspace Library

Search for datasets sequenced at the Broad Institute, or public datasets hosted at the Broad. Datasets are pre-loaded as workspaces. You can clone these, or copy data into the workspace of your choice.

Samples: > 158,629

[BROWSE DATASETS](#)



Framingham Heart Study Teaching Dataset

Since 1948, the Framingham Heart Study has been committed to identifying the common factors or characteristics that contribute to cardiovascular disease, over three generations of participants. This is a teaching dataset and may not be used for publication purposes.

Participants: 4,400

[BROWSE DATA](#)



Human Cell Atlas

The Human Cell Atlas (HCA) is made up of comprehensive reference maps of all human cells — the fundamental units of life — as a basis for understanding fundamental human biological processes and diagnosing, monitoring, and treating disease.

[BROWSE DATA](#)



Neuroscience Multi-Omic Archive

The Neuroscience Multi-Omic (NeMO) Archive is a data repository specifically focused on the storage and dissemination of omic data from the BRAIN Initiative and related brain research projects. NeMO operates in close partnership with the Broad Single Cell Portal, Terra, and the Brain Cell Data Center (BCDC).

Files: >= 210,000; Projects >= 5; Species >= 3

[BROWSE DATA](#)



Therapeutically Applicable Research to Generate Effective Treatments (TARGET) presented by the National Cancer Institute

The TARGET initiative employed comprehensive molecular characterization to determine the genetic changes that drive the initiation and progression of hard-to-treat childhood cancers. TARGET makes the data generated available to the research community with a goal to identify therapeutic targets and prognostic markers so that novel, more effective treatment strategies can be developed and applied.

Participants: 1,324

[BROWSE DATA](#)



The Cancer Genome Atlas Presented by the National Cancer Institute

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between the National Cancer Institute and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Participants: 11,000

[BROWSE DATA](#)



TopMed presented by NHLBI BioData Catalyst

Trans-Omics for Precision Medicine (TOPMed), sponsored by the National Institutes of Health's National Heart, Lung, and Blood Institute (NHLBI), is a program to generate scientific resources to enhance our understanding of fundamental biological processes that underlie heart, lung, blood, and sleep disorders (HLBS).

Participants: > 54,000

[BROWSE DATA](#)



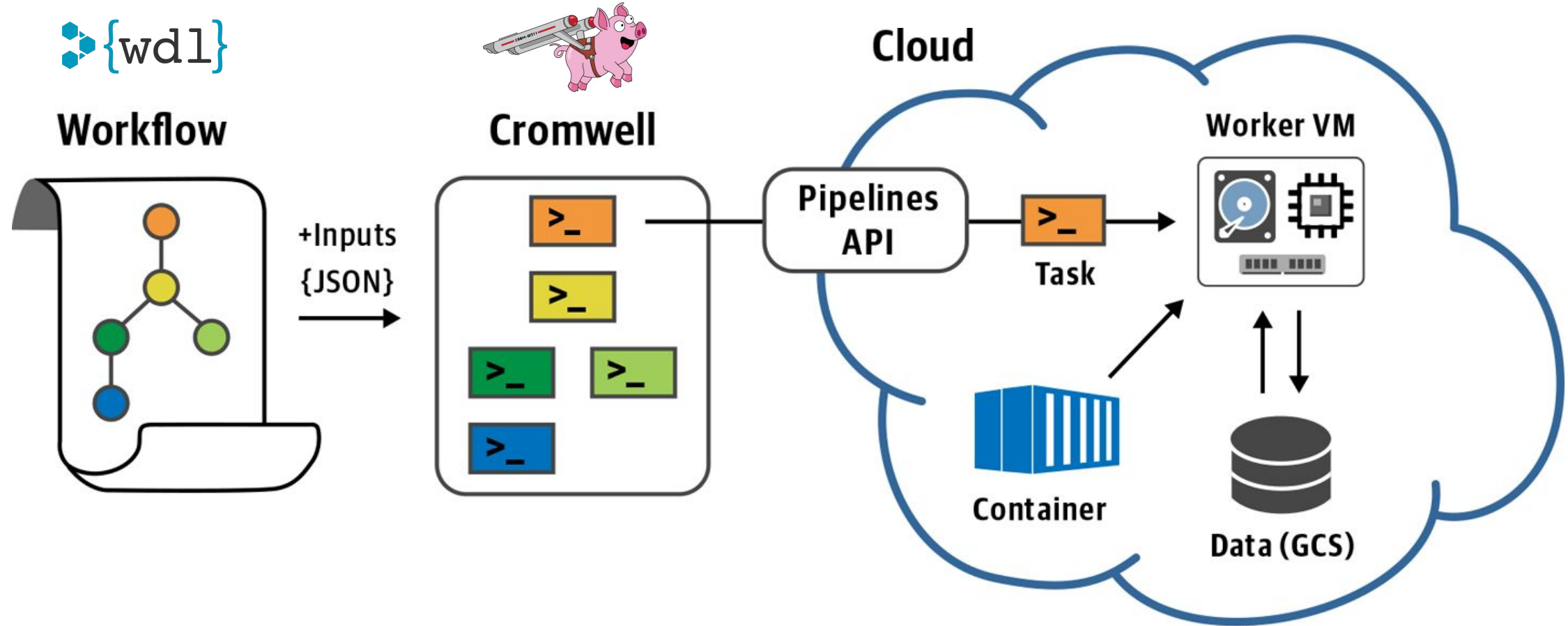
UK Biobank

UK Biobank is a national and international health resource with unparalleled research opportunities. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses. This Data Explorer is only available to specific early-access users at this time.

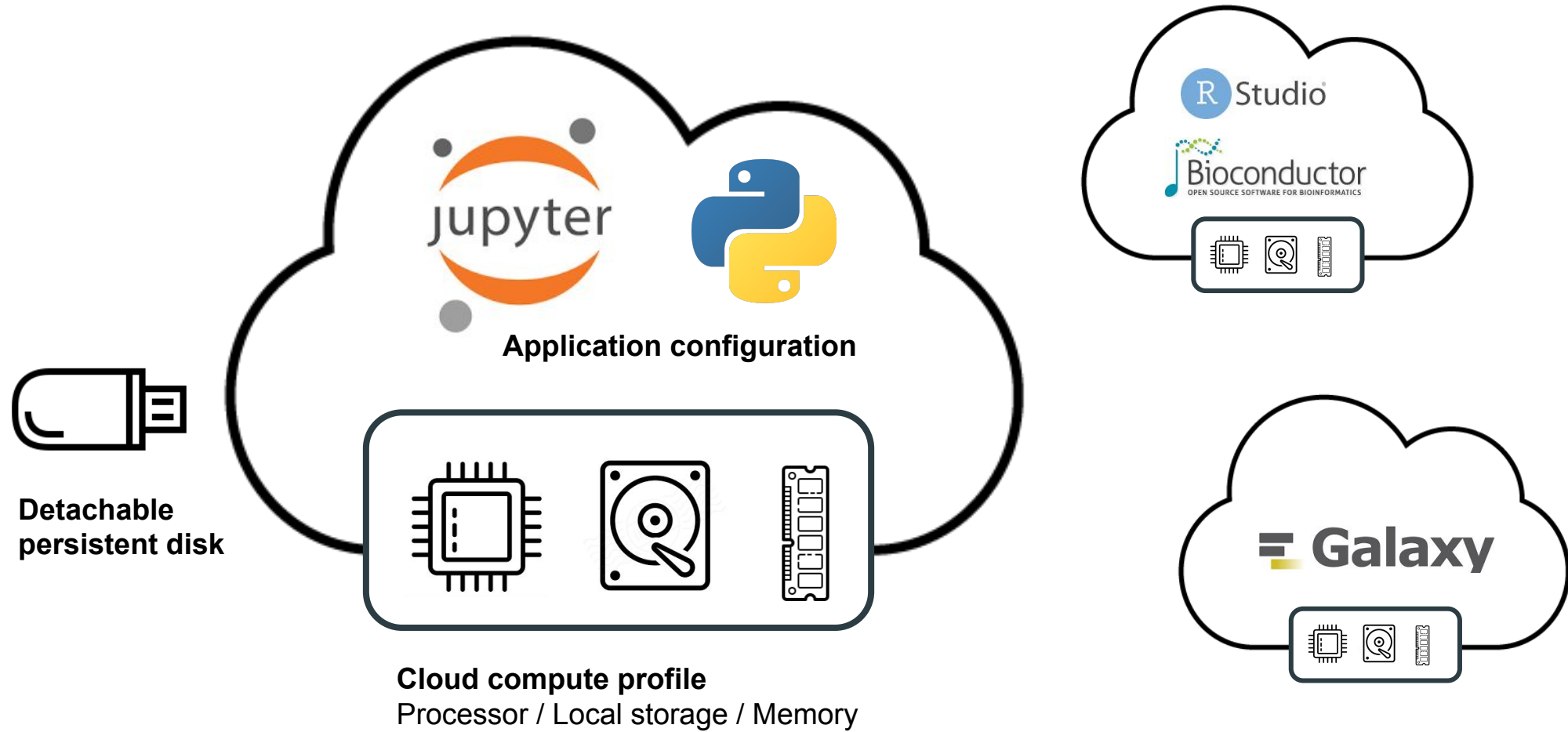
Participants: > 500,000

[BROWSE DATA](#)

A powerful workflow management system for big data processing



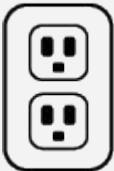
A highly flexible system of customizable cloud environments for interactive analysis



A set of foundational principles designed to benefit researchers



Open-source



Standards-based



Modular



Community-driven



**DATA
BIOSPHERE**



**Global Alliance
for Genomics & Health**



Built secure to work with highly sensitive data



- ✓ Human genomes
- ✓ Clinical data
- ✓ Many other data types

FISMA Moderate
FedRAMP Authorized

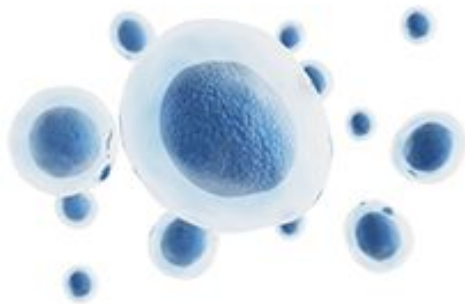


Thus equipped, FireCloud empowers the cancer research community at multiple levels



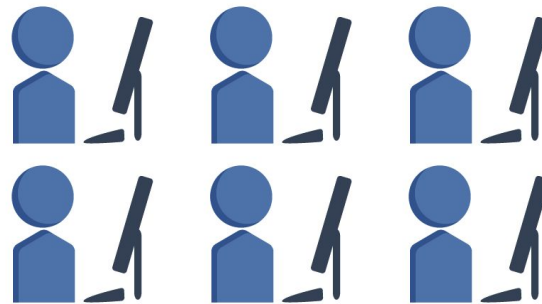
Data access & analysis

Insight generation



Publishable workspaces

Reproducible methods
Tool tutorials



Analysis portals

Suitable interfaces for
non-computational users



ABOUT THE WORKSPACE

This workspace de

Practice retrieving

create a panel-of-r

CNV_Somatic_Pai

What is Targ

The Therapeutical

Office of Cancer G

following quote is

The Therapeutic

molecular chara

childhood cance

therapeutic targ

applied.

Improved pedia

- Despite incr

ABOUT THE WORKSPACE

Practice accessing controlled-access TCGA data and run a simple MD5sum workflow on the data.

What is TCGA?


The Cancer Genome Atlas (TCGA), a landmark [cancer genomics](#) program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between the National Cancer Institute and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already lead to improvements in our ability to diagnose, treat, and prevent cancer, will remain [publicly available](#) for anyone in the research community to use.

(From TCGA project site on the [National Cancer Institute site](#))

The NCI's TCGA program involved a collaboration of scientists from 16 nations that has discovered nearly 10 million cancer-related mutations before the sequencing aspect ended after its decade long run (Ledford 2015). The program set out to characterize the molecular changes that occur within cancer cells and types. Additionally, it built a comprehensive model demonstrating a better understanding of protein pathways as well as their importance in the treatment of cancer. Their findings have improved the consolidation of data that helps diagnose, treat, and prevent cancer. This comprehensive understanding across cancer types has allowed researchers to use mutation, gene expression, and methylation data to create a

WORKSPACE INFORMATION

CREATION DATE 5/31/2019	LAST UPDATED 5/10/2021
SUBMISSIONS 0	ACCESS LEVEL Reader
GOOGLE PROJECT ID help-gatk 	




OWNERS

bshifaw@broadinstitute.org

TAGS

NCI TCGA

Google Bucket

Name: fc-0269d0d7-6919-4c6b-887... 
Location:  multi-region: US
[Open in browser](#) 



ABOUT THE WORKSPACE

This workspace shows you how to take a query result from the [NCI Genomic Data Commons](#) (GDC) data portal and use it as the input to a workflow (or Notebook) in FireCloud.

To get started with this tutorial, make sure you "clone" this workspace so you have your own copy to work with.

Note: Hold down the control button – or the command key on a Mac computer – to open any links below in a new tab.

Overview of the NCI Cancer Research Data Commons projects

The National Cancer Institute launched the [Cancer Research Data Commons](#) to provide researchers the means to accelerate discovery through the connection and harmonization of datasets with analytical tools in cloud native environments. From this, repositories of data known as data commons, were born. The number of data commons and the type of data being collected continue to grow, with new data commons launching almost yearly. Among the data commons, and their associated datasets, are:

- [Genomics Data Commons](#) (GDC): Supports hosting, standardization, and analysis of genomic, clinical, and biospecimen data. The GDC harmonizes raw sequencing data, identifies and applies bioinformatics methods for generating mutation calls, structural variants and other high-level data.
- [Proteomics Data Commons](#) (PDC): Advances the understanding of the role that proteins play in the cancer lifecycle. In-depth analysis of proteomic data allows the study of both how and why cancer develops and informs ways of tailoring treatment for patients.
- [Imaging Data Commons](#) (IDC): Connects researchers with publicly available cancer imaging data, often linked with other types of cancer data. IDC provides the tools to search and visualize cancer imaging data, define cohorts and use those cohorts to better understand the disease.

Other data commons within the CRDC, like the Integrated Canine Data Commons launched in 2020, will have similar workspaces when there is sufficient data and infrastructure to support dynamically querying and pointing the data to a workspace. The workspace in particular will focus on accessing data in GDC.

GDC data access

Account Linking

To use datasets from the Genomics Data Commons [GDC](#) in your FireCloud workspace you need to first link your GDC account with FireCloud. Go to your [Profile page](#) and look for "NCI CRDC Framework Services" under "IDENTITY & EXTERNAL SERVICES". Click the link to login using your eRA Commons ID (which you should have already if you

WORKSPACE INFORMATION

CREATION DATE 3/29/2021	LAST UPDATED 3/29/2021
SUBMISSIONS 1	ACCESS LEVEL Reader
GOOGLE PROJECT ID fc-product-d...	

OWNERS

akuan@broadinstitute.org
bshifaw@broadinstitute.org
boconnor@broadinstitute.org
cara@broadinstitute.org

TAGS

CRDC GDC Genomic Data Commons
NCI

Google Bucket

Name: fc-232922f4-e6a3-466a-8c39-...

Location: multi-region: US

[Open in browser](#)

Case 1: Research paper



Example of a typical study
done in part on FireCloud



Subtype Heterogeneity and Epigenetic Convergence in Neuroendocrine Prostate Cancer (NEPC)

Paloma Cejas *et al.*, <https://doi.org/10.1101/2020.09.13.291328>

- ▶ **Goal:** Understand molecular basis of NEPC in order to design more efficient targeted therapeutic options
 - ▶ Neuroendocrine prostate cancer (NEPC)
 - ▶ High grade tumors with aggressive clinical behavior and a poor prognosis
 - ▶ Treatment emergent phenotype from prostatic adenocarcinomas after anti-AR therapy
 - ▶ Improvements in anti-AR therapy has increased prevalence of NEPCs
- ▶ **Approach:** Epigenetic profiling of a range of Neuroendocrine Carcinomas to understand how the common phenotype is maintained across tumor types



The analysis involved a wide range of data types and computational tools




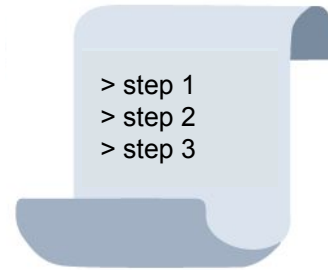
Data Types

Whole exome sequence (WES)
Epigenetic profiling
(ATAC-seq, CHIP-seq, RNA-seq)


Data Origins

Tissue samples from patients

 TCGA data (PRAD & LUAD)



Automated workflows

 WES variant calling & annotation
(Mutect, VEP, Oncotator)

Single-cell ATAC-seq & RNA-seq
data processing (CellRanger)

Bulk RNAseq
data processing (VIPER)

Bulk ATAC-seq & CHIP-seq
data processing (ChiLin)



Interactive analysis

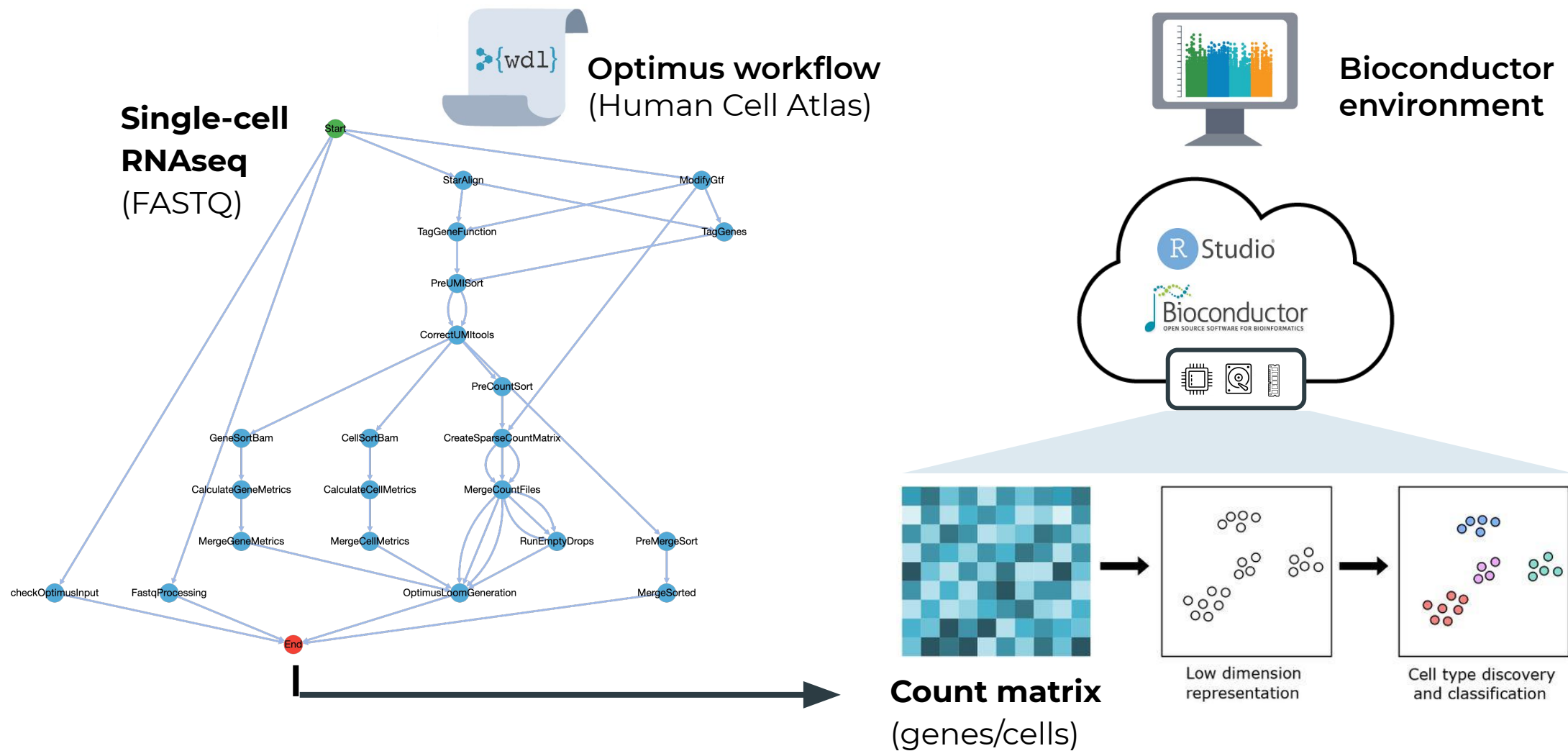
R tools
Bioconductor, Seurat

Python tools
MACS2

Online services
Cistrome Data Browser

Other visualization
IGV

Recent updates increase range of analyses that can be done in FireCloud




Done today, the majority of the work could be done in FireCloud

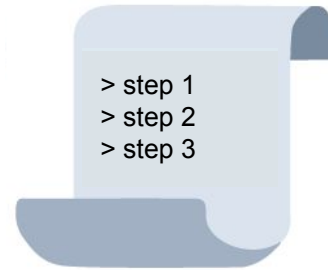


Data Types





Whole exome sequence (WES)
Epigenetic profiling
(ATAC-seq, CHIP-seq, RNA-seq)

Data Origins

Tissue samples from patients
 TCGA data (PRAD & LUAD)






Automated workflows

-  WES variant calling & annotation
(Mutect, VEP, Oncotator)
-  Single-cell ATAC-seq & RNA-seq
data processing (CellRanger)
-  Bulk RNA-seq
data processing (VIPER)
-  Bulk ATAC-seq & CHIP-seq
data processing (ChiLin)



Interactive analysis

-  *R tools*
Bioconductor, Seurat
-  *Python tools*
MACS2
- Online services*
Cistrome Data Browser
-  *Other visualization*
IGV

Case 2: PANOPLY



Example of an analysis
framework built on FireCloud



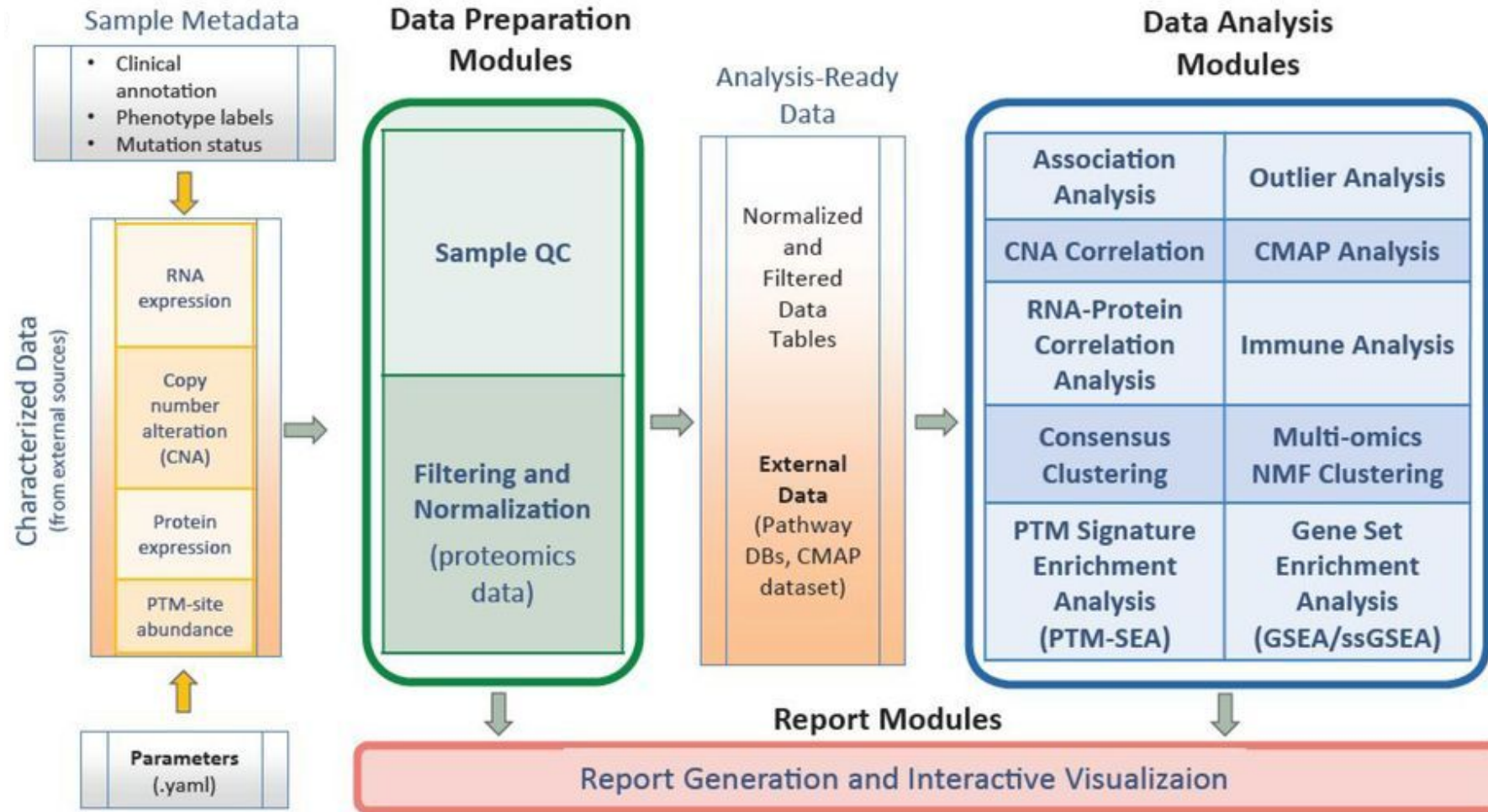
PANOPLY: A cloud-based platform for automated and reproducible proteogenomic data analysis

D.R. Mani *et al.*, <https://doi.org/10.1101/2020.12.04.410977>
<https://terra.bio/panoply-framework-for-cancer-proteogenomics/>

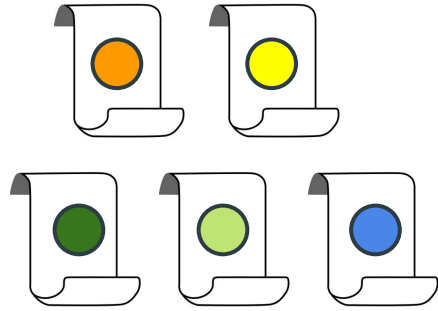
- ▶ **Goal:** Make it easier to transform multi-omic data from cancer samples into biologically meaningful and interpretable results
 - ▶ Proteogenomics = integrative analysis of genomic, transcriptomic, proteomic and post-translational modification data produced by high-throughput sequencing and mass spectrometry-based proteomics
- ▶ **Approach:** Standardize & make available a comprehensive collection of algorithms from CPTAC landmark proteogenomic studies
 - ▶ CPTAC = Clinical Proteomic Tumor Analysis Consortium



Overview of the PANOPLY framework



PANOPLY framework implementation: Workspaces & associated resources



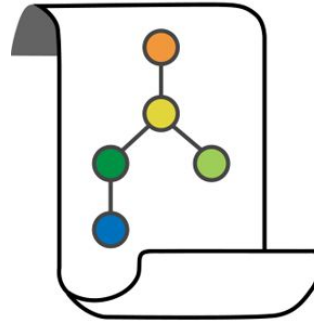
Modules workspace

Separate workflow per module

Maximum flexibility

Can compose new pipelines

Can add new modules

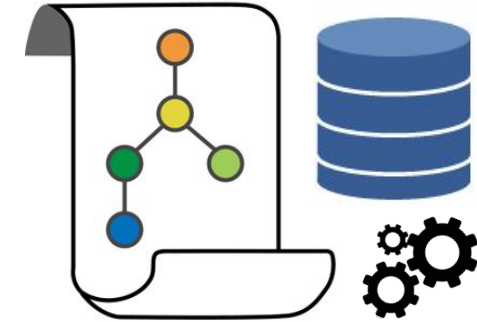


Pipelines workspace

Unified pipelines for standard use cases

Maximum reproducibility

Can run exactly as published



Tutorial workspace

Pre-run clone of the Pipelines WS

Includes preconfigured dataset (BRCA)

Job history shows execution results reproducing parts of Mertins *et al*, 2016*

+ **Jupyter Notebook in all WS providing step-by-step configuration and launch instructions**



* Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62 (2016)

Case 3: MOAlmanac



Example of an analysis portal
built on FireCloud



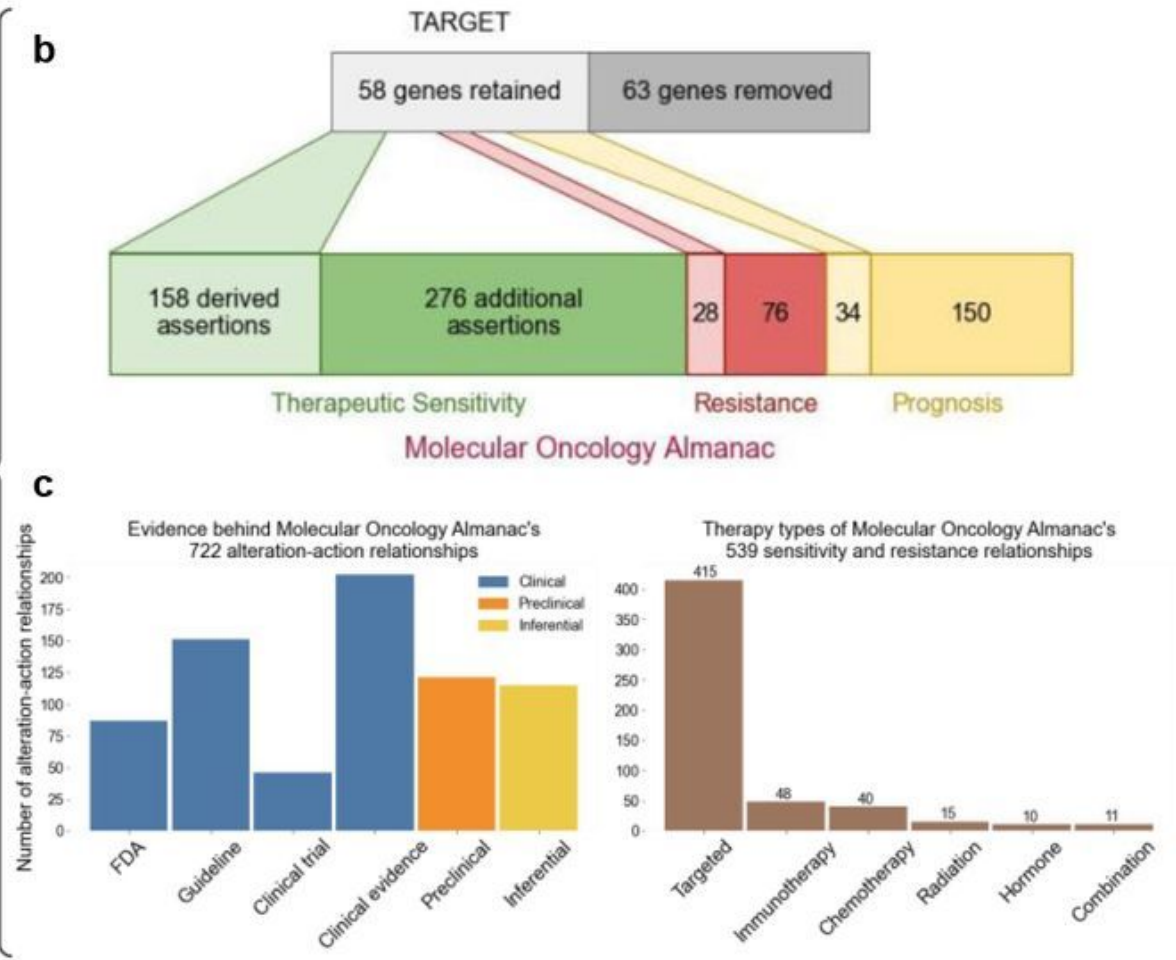
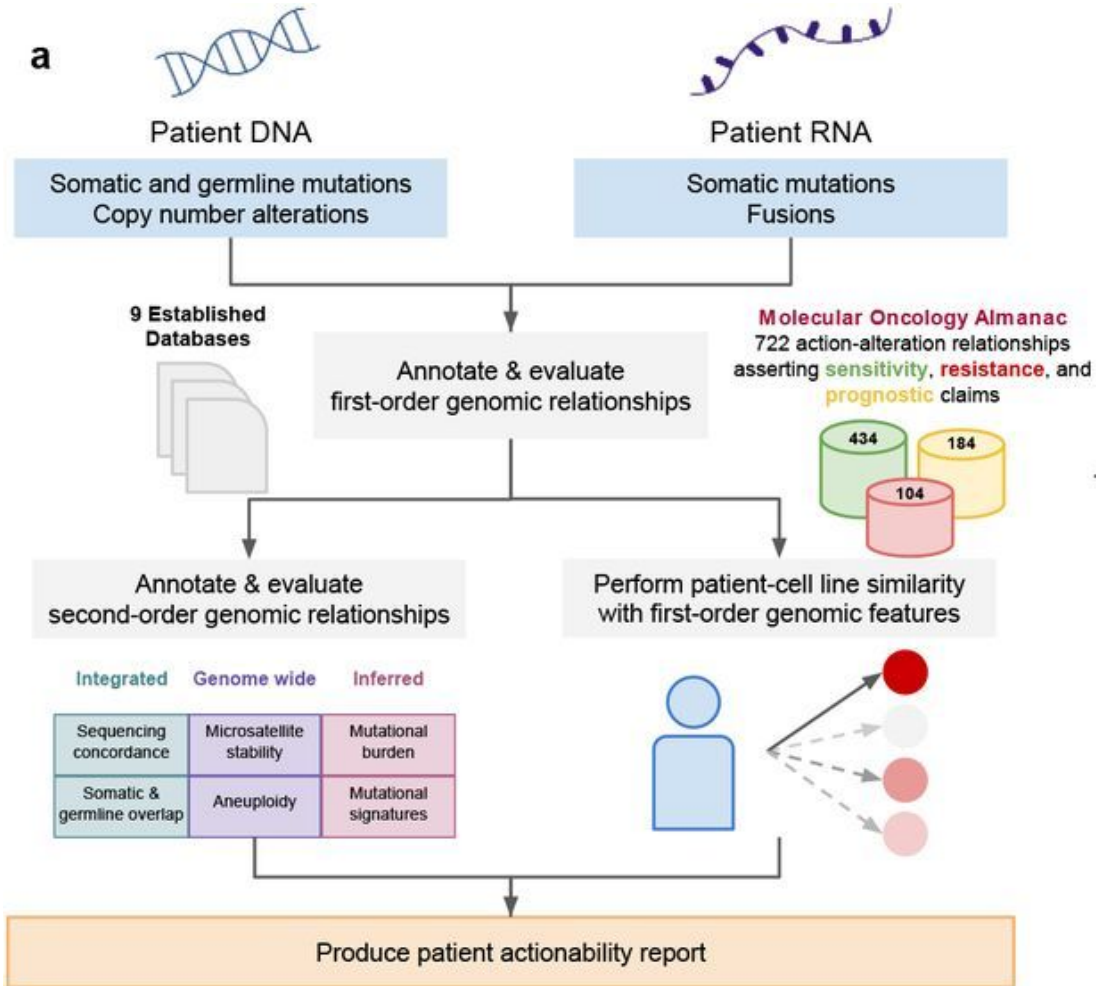
Clinical interpretation of integrative molecular profiles to guide precision cancer medicine

Brendan Reardon *et al.*, <https://doi.org/10.1101/2020.09.22.308833>

- ▶ **Goal:** Enable integrative interpretation of genomic and transcriptional cancer data for point-of-care treatment decision-making and translational hypothesis generation
 - ▶ Individual tumor molecular profiling is routinely used to detect single gene-variant (“first-order”) genomic alterations that may inform therapeutic actions
 - ▶ Interactions between such first-order events (e.g., somatic-germline) and global molecular features (e.g. mutational signatures) are increasingly associated with clinical outcomes, but these “second order” alterations are not yet generally accounted for in clinical interpretation algorithms and knowledge bases
- ▶ **Approach:** Provide a clinical interpretation method that evaluates individual patient molecular profiles based on associations between genes involved and cancer in published data sources
 - ▶ **Key challenge:** Make it accessible to non-computational researchers/physicians



Overview of the Molecular Oncology Almanac framework



MOAImanac framework implementation: Audience-specific channels

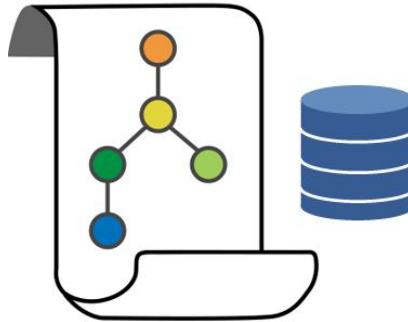


Knowledge base

Curated collection of putative alteration/action relationships based on published literature

Web browser interface + API

Data available for download



Pipeline workspace

MOAImanac method workflow

Includes preconfigured dataset

Can run as published



Interpretation portal

Simplified interface to run MOAImanac method with standard configuration

Suitable for non-computational researchers

Leverages the Firecloud API

Required fields

Add your de-identified sample name, tumor type from Oncotree, and select a billing project. The following inputs are required.

De-identified sample name:

Only letters, numbers, underscores, and are dashes allowed.

Tumor type: ⓘ

Terra billing project: ⓘ

Optional fields

Add a description or upload any combination of the file types below. The following inputs are optional.

Analysis description: ⓘ

Single nucleotide variants: ⓘ

 No file chosen

Suggested method: MuTect 1.0

Insertions or deletions: ⓘ

 No file chosen

Suggested method: Strelka

Somatic bases covered: ⓘ

 No file chosen

Suggested method: MuTect 1.0

Copy number alterations: ⓘ

 No file chosen

Suggested method: GATK CNV

Fusions from RNA: ⓘ

 No file chosen

Suggested method: Star Fusion

Single nucleotide variants (RNA): ⓘ

 No file chosen

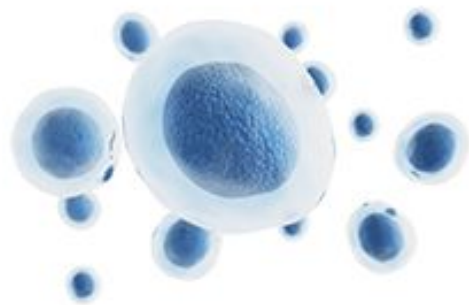
Suggested method: MuTect 1.0

Take-home: FireCloud empowers the cancer research community at multiple levels



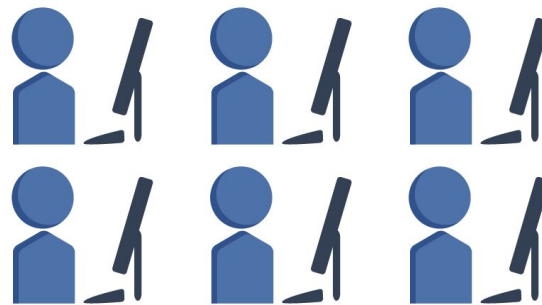
Data access & analysis

Insight generation



Publishable workspaces

Reproducible methods
Tool tutorials



Analysis portals

Suitable interfaces for
non-computational users





Get started today

<https://firecloud.terra.bio/>

<https://terra.bio/resources/getting-started>

User Guide and video tutorials:

<https://support.terra.bio>

[YouTube channel](#)

Contact Us:

support@terra.bio