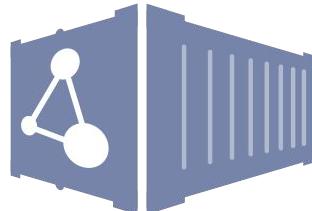


Reproducible Analysis in the Cloud with Dockstore & Terra

Beri Shifaw, Broad Institute
Louise Cabansay, UCSC Genomics Institute
Beth Sheets, UCSC Genomics Institute



AnVIL



Workshop Overview:

- Logistics (5 m)
- Overview of the Data Biosphere (10 m)
 - NHGRI's AnVIL & Interoperability across cloud research initiatives
- Introduction to Dockstore (20 m)
 - 'App store' for reproducible bioinformatics workflows
- Introduction to Terra (25 m)
 - User friendly workspaces for analysis in the cloud

Break (15 m)

- Integration of Dockstore and Terra: Analyzing COVID-19 viral genomes (60 m)
- Wrap up

Pre-workshop Requirements

If you don't have a Terra account setup already. During the break, you need to:

- 1) Create a Terra account now with a Google-backed email (see instructions from pre-workshop email).
- 2) Send that same email to Beth Sheets in a private message or email <esheets@ucsc.edu> so that you can be added to the workshop billing project.

Workshop Logistics

- **Please keep your mic on mute when you are not talking**
- **Ask questions in chat**

Members of the team will be available to answer questions in the chat box.
You can also privately message our support team: **Beri S., Beth S., Louise C., Mo H.**
- **Raise your hand feature**

If you would like to ask a question verbally, please use the raise your hand feature. Along the way we will pause to give time to those with raised hands as time permits.

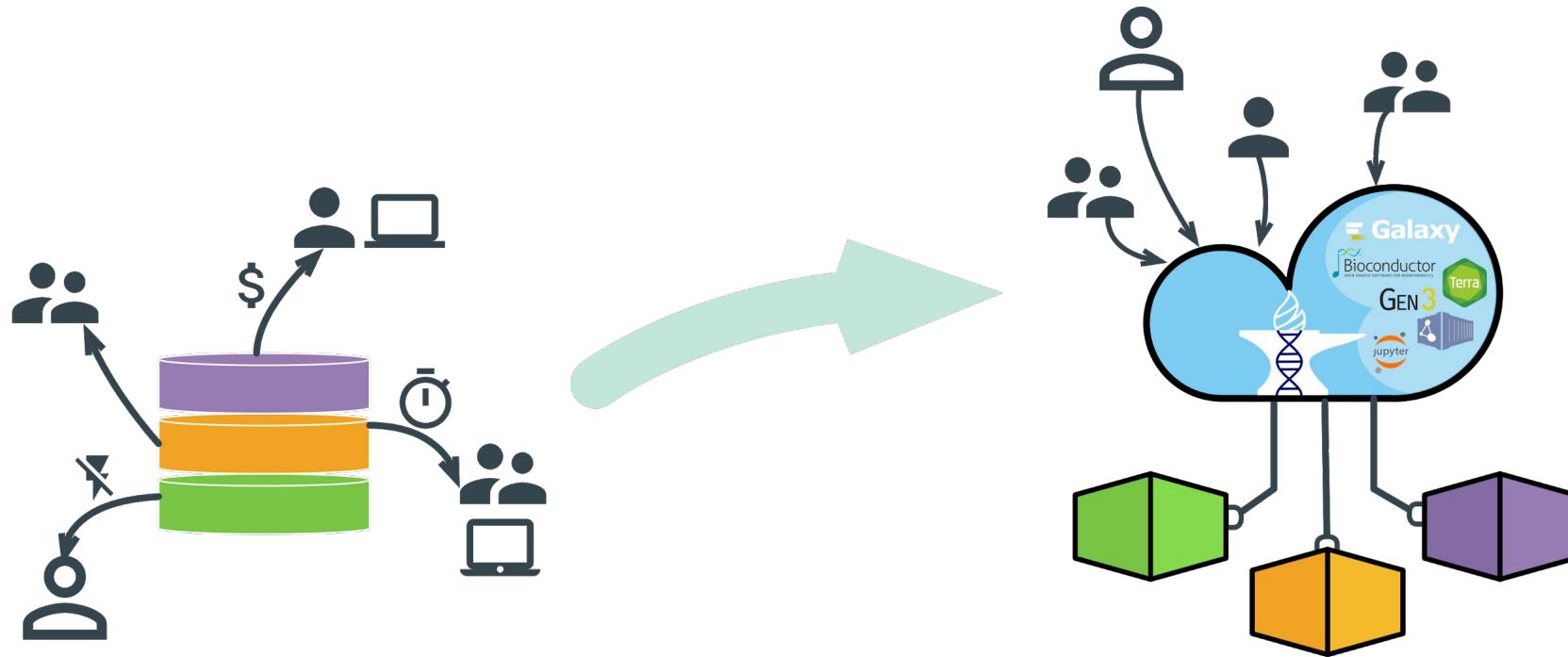
Introduction to the Data Biosphere

Beth Sheets
UCSC Genomics Institute



Problem: data is getting too big to practically download and store

Data Biosphere: Inverting the model of genomic data sharing



Traditional: Bring data to the researcher

- Copying/moving data is costly
- Harder to enforce security
- Redundant infrastructure
- Siloed compute

Goal: Bring researcher to the data

- Reduced redundancy and costs
- Active threat detection and auditing
- Greater accessibility
- Elastic, shared, compute

How should a data biosphere be structured?

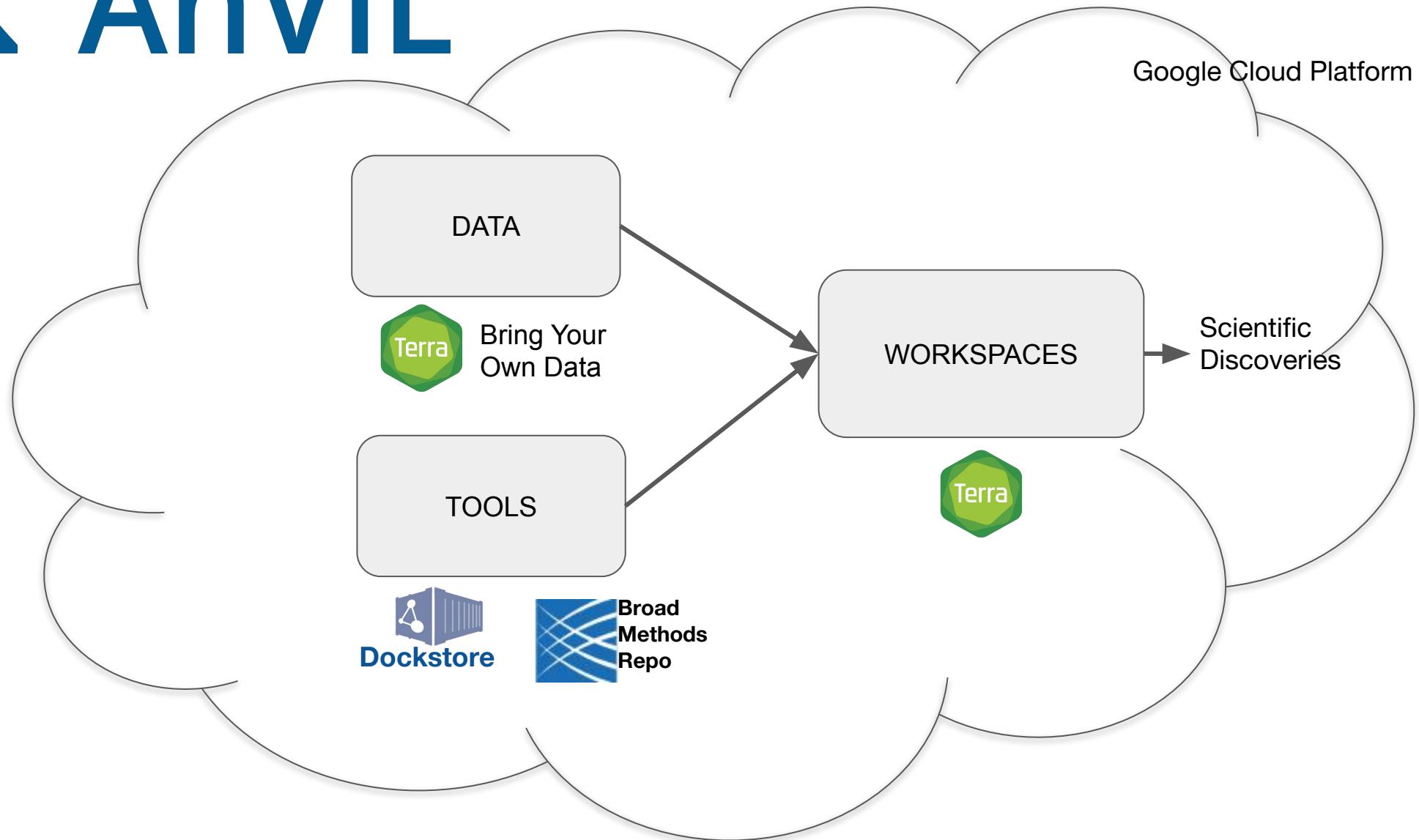
MODULAR	Comprised of functional components with well-specified interface
COMMUNITY FOCUSED	Created by many groups to foster a diversity of ideas
OPEN	Open-source licenses, software, architecture to enable extensibility
STANDARDS BASED	Consistent with standards developed by coalitions such as GA4GH

What is the NHGRI's AnVIL?

Scalable and interoperable computing resource for the genomics scientific community

- **Cloud-based infrastructure**
 - Highly elastic; shared analysis and computing environment
- **Data access and security**
 - Genomic datasets, phenotypes and metadata
 - Large datasets generated by NHGRI programs, as well as other initiatives / agencies
 - dbGaP Authenticated sharing of primary and derived datasets
- **Collaborative computing environment for datasets and analysis workflows**
 - Storage, scalable analytics, data visualization
 - Security, training & outreach, with new models of data access
 - ...for both users with limited computational expertise and sophisticated data scientist users

The screenshot shows the homepage of the AnVIL project at anvilproject.org. The header includes the AnVIL logo, navigation links for About, Data, Tools, Training, News, Events, FAQ, and Contact, and a lock icon indicating secure access. The main title "Migrate Your Genomic Analysis Workflows to the Cloud" is prominently displayed. Below the title, a subtitle reads: "Analyze large, open & controlled-access genomic datasets with familiar tools and reproducible workflows in a secure cloud-based execution environment." A "Terra" button with the text "Launch AnVIL in Terra >" is visible. Five circular icons provide key statistics: 5 CONSORTIA, 96 COHORTS, 72K SUBJECTS, 88K SAMPLES, and 1.1 PB SIZE. At the bottom, a dark blue footer features the "TOOLS" section, the "AnVIL API Library" link, and a command-line interface section with "Learn More >" and a "..." ellipsis.

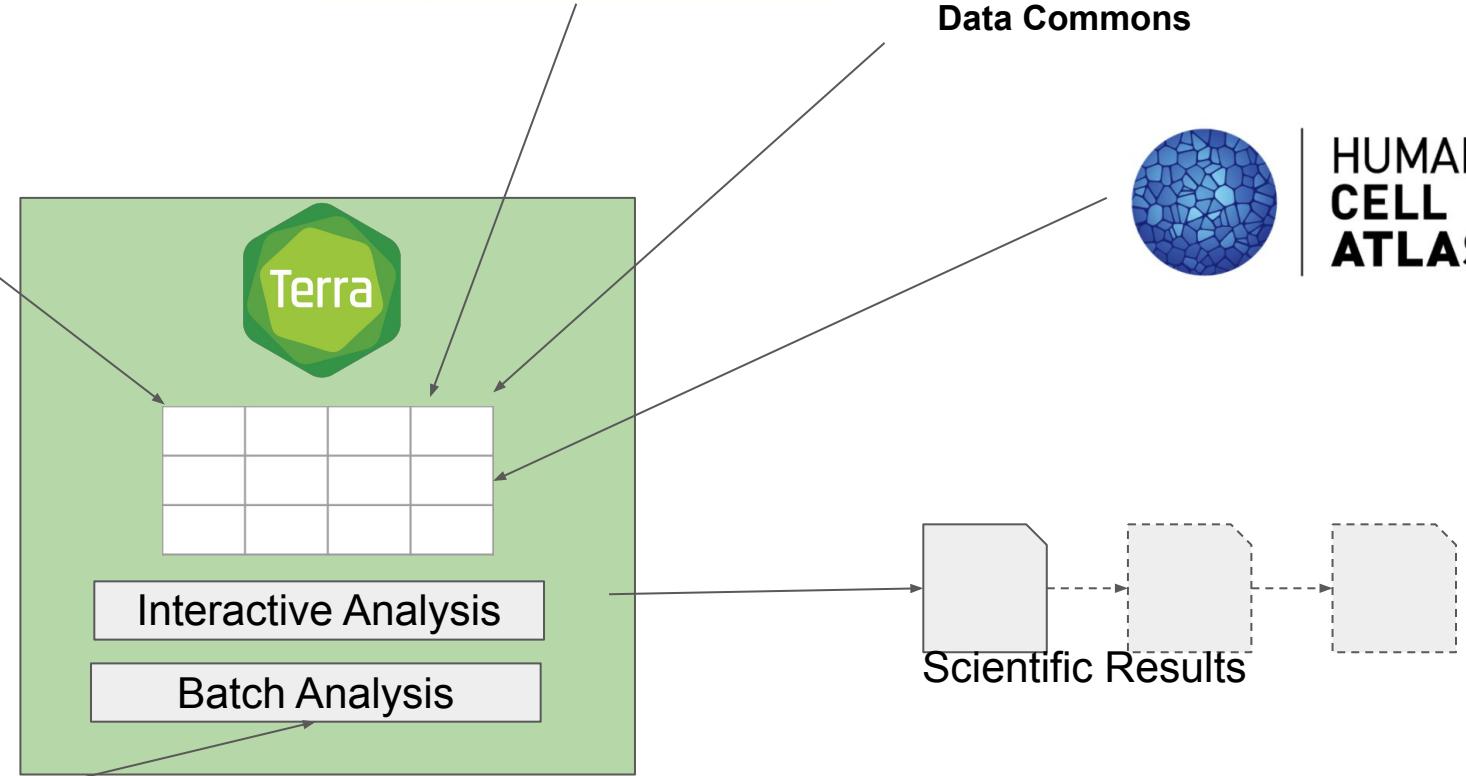




NCI Cancer Research
Data Commons



GEN3
DATA COMMONS



How to start using the AnVIL?

anvilproject.org

The screenshot shows the homepage of anvilproject.org. The header features the AnVIL logo and a navigation bar with links to About, Data, Tools, Training, News, Events, FAQ, and Contact. The main title "Migrate Your Genomic Analysis Workflows to the Cloud" is prominently displayed. Below it is a subtitle: "Analyze large, open & controlled-access genomic datasets with familiar tools and reproducible workflows in a secure cloud-based execution environment." A "Terra" button with the text "Launch AnVIL in Terra >" is shown above five circular statistics: 5 CONSORTIA, 96 COHORTS, 72K SUBJECTS, 88K SAMPLES, and 1.1PB SIZE. At the bottom, there's a "TOOLS" section for the "AnVIL API Library", which includes a command-line interface for interacting with AnVIL data and a "Learn More" link.

Migrate Your Genomic Analysis Workflows to the Cloud

Analyze large, open & controlled-access genomic datasets with familiar tools and reproducible workflows in a secure cloud-based execution environment.

Terra [Launch AnVIL in Terra >](#)

5 CONSORTIA

96 COHORTS

72K SUBJECTS

88K SAMPLES

1.1PB SIZE

TOOLS

AnVIL API Library

Interact with AnVIL data, analysis solutions, and workflows via a command line interface.

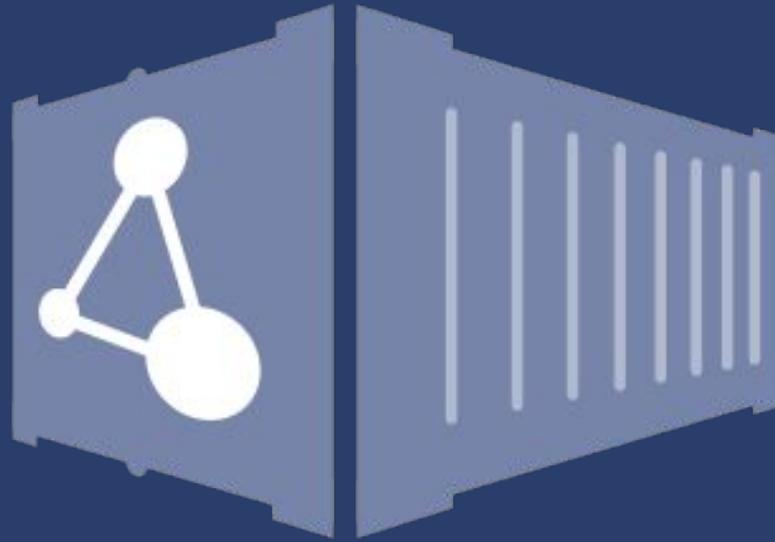
Learn More >>

Workshop Overview:

- Logistics (5 m)
- Overview of the Data Biosphere (10 m)
 - The AnVIL as an example
- **Introduction to Dockstore (20 m)**
 - ‘App store’ for reproducible bioinformatics workflows
- Introduction to Terra (25 m)
 - User friendly workspaces for analysis in the cloud

Break (15 m)

- Integration of Dockstore and Terra: Analyzing COVID-19 viral genomes (60 m)
- Wrap up



Dockstore: An Introduction

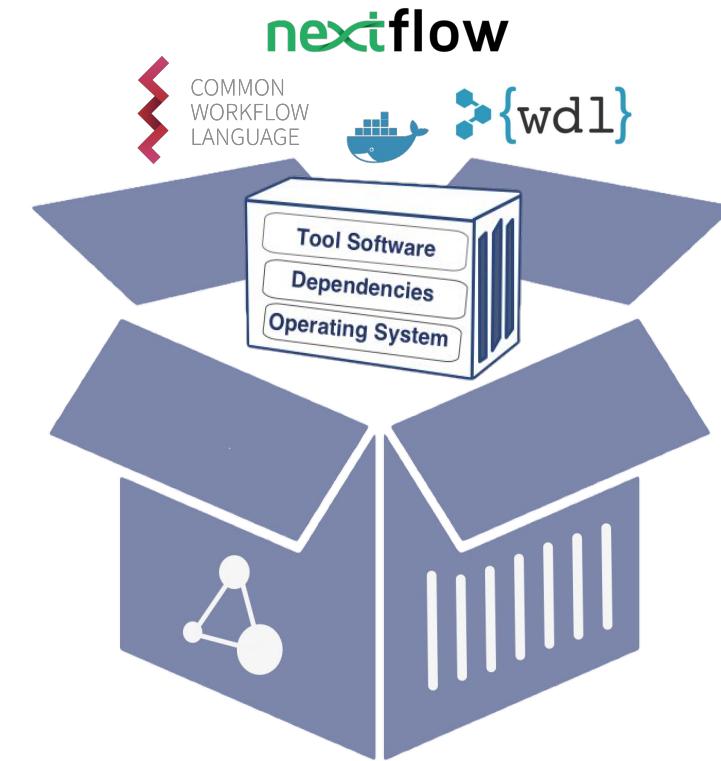
Louise Cabansay

UC Santa Cruz, Genomics Institute
Software Engineer, Dockstore

What is Dockstore?

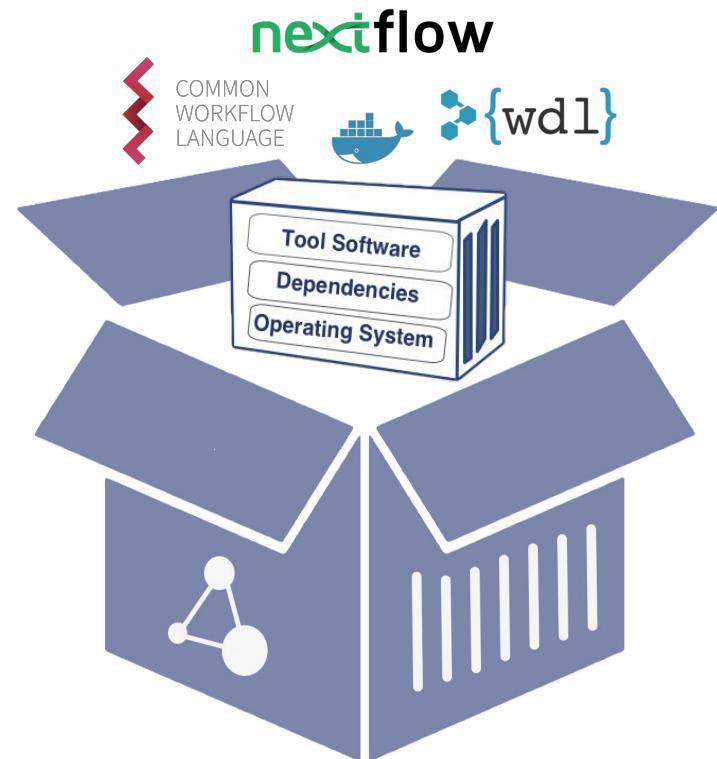
Dockstore is a free and open source platform for sharing scientific tools and workflows. It is a registry of Docker-based resources described using popular workflow languages CWL, WDL, Nextflow and now Galaxy!

- **Portability**
 - Run workflows in any environment that supports Docker
- **Interoperability**
 - Standardize computational analysis through GA4GH APIs
- **Reproducibility**
 - Create, Share, Use
 - Containers + Popular descriptor languages



What is Dockstore? (simple, tl;dr)

- A platform for sharing research software
- An ‘app store’ for bioinformatics



How is Dockstore useful? What problems does it solve?

Simplest Use Case: Search and Findability

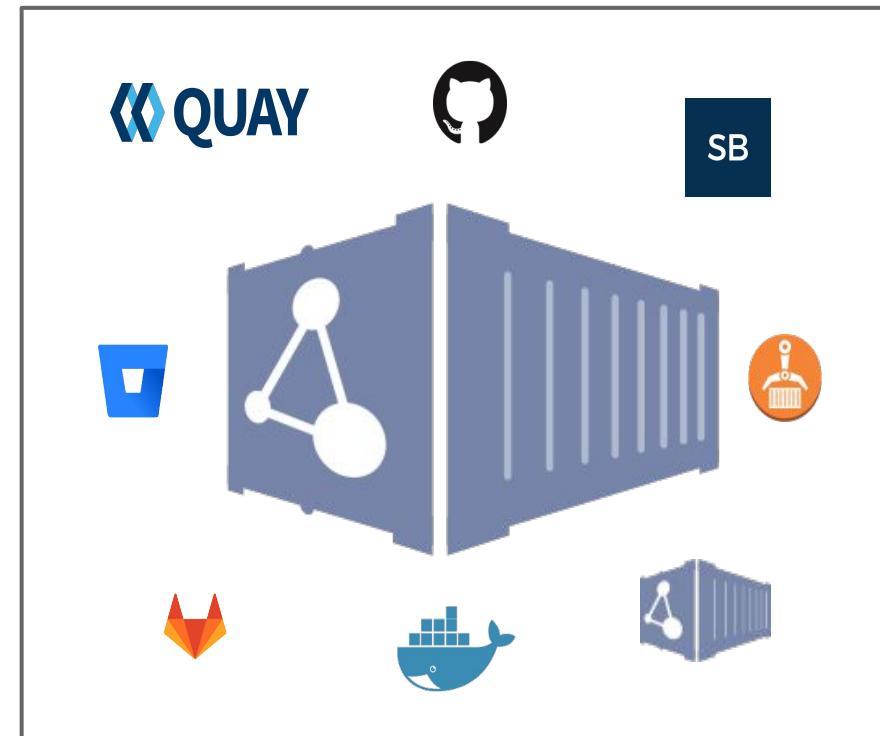
A researcher wants to do some type of genomics work (assembly, alignment, variant calling, GWAS, etc.)

- What software can be used for the job?
- What works with their data?
- Where do they download the software? What version?

Dockstore as a Registry

Searchable and Centralized catalogue of bioinformatics tools and workflows

- Many ways to get tools and workflows into Dockstore!
 - External Hosting: Github, Bitbucket, Gitlab, Quay.io, Docker Hub
 - Direct Hosting (descriptors)



Search Functionality

- Search by name, author, organization, descriptor, etc
- Filter results using facets
 - Descriptor Language
 - Labels

The screenshot shows the 'Advanced Search' interface. At the top, it says 'Find tools and workflows with files that have ...'. Below are four search fields:

- All these words:** e.g. bcbio
- The exact word or phrase:** e.g. cgp
- Any of these words:** e.g. WXS
- None of these words:** e.g. wgs

At the bottom are two buttons: 'Clear All' and 'Advanced Search'.

The screenshot shows the search results page for 'Topmed'. At the top, there are 'Expand All' and 'Collapse All' buttons. The search term 'Topmed' is entered in the search bar. Below the search bar are facets:

- Entry Type:** workflow (7)
- Language:** CWL (4), WDL (3)
- Author:** Walt Shands (2), Ruchi Munshi (1), Seven Bridges (1), Vladimir Obucina (1), Yilin Xu (1)
- Workflow: Source Control:** github.com (7)
- Workflow: Organization:** DataBiosphere (7)

On the right, there is a 'Tag Cloud' section with the text: 'A workflow is a series of tools strung together, with an associated descriptor describing how to run it.' Below the tag cloud is a table of search results:

Name	Author	Format	Project Links	Stars
DataBiosphere/topmed-workflows/UM_variant_caller_wdl	Walt Shands	WDL	GitHub	6★
DataBiosphere/topmed-workflows/UM_aligner_wdl	Walt Shands	WDL	GitHub	2★
DataBiosphere/topmed-workflows/CCDG_aligner_functional	Ruchi Munshi	WDL	GitHub	2★
DataBiosphere/topmed-workflows/UM_aligner_cwl	Seven Bridges	CWL	GitHub	1★
DataBiosphere/topmed-workflows/gatk-vcf-comparator	n/a	CWL	GitHub	
DataBiosphere/topmed-workflows/CCDG_aligner_functional	Yilin Xu	CWL	GitHub	
DataBiosphere/topmed-workflows/UM_variant_caller_cwl	Vladimir Obucina	CWL	GitHub	

Organizations and Collections

- Organizations
 - A place for groups, labs, consortiums, etc to showcase their projects
 - Markdown descriptions
 - Permission-based membership roles
- Collections
 - Group of tools/workflows into highlighted by an organization
 - Markdown descriptions

The screenshot shows the AnVIL platform interface. At the top, there is a dark blue header with a search bar, an 'Organizations' button, a 'Docs' button, and a 'Login' button. Below the header, the word 'Organizations' is displayed. The main content area features the AnVIL logo and the text 'Analysis, Visualization and Informatics Lab-space'. It states that the project is funded by the NIH to create a managed platform for genomics researchers. Below this, there is a link to the website (<https://anvilproject.org/>) and a star rating section. The interface includes tabs for 'Collections', 'Members', and 'Events'. Under the 'Collections' tab, two items are listed: 'GATK4' (Variant caller) and 'Single Cell Pipelines' (Collection of Single Cell processing pipelines). To the right of these collections, there is a detailed sidebar with sections for 'About the Project', 'Project Aims', 'Operate services', and 'Organize and host key NHGRI datasets'. The sidebar also includes a '0' rating indicator.

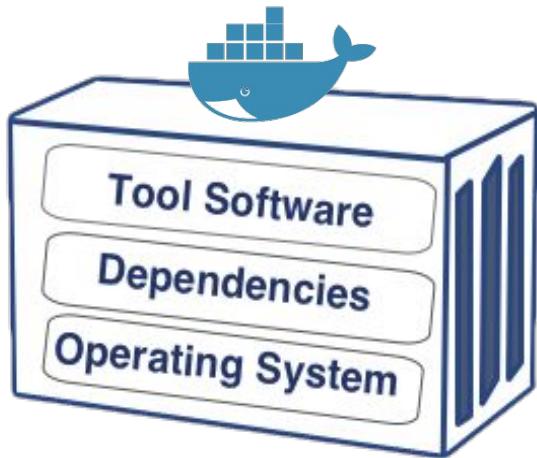
Common Use Case: Simplifying Installation Process

A researcher already knows what software they want to use, but...

- Installation problems
 - Software was built on a different system
 - compiler or build issues
 - Install documentation is incomplete or out of date
- Dependency problems
 - Software requires a different version than on your environment (ex. Java, Python)
 - Multiple tools have conflicting version requirements of a dependency

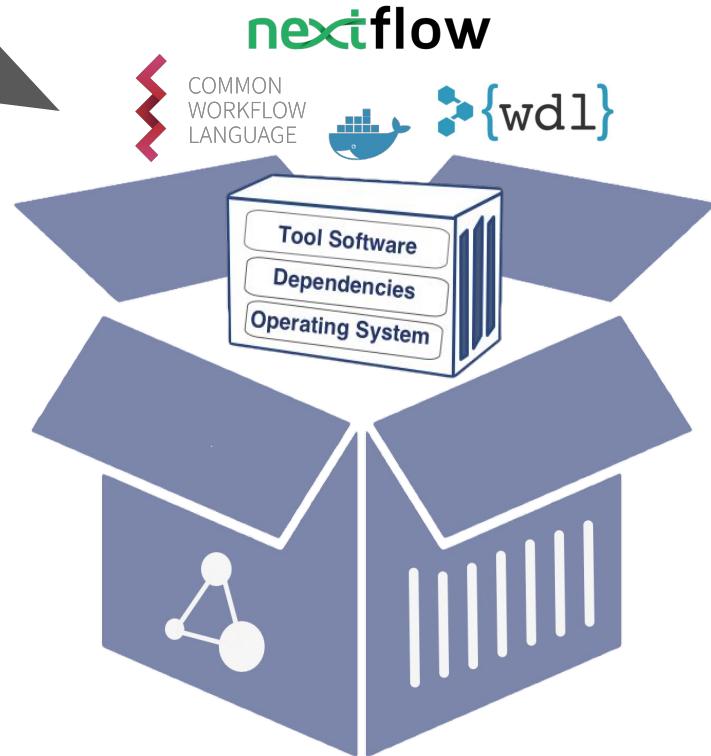
How It Works: Packaging Content

Container



+

Descriptor



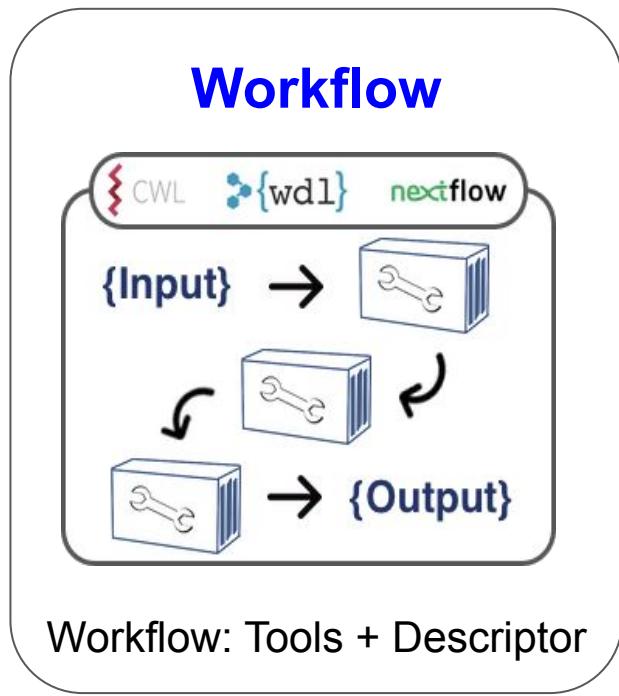
Software is “packaged” using containerization technology and described using descriptor language

- Analysis can be moved from cloud-to-cloud, VM-to-VM, server-to-server and yet be guaranteed to run on anything that supports Docker

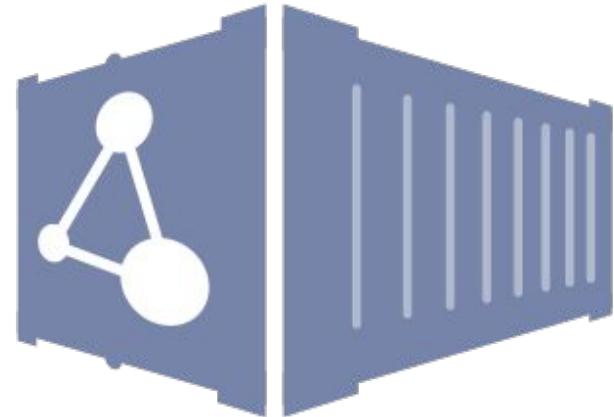
Content: tool vs workflow



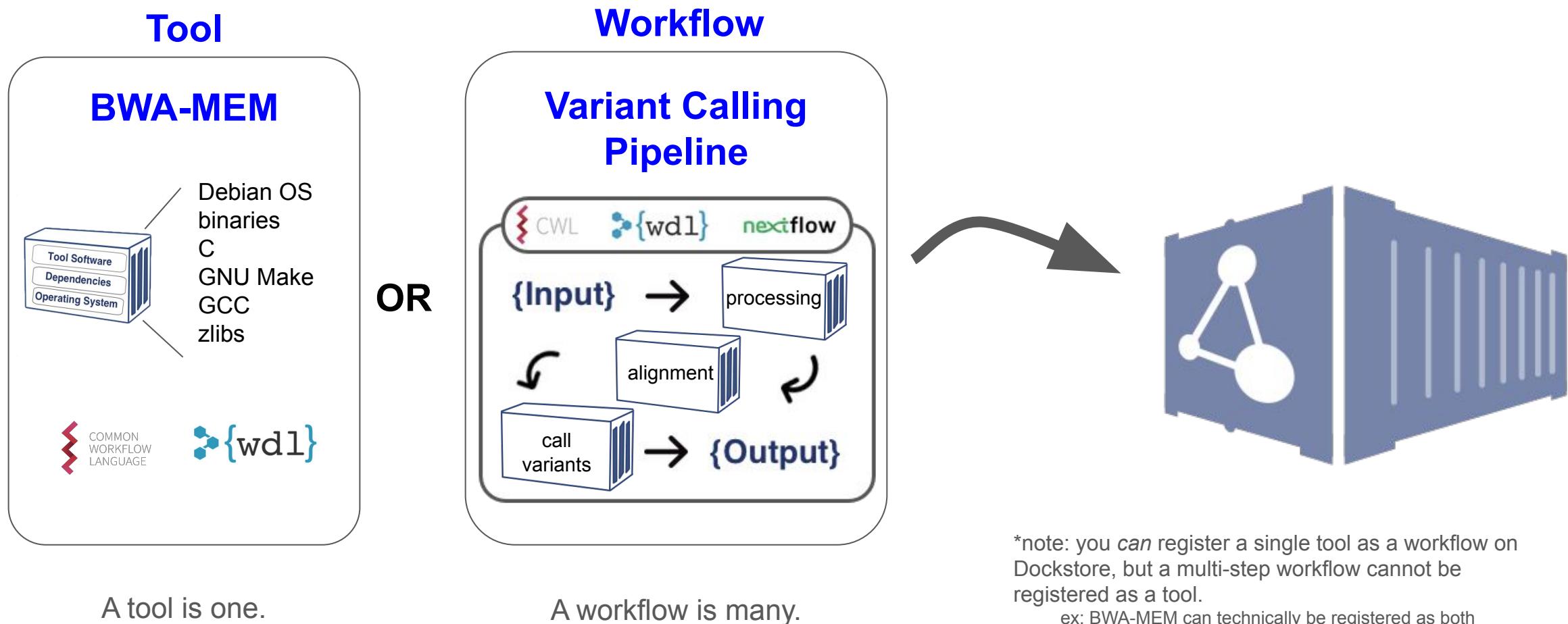
A tool uses a single container and performs a single action that is outlined by a descriptor.



A workflow can use multiple containers and executes multiple actions or steps, still outlined by a descriptor



Content: tool vs workflow (example)



Containers, Descriptors, and Parameter Files

Container:

Packaged up code with all of its dependencies. This allows for portable software that runs quickly and reliably from one computing environment to another.



Containers, Descriptors, and Parameter Files

Container:

Packaged up code with all of its dependencies. This allows for portable software that runs quickly and reliably from one computing environment to another.



Descriptor:

A workflow language used to *describe* how to run your pipeline.

- Which containers
- What steps and when
- Define parameters
 - I/O data
 - compute requirements
- Metadata



Containers, Descriptors, and Parameter Files

Container:

Packaged up code with all of its dependencies. This allows for portable software that runs quickly and reliably from one computing environment to another.



Descriptor:

A workflow language used to *describe* how to run your pipeline.

- Which containers
- What steps and when
- Define parameters
 - I/O data
 - compute requirements
- Metadata



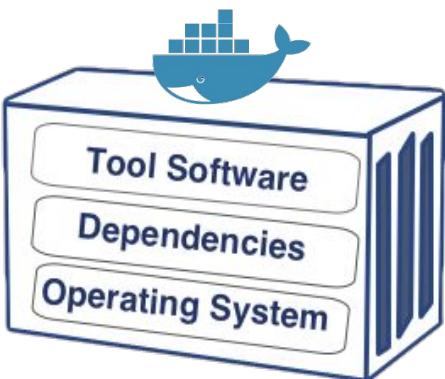
Parameter File (wdl, cwl):

- Specifies the actual input/output files (local, ftp, http, or cloud)
- Set compute resources
- JSON, YAML

```
{  
    "bam_input": {  
        "class": "File",  
        "format": "http://edamontology.org/format_  
2572",  
        "path": "/tmp/NA12878.chrom20.ILLUMINA.bw  
a.CEU.low_coverage.20121211.bam"  
    },  
    "bamstats_report": {  
        "class": "File",  
        "path": "/tmp/bamstats_report.zip"  
    }  
}
```

Example: BWA

- Images of containers are built from Dockerfiles
- Dockerfiles describe the packaged up environment:
 - operating system or base image to build upon
 - Metadata (ex: authorship)
 - dependencies needed for the software
 - the actual analysis software



```
Dockerfile  x  
1 #####  
2 # Dockerfile to build a sample container for bwa  
3 #####  
4  
5 # Start with a base image  
6 FROM ubuntu:18.04  
7  
8 # Add file author/maintainer and contact info (optional)  
9 MAINTAINER Foo Bar <foobar@institute.edu>  
10  
11 #set user you want to install packages as  
12 USER root  
13  
14 #update package manager & install dependencies (if any)  
15 RUN apt update -y  
16  
17 # install analysis software from package manager  
18 RUN apt -y install bwa  
19  
20
```

Example: Metrics.wdl

A simple example workflow that generates some statistics about an alignment.

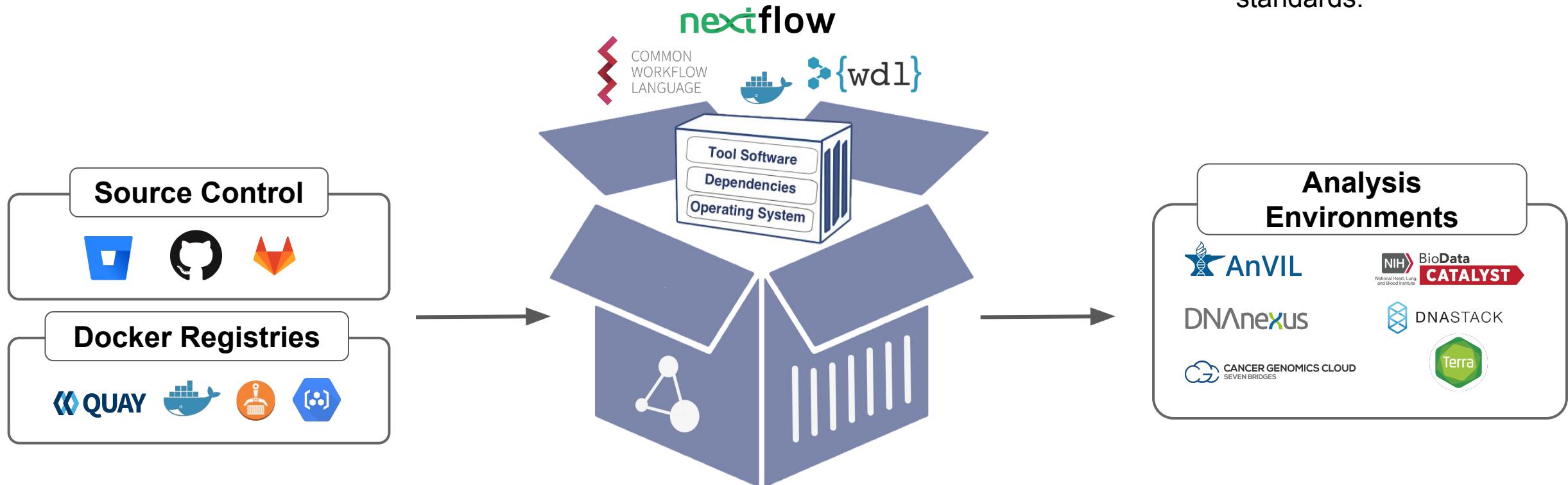
- Specify WDL version
- Define workflow, call, and task(s)
- Define parameters
 - Input and output
 - Parameterized command(s)
 - Runtime environment
 - Containers
 - Compute resource
- Metadata
 - Authorship, contact information, etc

```
metrics.wdl  x
1  version 1.0
2
3  workflow metrics {
4      call flagstat
5      output { File align_metrics = Flagstat.metrics }
6  }
7
8  task flagstat {
9      input {
10         File input_bam
11     }
12
13     String stats = basename(input_bam) + ".metrics"
14     command {
15         samtools flagstat ${input_bam} > ${stats}
16     }
17     output { File metrics = "${stats}" }
18
19     # set some parameterized runtime parameters
20     runtime {
21         docker: "quay.io/ldcabansay/samtools:latest"
22     }
23 }
24 meta {
25     author: "Foo Bar"
26     email: "foobar@institute.edu"
27 }
```

Dockstore Ecosystem



This powerful interoperability is possible through GA4GH API standards.



Store your descriptors and containers and descriptors on your preferred sites

Register these as tools and workflows on Dockstore, allowing for a centralized bioinformatics catalog of resources

Dockstore's launch-with feature enables users to export tools and workflows to a variety of cloud compute platforms

Launching Analysis

The screenshot shows the Galaxy Workflows interface. On the left, a workflow descriptor file named `vg_map_call_sv.wdl` is displayed, containing configuration for structural variant calling using VG. The code includes details like sample names, input files (FASTQ, CRAM), reference genome, and various parameters for the analysis. On the right, a "Launch with" interface is shown, listing several partner platforms: DNAstack, DNAexus, Terra, CGC, AnVIL, and NHLBI BioData Catalyst. Below these are sections for "Recent Versions" (svpack Jan 28, 2020) and "Source Repositories" (GitHub). A "Collections" section contains a link to "Structural Variant Calling using Graph Genomes". At the bottom, there are social sharing icons for Facebook, Twitter, LinkedIn, and Reddit.

Partner Platforms

- DNAstack
- DNAexus
- Terra
- CGC : Cancer Genomics Cloud (Seven Bridges)
- AnVIL
- BioData Catalyst

Structural Variant Calling using Graph Genomes

Contributed by: Jean Monlong and Charles Markello
(VG Team, UC Santa Cruz, Genomics Institute)

Launching Analysis - Example



A screenshot of the Terra Import Workflow interface. At the top, it says "POWERED BY Terra BETA IMPORT WORKFLOW". Below that, there's a section titled "Importing from Dockstore" with the URL "github.com/klarman-cell-observatory/cumulus/Cumulus v.0.15.0". A warning message states: "Please note: Dockstore cannot guarantee that the WDL and Docker image referenced by this Workflow will not change. We advise you to review the WDL before future runs." To the right, there are input fields for "Workflow Name" (set to "Cumulus") and "Destination Workspace" (set to "Dockstore-Webinar"). A large "IMPORT" button is at the bottom of the right panel. The background features a pattern of overlapping hexagons.

Cumulus: Cloud-based single-cell and single-nucleus genomics analysis workflows

AnVIL Organization, Cumulus Collection: <https://dockstore.org/organizations/anvil/collections/Cumulus>

Contributed by: Bo Li & Yiming Yang (Cumulus Team, Broad Institute)

Earlier Question:
What problems does this solve?

Collaborations, Scale, and Reproducibility

Ex: A team of collaborators from X different institutions are working together on a study that analyzes the genomes of thousands of individuals.

- Each institution uses a different compute environment for analysis
 - workstations, local machines, small servers, HPCs, or various cloud platforms
- Dockstore's way of packaging content makes sure that pipelines work across all the compute environments used by the collaborators.
- Algorithm portability let's you move and scale analysis in the cloud as the datasets grow. This enables research to go beyond compute resources that are traditionally available locally.
- Organizations, Collections, and Digital Object Identifiers (DOIs) allow others to find, share, and precisely reproduce the research done in the resulting publications from these collaborators.

Documentation and Tutorials

- Example Topics:
 - Creating Tools and Workflows
 - Launching Tools and Workflows
 - Creating snapshots and Digital Object Identifiers (DOIs)
 - Creating Organizations
 - Writing checker workflows
 - And many more!

<https://docs.dockstore.org/>

Developer Tutorial
Go through the process of creating a tool and registering it on Dockstore.

End User Tutorials
Learn how to use Dockstore from the perspective of a user who runs tools and workflows.

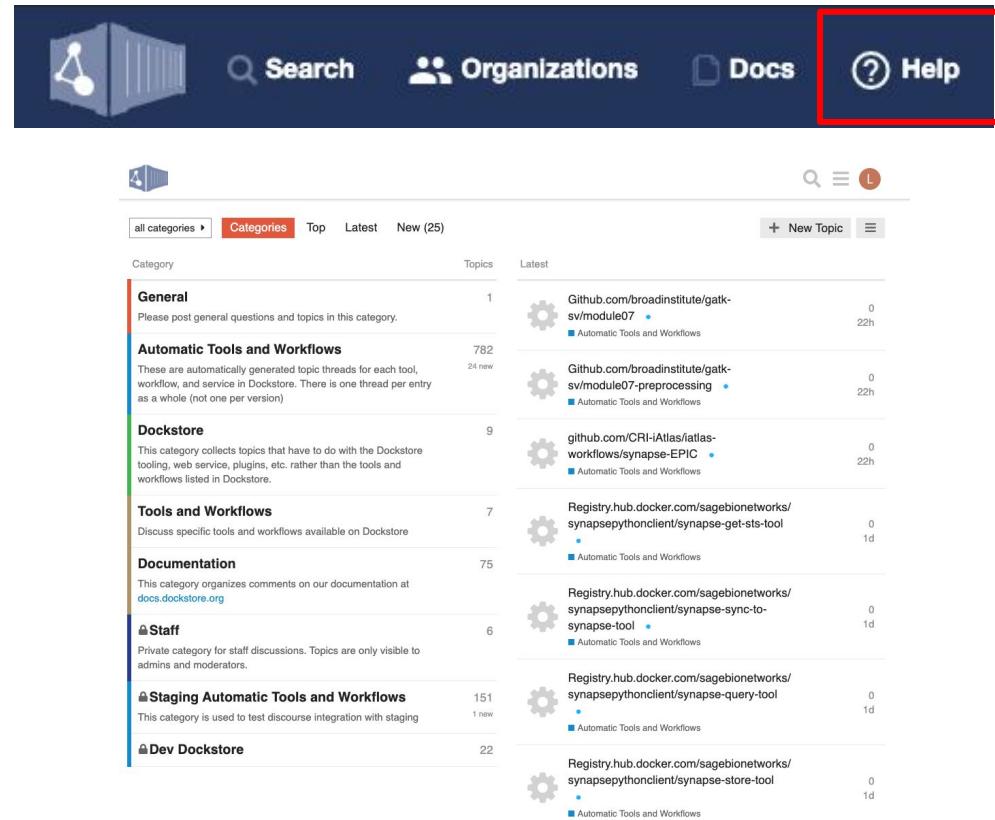
Advanced Tutorials
A collection of articles and tutorials regarding advanced Dockstore topics

The screenshot shows the Dockstore documentation website. On the left, there's a sidebar with navigation links for 'ABOUT', 'GETTING STARTED GUIDE', 'LAUNCH', and 'ADVANCED DEVELOPER TOPICS'. The main content area has a heading 'Organizations and Collections' with a sub-section 'Organizations'. It explains that organizations are landing pages for collaborations and provides a link to the 'organizations' page. Below this is a 'Create Organization Request' form. The form fields include 'Name *' (set to 'OICR'), 'Display Name *' (set to 'Ontario Institute for Cancer Research'), 'Topic *' (set to 'OICR is a collaborative, not-for-profit research institute accelerating the development of n...'), 'Organization website' (set to 'https://oicr.on.ca/'), 'Link to organization website' (empty), 'Location' (set to 'Toronto, ON'), 'The location of the organization' (empty), 'Contact Email Address' (empty), 'Once approved, this email address will be publicly visible', 'Image URL' (set to 'https://oicr.on.ca/wp-content/uploads/2017/01/OICR_Logo.png'), and 'Link to your organization's logo. Link must end in jpg, jpeg, png, or gif' (empty). At the bottom right of the form are 'Cancel' and 'Create Organization Request' buttons.

Getting Help on Dockstore

User forum at <https://discuss.dockstore.org/>

- Topics embedded with each tool, workflow, and documentation page.
- Talk about bioinformatics, workflows, and get help on development



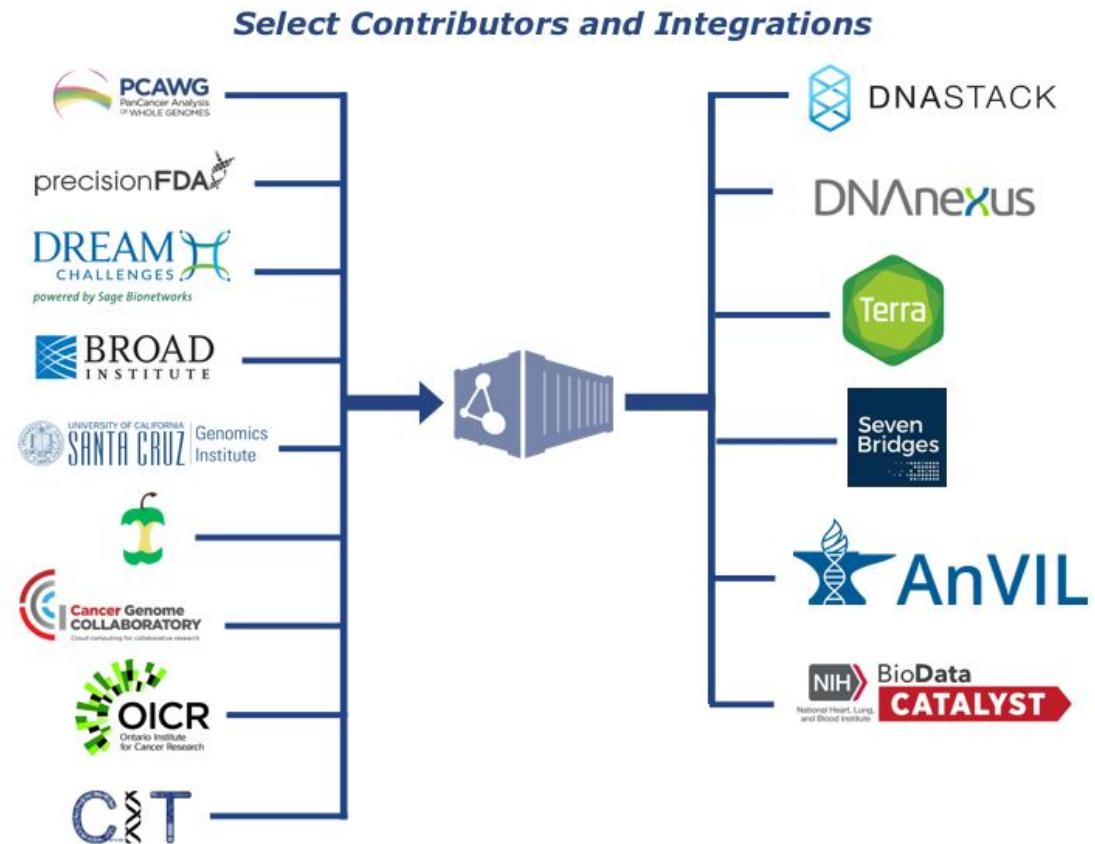
The screenshot shows the Dockstore user forum homepage. At the top, there is a dark blue header with the Dockstore logo, a search bar, an 'Organizations' link, a 'Docs' link, and a 'Help' link which is highlighted with a red box. Below the header, there is a navigation bar with links for 'all categories', 'Categories', 'Top', 'Latest', and 'New (25)'. The main content area displays a list of categories and their corresponding topics. The categories include:

- General**: Please post general questions and topics in this category. 1 topic.
- Automatic Tools and Workflows**: These are automatically generated topic threads for each tool, workflow, and service in Dockstore. There is one thread per entry as a whole (not one per version). 782 topics.
- Dockstore**: This category collects topics that have to do with the Dockstore tooling, web service, plugins, etc. rather than the tools and workflows listed in Dockstore. 9 topics.
- Tools and Workflows**: Discuss specific tools and workflows available on Dockstore. 7 topics.
- Documentation**: This category organizes comments on our documentation at docs.dockstore.org. 75 topics.
- Staff**: Private category for staff discussions. Topics are only visible to admins and moderators. 6 topics.
- Staging Automatic Tools and Workflows**: This category is used to test discourse integration with staging. 151 topics.
- Dev Dockstore**: 22 topics.

Each topic entry includes a gear icon, the topic title, the number of replies, and the last update time.

Dockstore Community

Dockstore is thankful to its many contributors, users, and partners. This community has pulled together a library of over **700** tools and workflows. In the diagram to the right we've highlighted a few select contributors to give a sense of what has been occurring in this space.



The Dockstore Team



UNIVERSITY OF CALIFORNIA
SANTA CRUZ Genomics
Institute

Lincoln Stein

Denis Yuen

Andrew Duncan

Gary Luu

Gregory Hogue

Benedict Paten

Elnaz Sarbar

Charles Overbeck

Walt Shands

David Steinberg

Nneka Olunwa

Beth Sheets

Louise Cabansay

Natalie Perez

Melaina Legaspi

Charles Reid

Emily Soth

Andy Chen

Acknowledgements



This work was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-168).



Funded by:



Workshop Overview:

- Logistics (5 m)
- Overview of the Data Biosphere (10 m)
 - The AnVIL as an example
- Introduction to Dockstore (20 m)
 - ‘App store’ for reproducible bioinformatics workflows
- **Introduction to Terra (25 m)**
 - User friendly workspaces for analysis in the cloud

Break (15 m)

- Integration of Dockstore and Terra: Analyzing COVID-19 viral genomes (60 m)
- Wrap up



A Platform for Biomedical Researchers to Access Data, Run Analysis Tools, and Collaborate

Data Sciences Platform, Broad Institute of MIT and Harvard

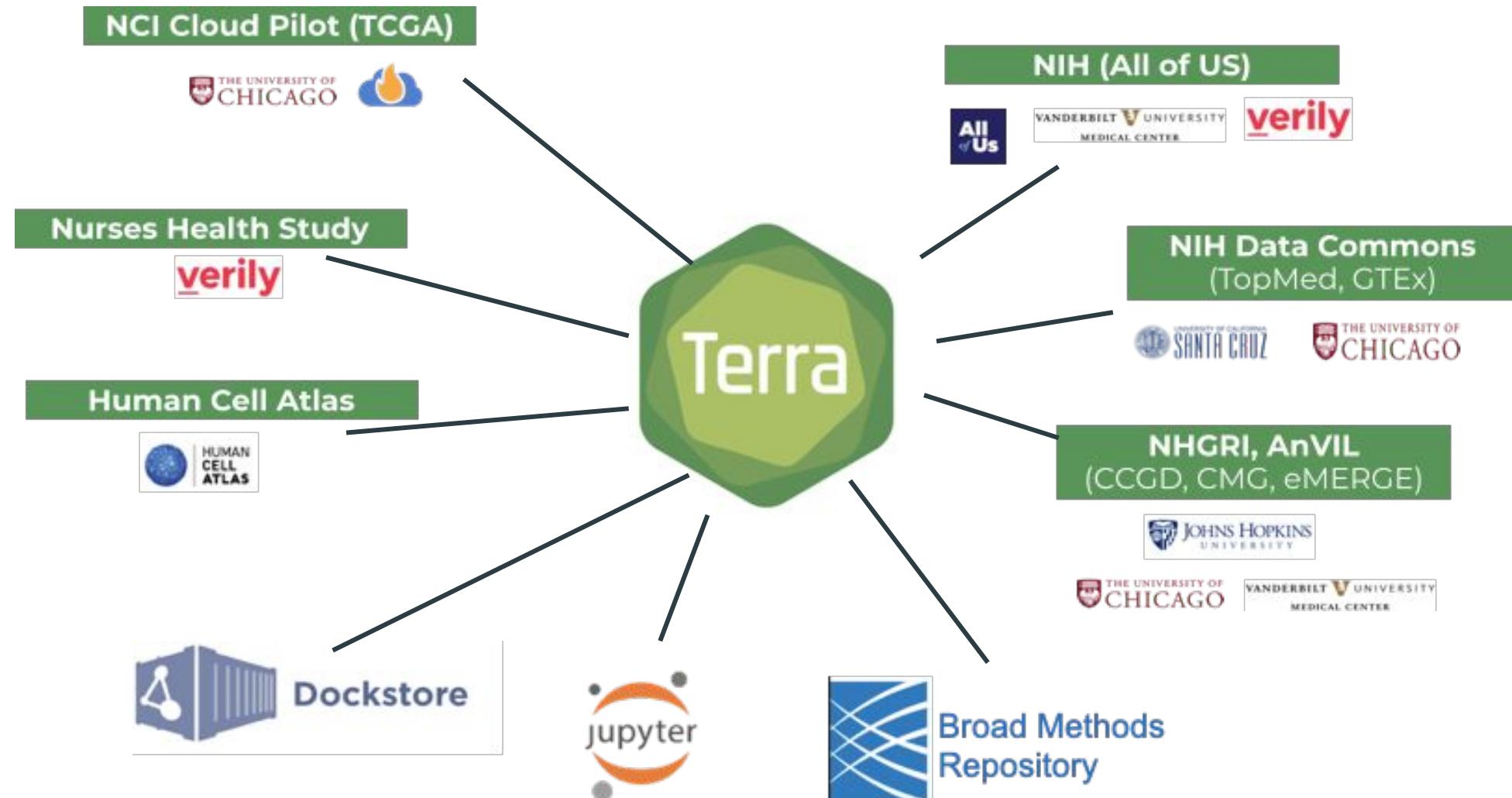


Learning Objectives

- Understand the components of Terra for research
 - Data
 - Bulk Analysis
 - Interactive analysis
- Understand the use of containers for portability and reproducibility
- Understanding Terra's security compliance
- Understanding Costs associated with Terra



A cloud-based analysis platform to access and analyze data





Terra: An end-to-end, cloud-native platform



<https://terra.bio>



The Workspace - The fundamental unit in Terra

The screenshot shows the Terra Workspaces interface. At the top, there's a navigation bar with the Terra logo, a 'WORKSPACES' section labeled 'BETA', and links for 'Workspaces > help-gatk/Reproducibility_Case_Study_Tetralogy_of_Fallot'. The main menu includes 'DASHBOARD' (selected), 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. On the right, there are icons for 'Notebook Runtime STOPPED (\$0.03 hr)' and a gear icon.

ABOUT THE WORKSPACE

Reproducing the paper: Variant analysis of Tetralogy of Fallot

Overview

This workspace reproduces the work described by Matthieu Miossec and collaborators in the bioRxiv preprint "Deleterious genetic variants in NOTCH1 are a major contributor to the incidence of non-syndromic Tetralogy of Fallot (ToF)." The original ToF study is a classic example of a study to understand the genetics that underlie a particular phenotype. The workspace reproduces all steps in the study as closely as possible, from processing the raw data (BAM) files, to calling variants, to the clustering analysis that led to the final result. The workspace serves as a template of best practices for making your own work easily reproducible with a detailed explanation of how we reproduced the ToF study using a cloud-based analysis platform. Sample data and notebooks allow users to reproduce the process themselves.

Summary of original ToF study

By analysing high-throughput exome sequence data from 867 cases and 1252 controls, the authors identified 49 deleterious variants within the NOTCH1 gene that appeared to correspond with the ToF congenital heart disease. Others had previously identified NOTCH1 variants in families with congenital heart defects, including ToF. However, the work by Miossec et al. is the first to scale variant analysis of ToF to a cohort of nearly a thousand case samples and show that NOTCH1 is a significant contributor to ToF risk.

WORKSPACE INFORMATION

CREATION DATE 3/15/2019	LAST UPDATED 3/21/2019
SUBMISSIONS 0	ACCESS LEVEL Proj. Owner
EST. \$/MONTH \$0.00	

OWNERS

bshifaw@broadinstitute.org
jhajian@broadinstitute.org
schalova@broadinstitute.org

TAGS

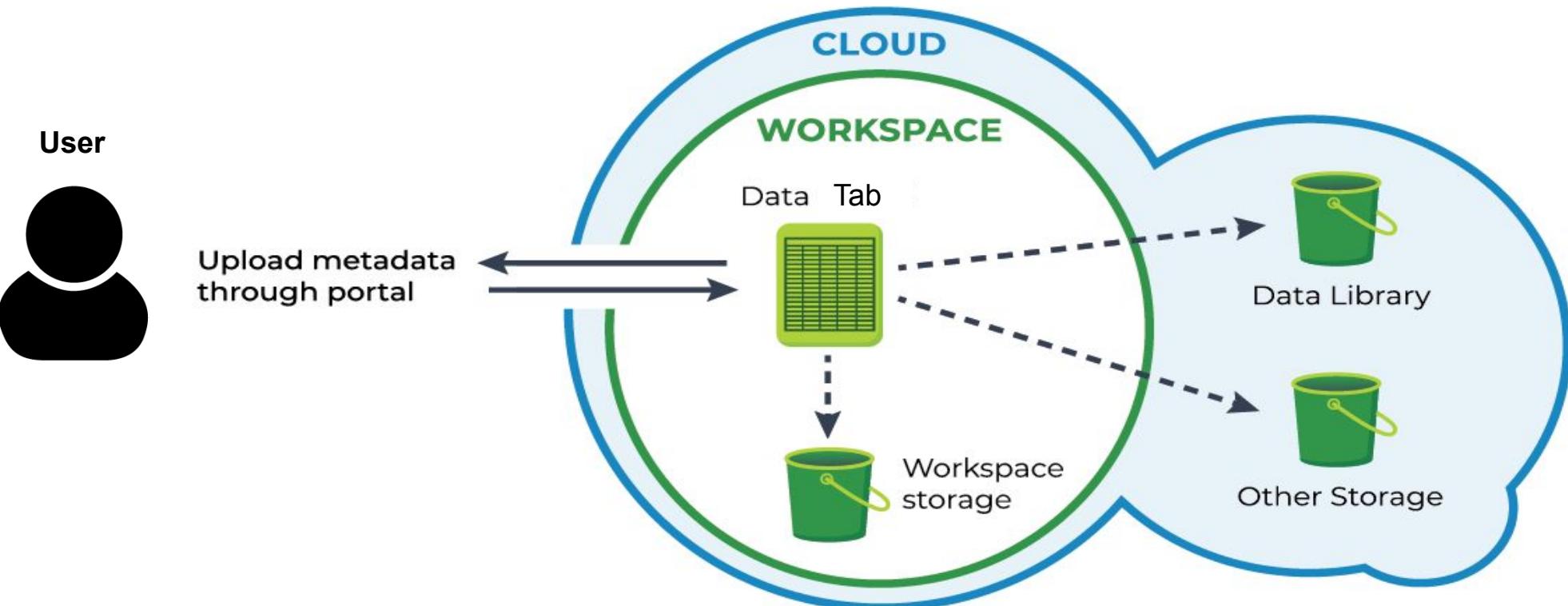
Add a tag

ASHG18 exome germline
reproduction workshop

Google Bucket
fc-8ccba5f8-0e73-424e-8c8f-1b3315...
[Open in browser](#)



Organize data and tools in the Terra workspace





Find and access data in the Terra Data Library

TERRA BETA **LIBRARY**

[DATASETS](#) [SHOWCASE & TUTORIALS](#) [CODE & WORKFLOWS](#)

1000 Genomes High Coverage
presented by NHGRI AnVIL

1000 Genomes project phase 3 samples sequenced to 30x coverage. This dataset is delivered as a workspace. You may clone this workspace to run analyses or copy specific samples to a workspace of your choice.

Participants: 2,504

[BROWSE DATA](#)

1000 Genomes Low Coverage

The 1000 Genomes Project ran between 2008 and 2015, creating the largest public catalogue of human variation and genotype data. The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the populations studied.

Participants: 3,500

[BROWSE DATA](#)

AMP Parkinson's Disease

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... [READ MORE](#)

Participants: > 4,700

[BROWSE DATA](#)

Project Baseline by verily

Baseline Health Study 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

[BROWSE DATA](#)

CCDG presented by NHGRI AnVIL

The Centers for Common Disease Genomics (CCDG) are a collaborative large-scale genome sequencing effort to comprehensively identify rare risk and protective variants contributing to multiple common disease phenotypes.

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.

ENCODE Project

The Encyclopedia Of DNA Elements (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.

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.



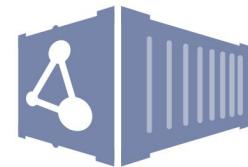
Two different modes of computation

- ▶ Bulk analysis workflows
 - Some jobs take a while, so launch them and come back later! (WDL and Cromwell)
- ▶ Interactive analysis
 - ▶ Run R or Python via Jupyter Notebooks, or run shell scripts on the terminal
 - ▶ Interrogate data in real time
 - ▶ Perform GWAS with Hail



Accessible, findable bulk analysis tools

The screenshot shows the Terra Workspaces interface. At the top, there's a navigation bar with icons for Home, Help, and Logout, followed by the text "WORKSPACES". Below the navigation bar, the path "Workspaces > help-gatk/Exome-Analysis-Pipeline > Workflows (read only)" is displayed. A horizontal menu bar below the path includes "DASHBOARD", "DATA", "NOTEBOOKS", "WORKFLOWS" (which is highlighted in green), and "JOB HISTORY". The main content area is titled "WORKFLOWS". On the left, there's a button "Find a Workflow" with a plus sign icon. To its right, a workflow card for "ExomeGermlineSingleSample" is shown, indicating it was created by "V. dev" and sourced from "dockstore". There's also a small three-dot menu icon next to the card.



Dockstore



**Broad
Methods
Repository**

- ▶ Align & QC sequence data per sample
- ▶ Call short variants per sample
- ▶ Joint-call across population
- ▶ Filter & QC variants



Bulk analysis: GATK4 workflows preloaded in showcase workspaces

Screenshot of the Terra platform interface showing preloaded GATK4 workflows in showcase workspaces.

The interface includes a navigation bar with sections: DATASETS, SHOWCASE & TUTORIALS (selected), and CODE & WORKFLOWS.

The main content area is divided into three columns:

- New and interesting:**
 - COVID-19_Broad_Viral_NGS**: Massachusetts has been severely impacted by the COVID-19 pandemic, with 101,163 cases and 7,085 deaths as of June 2, 2020. Seventy percent of the state's 6.9 M population lives in the city.
 - COVID-19**: This workspace contains COVID-19 genomic data and workflows that will enable you to perform viral genomic analysis. This workspace will be routinely updated with new, additional data as it
 - COVID-19_cross_tissue_analysis**: The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, underscores the urgent need to identify molecular mechanisms that mediate viral entry, propagation, and tissue
 - Genomics-in-the-Cloud-v1**: Companion workspace for Genomics in the Cloud, an O'Reilly animal book by Geraldine A. Van der Auwera and Brian D. O'Connor.
- Featured workspaces:**
 - Introduction-to-TCGA-Dataset**: # Workspace Overview. Practice accessing and analysing controlled-access TCGA data with example analysis Tools. Data processing.
 - DNA-methylation-preprocessing**: Suite of tools to conduct methylation data analysis. Methods from this workspace can be used for alignment.
 - Bioconductor**: Explore common Bioconductor packages that can be used to perform bulk RNA differential expression analyses or manipulate single-cell RNA-seq data.
 - Waddington-OT**: This tutorial provides a practical, hands-on introduction to inferring developmental trajectories with [Waddington-OT] (<https://broadinstitute.github.io/wot/>).
- GATK4 example workspaces:**
 - Germline-CNVs-GATK4**: ### GATK Best Practices for Germline Copy Number Variation. An analysis to detect germline copy number variants in exome sequence.
 - Variant-Functional-Annotation-With-Funcotator**: ### GATK Best Practices for Funcotator **Funcotator** (FUNCTIONal annOTATOR) analyzes variants for their function and writes the analysis to a specified output file.
 - Variant_Calling_Spark_Multicore**: ### GATK Best Practices for Variant Calling with Spark on a Multicore Machine. This workspace highlights a pipeline for
 - Exome-Analysis-Pipeline**: ### GATK Best Practices for Germline SNPs and Indels as used at the Broad Institute. A fully reproducible example workflow



Interactive analysis with Jupyter Notebooks

Screenshot of a Jupyter Notebook interface running on Terra Workspaces.

Header: Workspaces > fc-product-demo/2019_ASHG_Reproducible_GWAS_AH_01_13_2020 > notebooks > GWAS_initial_analysis_completed.ipynb
Notebook Runtime: RUNNING (\$0.39 hr)

Toolbar: File, Edit, View, Insert, Cell, Kernel, Widgets, Help
Not Trusted, Playground Mode (Edits not saved), PySpark 3

Contents:

- 1 Introduction
- 2 Set up your notebook
 - 2.1 Set runtime values
 - 2.2 Check kernel type
 - 2.3 Enable useful notebook extensions
 - 2.4 Load Python packages
- 3 Load phenotypes
 - 3.1 Set environment variables
 - 3.2 Load phenotype data
- 4 Examine phenotype data
 - 4.1 Generating distribution plots
 - 4.1.1 Exercise: Univariate distributions
 - 4.1.2 Exercise: Bivariate distributions
 - 4.1.3 Exercise: Boxplots
- 5 Working with genotype data using Hail
 - 5.1 Query workspace storage for VCF files
 - 5.2 Import packages and start a Hail session
 - 5.3 Load VCF data and perform variant QC
 - 5.3.1 Exercise: Load 1000 Genomes data
 - 5.3.2 Exercise: View matrix table structure
 - 5.3.3 Exercise: Merge phenotype and VCF data
 - 5.3.4 Exercise: Generate variant level summary statistics
- 6 Understanding population structure within our sample
 - 6.1 Variant filtering
 - 6.1.1 Exercise: Filter variants
 - 6.2 LD-pruning
 - 6.2.1 Exercise: Run LD pruning
 - 6.2.2 Exercise: Filter the matrix table
 - 6.3 Principal Component Analysis
 - 6.3.1 Exercise: Run PCA
 - 6.3.2 Exercise: Add PCA values to matrix table
 - 6.3.3 Exercise: Visualize samples in PCA space
- 7 Save sample metadata and update the data table
 - 7.1 Convert the phenotype data to the correct format
 - 7.1.1 Extract sample metadata
 - 7.1.2 Convert data format
 - 7.1.3 Flatten the PC array
 - 7.1.4 Convert the phenotype table to a Pandas data frame
 - 7.1.5 Convert data and check results
 - 7.1.6 Write the phenotypes out to a file and upload to the workspace

Text: *Univariate distributions* are easily visualized in histograms or density plots. We provide a function (`kdplot`) that will generate both types of plots, overlaid in a single figure. A continuously-valued variable corresponding to a column in the phenotype dataframe should be used as input, `ldl` in this example. The function is called with the following syntax:

```
kdPlot(samples, var = "ldl")
```

Figure: A histogram of the `ldl` phenotype with a normal distribution curve overlaid.

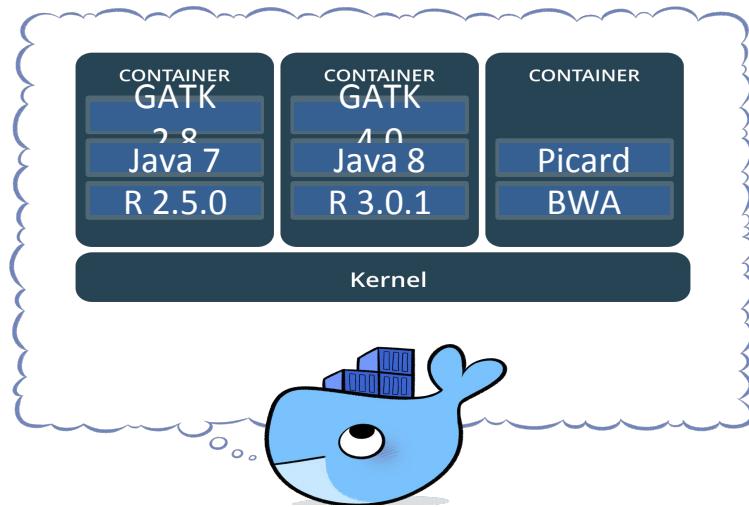
Text: *Bivariate distributions* can be visualized using a scatterplot. Use the function `bivariateDistributionPlot` to visualize two continuously valued variables. The `type` argument determines the type of plot generated and can be one of: "scatter", "reg", "resid", "kde", and "hex".

```
bivariateDistributionPlot(samples, var1 = "hdl", var2 = "whr", kind = "scatter")
```

Figure: A scatterplot showing the relationship between `hdl` (x-axis) and `whr` (y-axis).



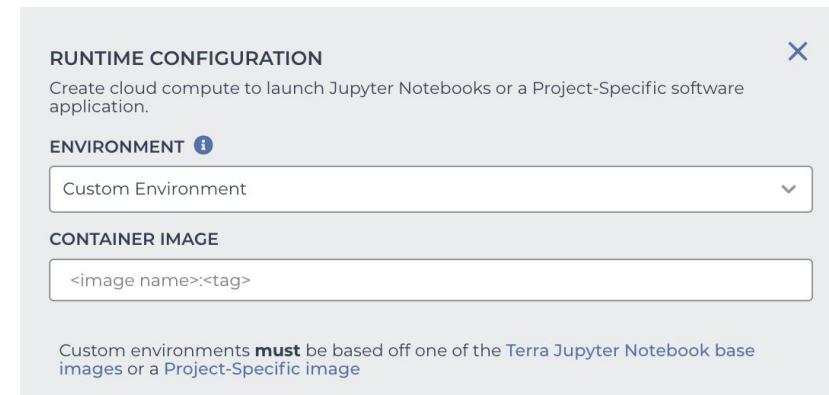
Containers for portability and reproducibility



Standardize the analysis environment by specifying a Docker container that includes the exact libraries and packages used for your analysis.

A Docker container encapsulates all the software dependencies associated with running a program

Takes the guesswork out of running workflows or notebooks on different platforms.



Anyone using the same Docker image will get the same results



How to collaborate and share in Terra

Workspaces are shareable and you're in control

Invite people to work with you; control who has access to the research assets in your workspaces

Workspaces are private by default. Sharing permissions include:

- ▶ Reader
- ▶ Writer
- ▶ Owner

Enforce privacy and security with Authorization Domains

Owners of controlled-access data can restrict the list of people with whom any derived workspaces can be shared



Security is a priority that's built in to the platform

- ▶ Certified FISMA moderate from the NCI for hosting controlled-access data owned by NCI and the NIH
- ▶ Leverages security features in Google Cloud Platform for Federal Risk and Authorization Management Program (FedRAMP) authorization



How much it costs and how to pay

- **Terra itself won't cost you anything** - To enable scientists to get work done, the open-source platform is available and free for everyone
- **Pay only for Google Cloud Platform** - Storage, compute, and egress
- **Terra Billing Accounts linked to Google Billing** - Securely-configured billing project linked to Billing Account on the Google Cloud Platform can be shared with members of your lab or department for use in Terra.



Spend time doing science... Not wrangling software

Documentation

- ▶ [Terra KnowledgeBase](#)

Support Forum

- ▶ [Community forum](#)

For hands-on practice try

- ▶ [Workflows-QuickStart](#)
- ▶ [Notebooks-QuickStart](#)

Terra Landing Page

- ▶ <https://terra.bio/>

15 minute break!

Reminder:

- Do you have a Terra account?
- Do you have access to a billing account on Terra?

Workshop Overview:

- Logistics (5 m)
- Overview of the Data Biosphere (10 m)
 - The AnVIL as an example
- Introduction to Dockstore (20 m)
 - ‘App store’ for reproducible bioinformatics workflows
- **Introduction to Terra (25 m)**
 - User friendly workspaces for analysis in the cloud

Break (15 m)

- Tutorial: Analyzing COVID-19 viral genomes with Dockstore & Terra (60 m)
- Wrap up

Tutorial

Analyzing COVID-19 viral genomes with Dockstore & Terra