



# A open-source, cloud-native platform for biomedical researchers

## One place to

- Access data
  - Run analysis tools
- Collaborate

Designed with built-in security so you can focus on science

### MODULAR

Comprised of functional components with well-specified interface

### COMMUNITY FOCUSED

Created by many groups to foster a diversity of ideas

### OPEN

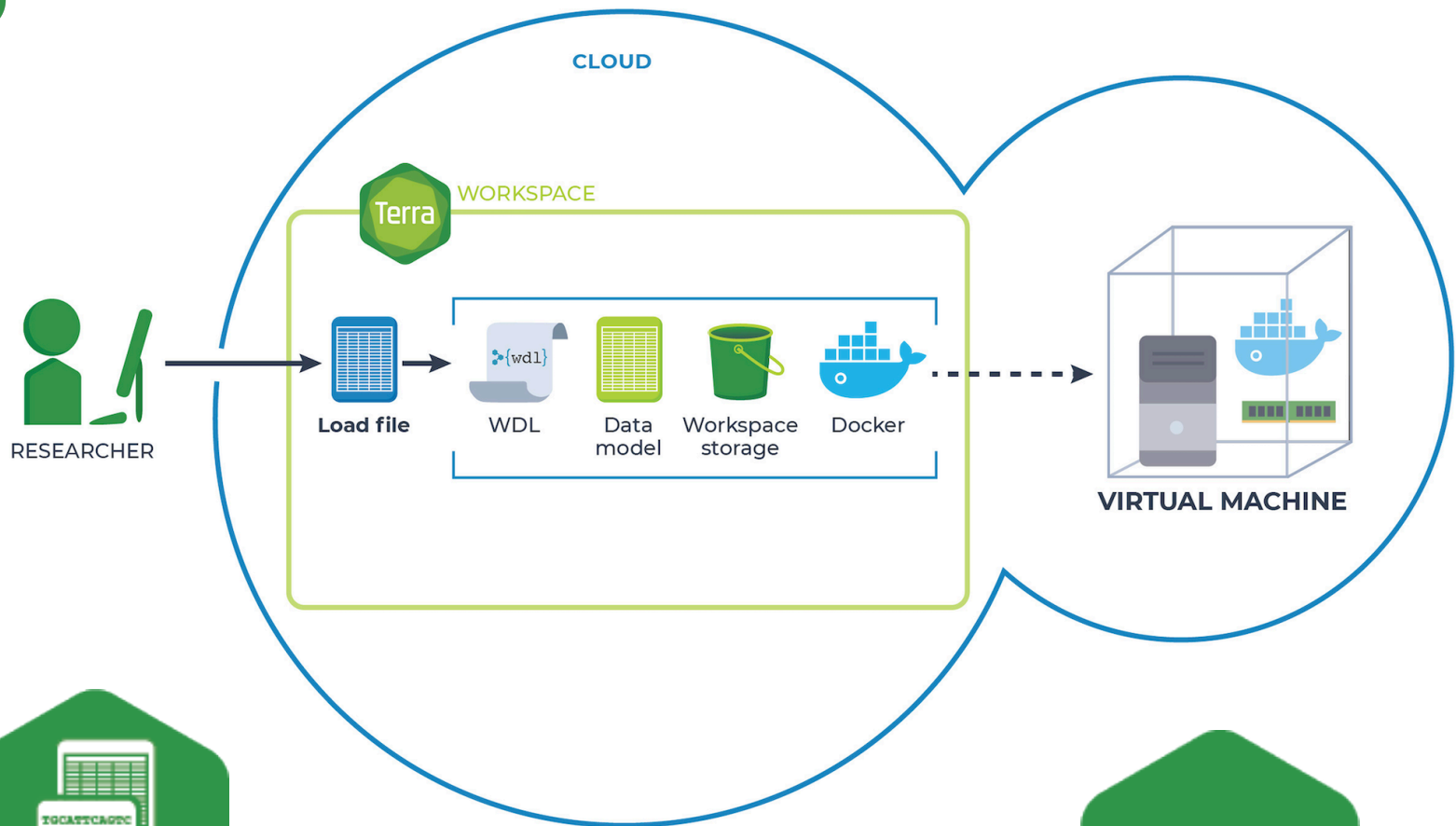
Open-source licenses, software, arch to enable extensibility

### STANDARDS BASED

Consistent with standards developed by coalitions such as GA4GH



# Core Functionality and organization





# Workspaces for secure analysis and collaboration

WORKSPACES Workspaces > fc-product-demo/SISG2019\_Terra\_Quickstart\_Workspace

Dashboard | DATA | NOTEBOOKS | WORKFLOWS | JOB HISTORY

## ABOUT THE WORKSPACE

This hands-on practice workspace helps users get started processing and analyzing data on Terra. You'll get familiar with the Data, Notebooks, Workflows, and Job History tabs in your workspace. You'll import BigQuery data from the Data Library and run a quick analysis in a notebook, and run your first workflow. The workspace is intended to be useful for a broad audience, streamlined so you can see results quickly with hands-on practice that will give you the foundation to start your own work on Terra.

Part 1 focuses on using Terra's Data Explorer to create cohorts of data and access and analyze the BigQuery data in a Jupyter notebook. Part 2 focuses on running a simple workflow, including linking genomics data in a Google bucket to the workspace Data table for access by the workflow, and configuring the workflow to run on your data.

Scroll down to the end for links to additional resources. Note that you will need to clone your own copy of this workspace to run the notebooks and workflows.

---

## Part 1: How to use Jupyter Notebooks to access and analyze data, including data in Terra's data library

### Notebooks overview

**Workspace Runtime**  
STOPPING (\$0.19 hr)

### WORKSPACE INFORMATION

CREATION DATE 7/9/2019	LAST UPDATED 7/15/2019
SUBMISSIONS 0	ACCESS LEVEL Owner
EST. \$/MONTH \$0.00	

### OWNERS

[jhajian@broadinstitute.org](mailto:jhajian@broadinstitute.org)

### TAGS

Add a tag

conversion x cramtobam x  
fastqtobam x validatebam x



# Integrated Data Library, associated Google bucket



**DATASETS** | SHOWCASE & TUTORIALS | CODE & TOOLS

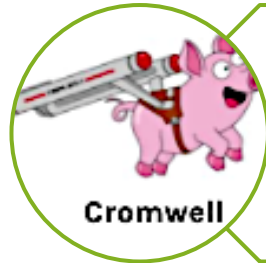
 <b>AMP Parkinson's Disease</b> The Accelerating Medicines Partnership (AMP) is a public-private partnership between the National Institutes of Health (NIH), multiple biopharmaceutical and life sciences companies, and non-profit organizations to identify... <a href="#">READ MORE</a> Participants: > 4,700 <a href="#">BROWSE DATA</a>	 <b>Baseline Health Study</b> <b>Baseline Health Study</b> is a longitudinal study that will collect broad phenotypic health data from approximately 10,000 participants, who will each be followed over the course of at least four years. The study is part of a broader effort designed to develop a well-defined reference, or "baseline," of health. Participants: > 1,500 <a href="#">BROWSE DATA</a>	 <b>ENCODE Project</b> The <b>Encyclopedia Of DNA Elements (ENCODE)</b> 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 <a href="#">BROWSE DATA</a>	 <b>FireCloud Dataset Library</b> 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 <a href="#">BROWSE DATASETS</a>
 <b>GTEx presented by NIH Commons</b> The Genotype-Tissue Expression (GTEx) Program established a data resource and tissue bank to study the relationship between genetic variation	 <b>Human Cell Atlas</b> 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	 <b>Nurses' Health Study</b> The Nurses' Health Study and Nurses' Health Study II are among the largest investigations into the risk factors for major chronic diseases in	 <b>TopMed presented by NIH Commons</b> Trans-Omics for Precision Medicine (TOPMed),



# Analysis Tools including both interactive and batch processing



Interactive development environment for working with notebooks, code and data. Supports 100 programming languages with 1.7M public notebooks available



Scientific workflow engine designed for simplicity & scalability. Supports WDL and CWL, the two workflow languages adopted by GA4GH



Open-source, scalable framework for exploring and analyzing genomic data. Hail can generate variant/sample annotations, and perform variant, gene-burden, eQTL association analyses



# Workflows to align, QC, call short variants per sample, joint-call across populations, and filter

Terra BETA WORKSPACES Workspaces > help-gatk/Somatic-CNVs-GATK4 > workflows > 2-CNV\_Somatic\_Pair

DASHBOARD DATA NOTEBOOKS **WORKFLOWS** JOB HISTORY

### 2-CNV\_Somatic\_Pair

Snapshot 1.3.0  
Source: [github.com/gatk-workflows/gatk4-somatic-cnvs/cnv\\_somatic\\_pair\\_workflow](https://github.com/gatk-workflows/gatk4-somatic-cnvs/cnv_somatic_pair_workflow)  
Synopsis:  
*No documentation provided*

- Process single workflow from files
- Process multiple workflows from:  [Select Data](#)  
all 2 pairs (will create a new set named "2-CNV\_Somatic\_Pair\_2019-07-15T14-02-49")
- Use call caching

SCRIPT .. **INPUTS** .. OUTPUTS .. [RUN ANALYSIS](#)

Show optional inputs

Task name	Variable	Type	Attribute
CNVSomaticPairWorkflow	common_sites	File	<input type="text" value="workspace.common_sites"/>
CNVSomaticPairWorkflow	gatk_docker	String	<input type="text" value="workspace.gatk_docker"/>



# Jupyter Notebooks for interactive visualization and analysis

*“An open-source web application that allows users to create and share documents containing live code, equations, visuals, and narrative text”*

## Interactive Analysis

- Visualize and iterate in real time
- Couple with BigQuery to analyze data of any size or format in real time

## Sharing

- Code and visualization tools, data, and narrative text in one package
- Lets your collaborators easily access and understand your results
- Open and interoperable standards



# Our Ecosystem Partners

Terra is product of  **BROAD**  
INSTITUTE and **verily**

We are privileged to be part of the  **DATA**  
**BIOSPHERE**

Our goal is to build momentum for an *open, compatible, and secure* approach to data within the larger research community.



# Acknowledgements

We are grateful to the  team for their support.

This walkthrough features a reproduction of the subset of analysis in:

Haas, ME et al. (2018) **Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure**. AJHG volume 103, issue 4, p461-473. [doi:10.1016/j.ajhg.2018.08.004](https://doi.org/10.1016/j.ajhg.2018.08.004)

This reproduction was developed in collaboration with members of the Kathiresan lab and specific thanks go to:

Mary Haas  
Krishna Aragam  
Sam Bryant  
James Pirucello  
Sek Kathiresan