Reproducible Analysis in the Cloud with Dockstore & Terra

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

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

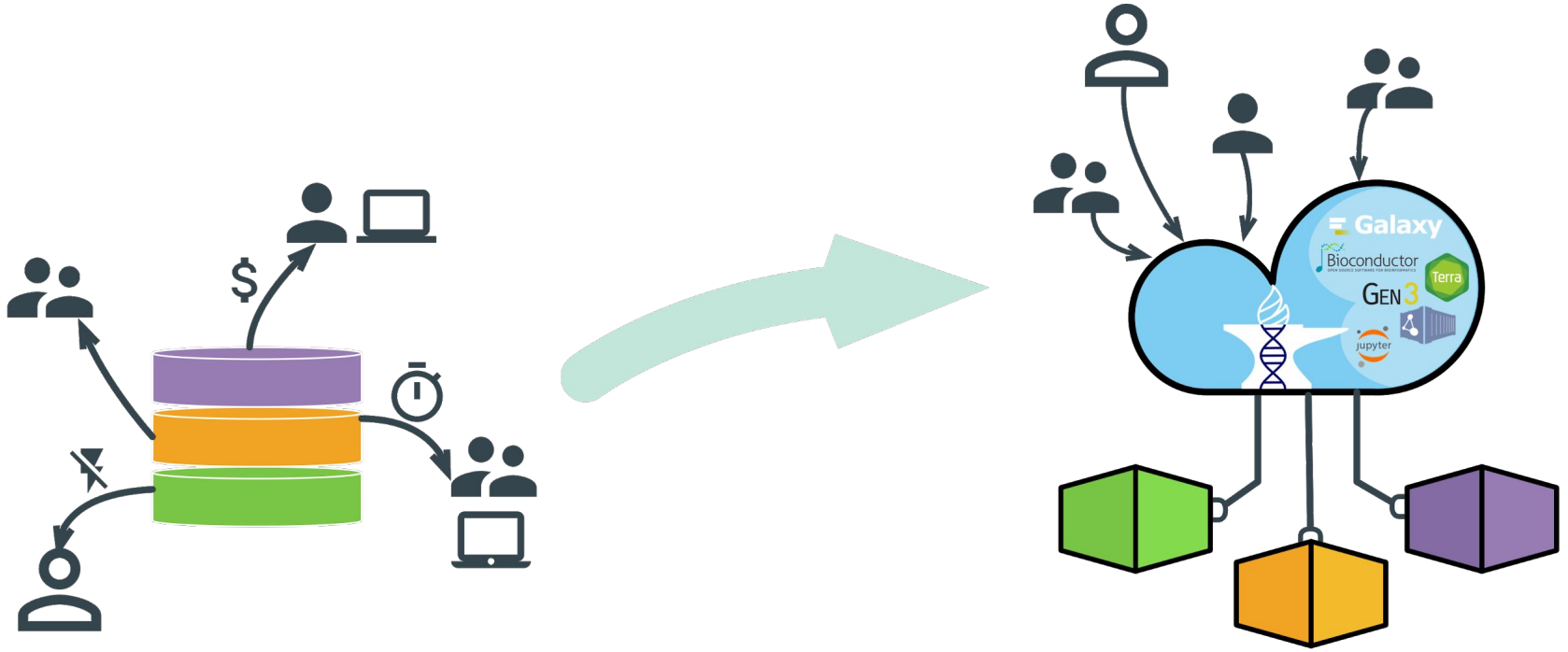
- **Please keep your mic on mute when you are not talking**
- **Ask questions in chat**
  Members of the team will be available to answer questions in the chat box.
  You can also privately our support team: **Beri S., Beth S., Louise C., Mo H.**
- **Raise your hand feature**
  If you would like to ask 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

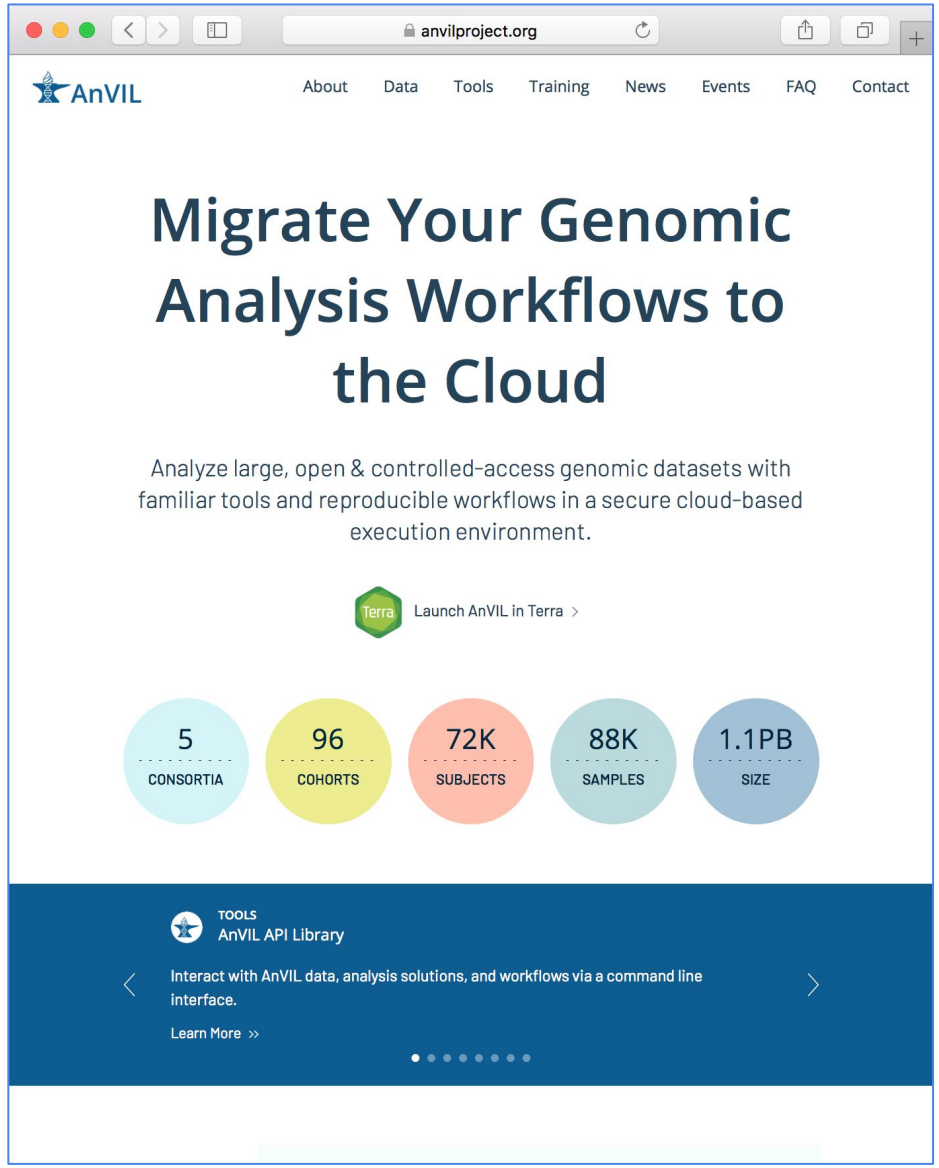
How should a data biosphere be structured?

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

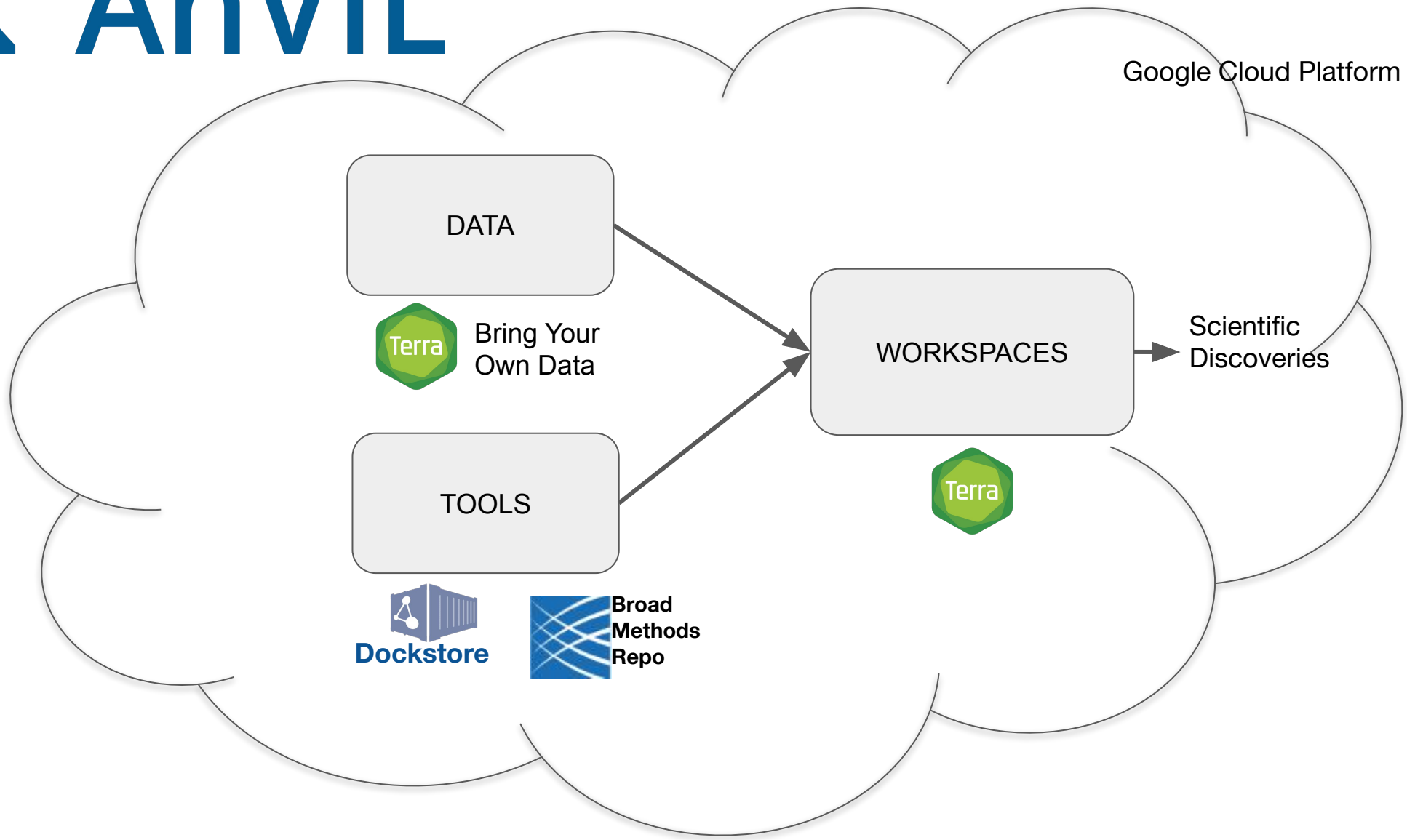
What is the NHGRI's AnVIL?

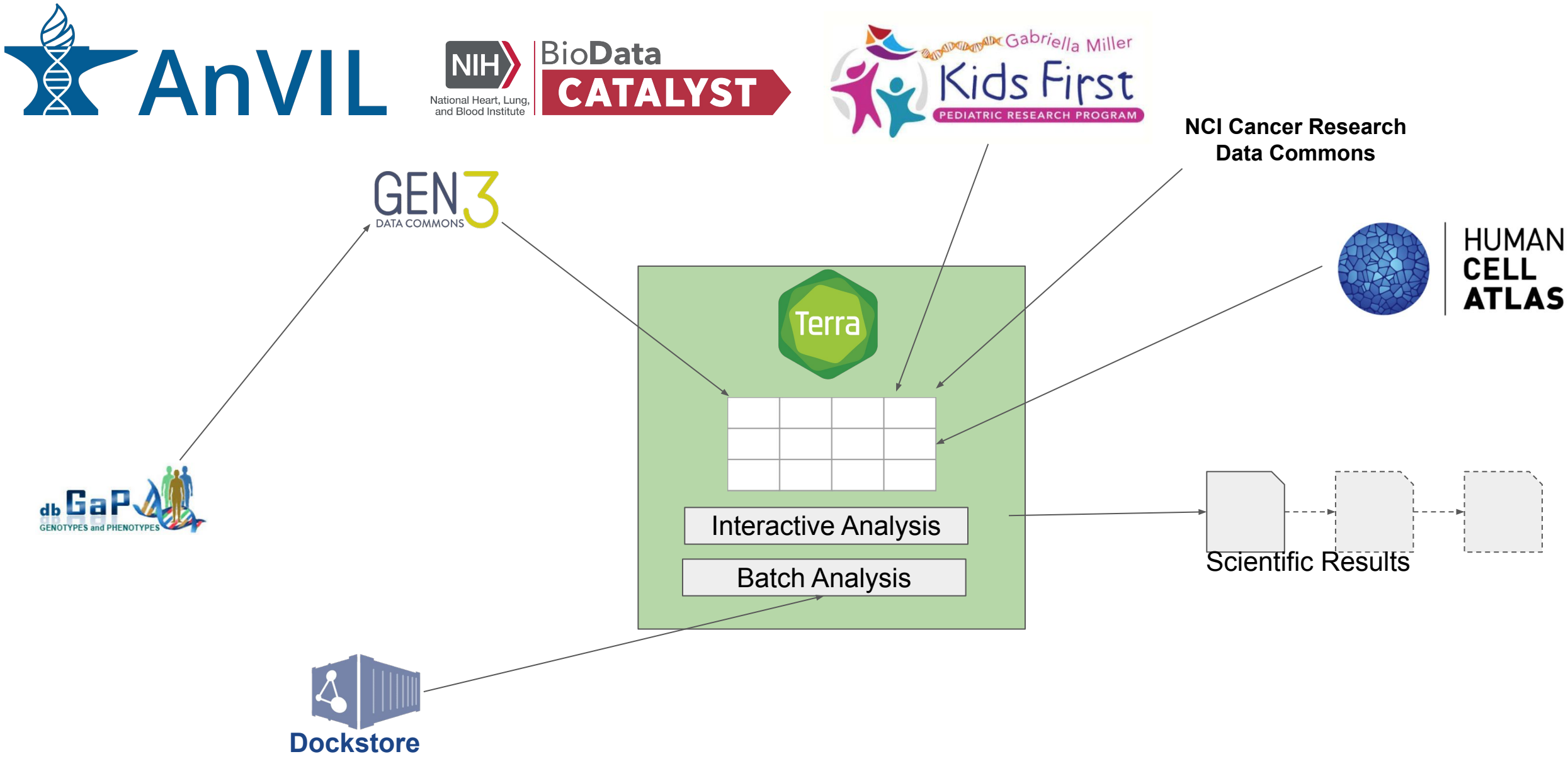
Scalable and interoperable computing resource for the genomics scientific community

- **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



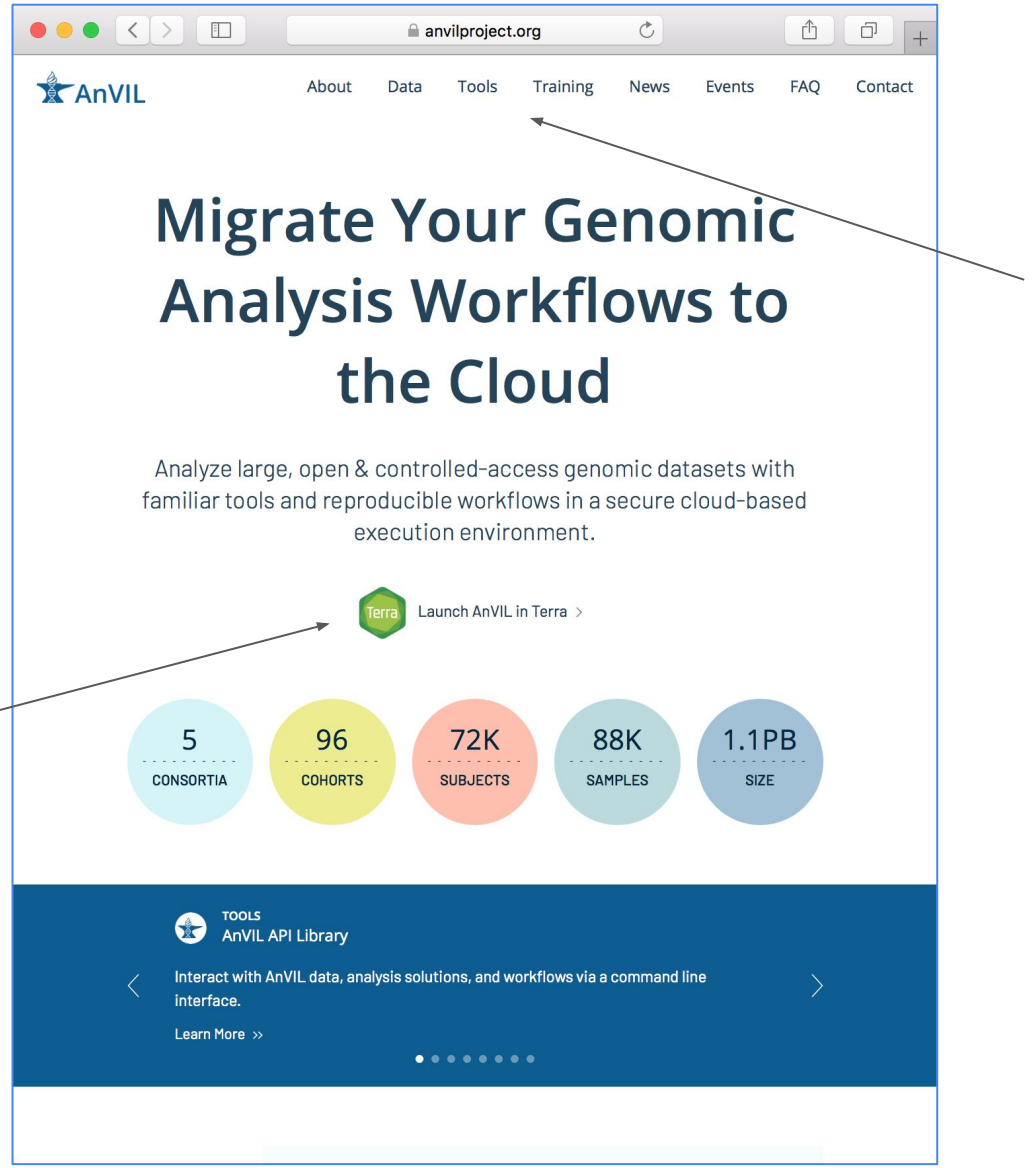
AnVIL


Google Cloud Platform
DATA
Terra
Bring Your
Own Data
TOOLS
Dockstore
Broad
Methods
Repo
WORKSPACES
Terra
Scientific
Discoveries


AnVIL
NIH
National Heart, Lung, and Blood Institute
BioData CATALYST
Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
NCI Cancer Research Data Commons
GEN3
DATA COMMONS
HUMAN CELL ATLAS
Terra
db GaP
GENOTYPES and PHENOTYPES
Interactive Analysis
Batch Analysis
Scientific Results
Dockstore

How to start using the AnVIL?

anvilproject.org

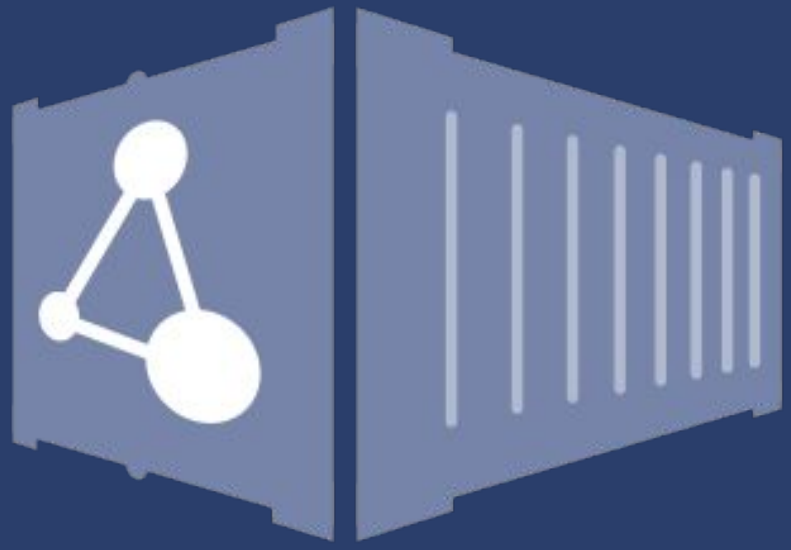


Workshop Overview:

- 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

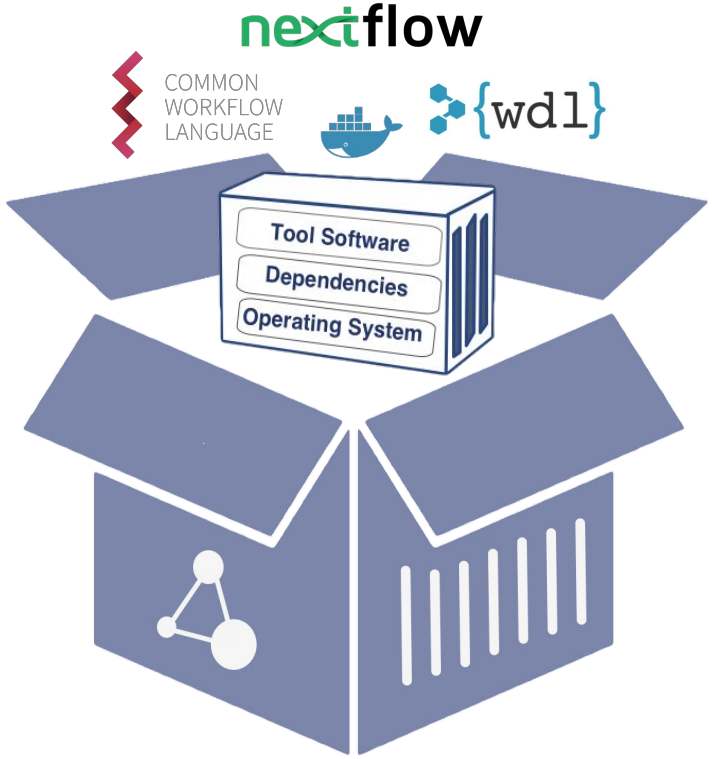
OICR Ontario Institute for Cancer Research

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

What is Dockstore?

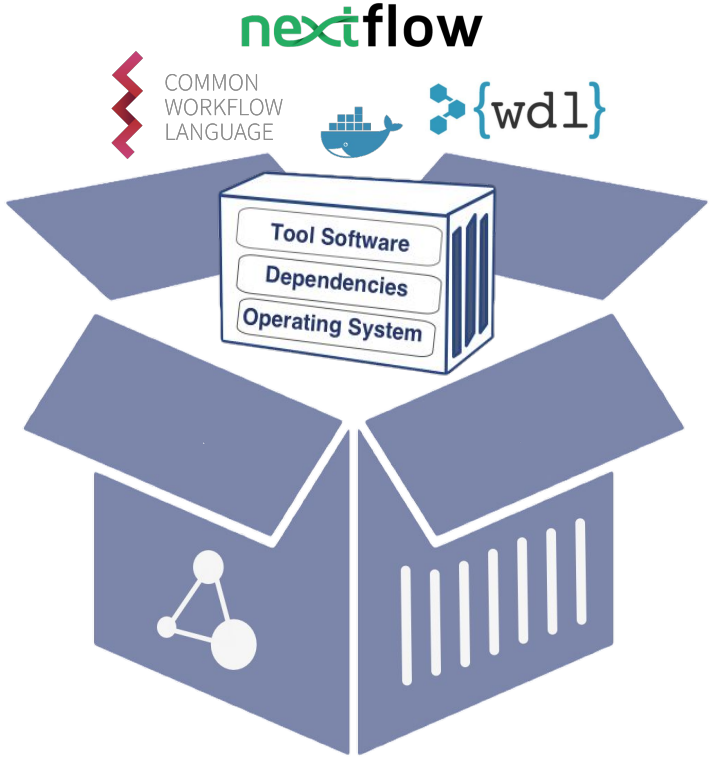
Dockstore is a free and open source platform for sharing scientific tools and workflows. It is a registry of Docker-based resources described using popular workflow languages CWL, WDL, Nextflow and now Galaxy!

- **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

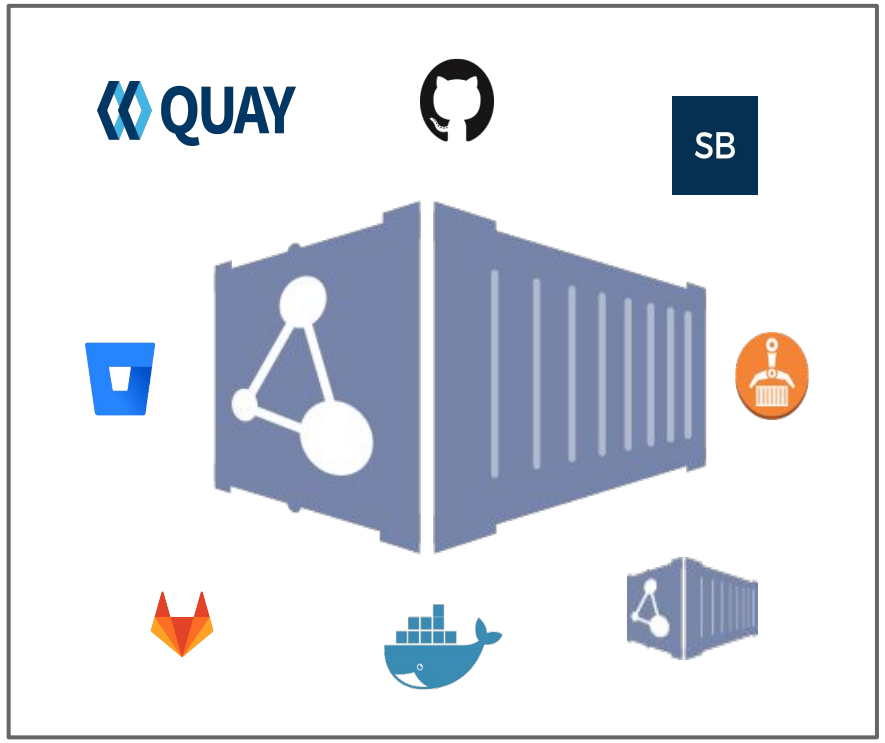
A researcher wants to do some type of genomics work (assembly, alignment, variant calling, GWAS, etc.)

- 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

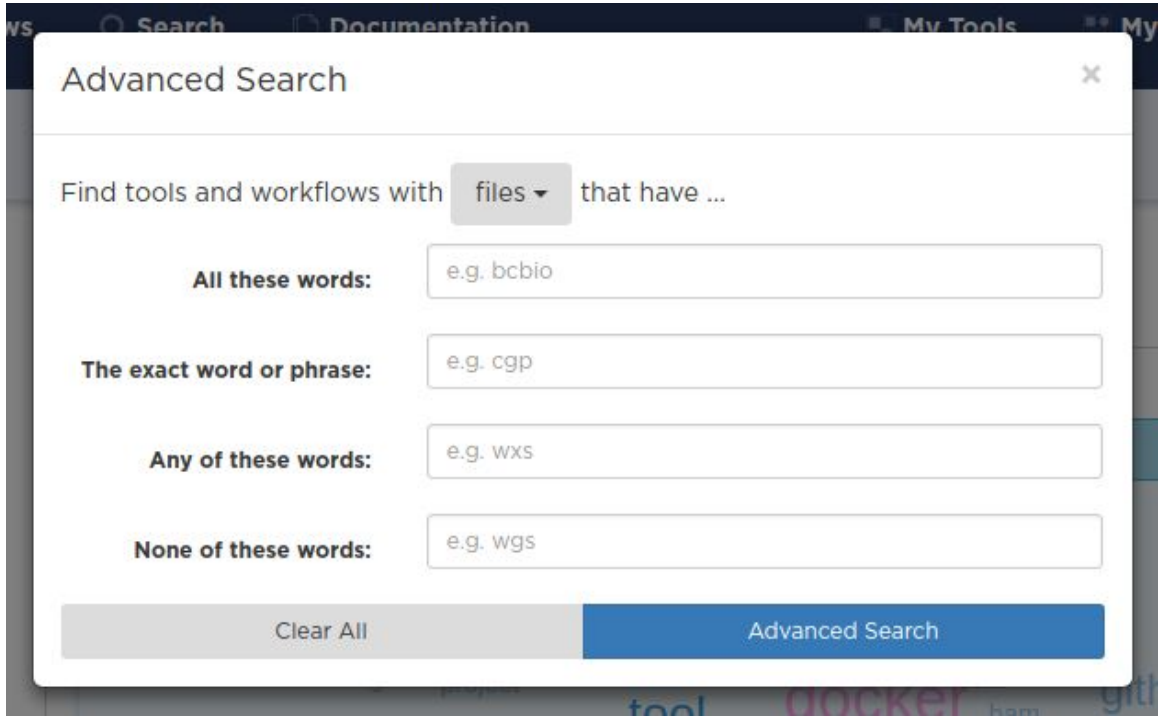
Searchable and **Centralized** catalogue of bioinformatics tools and workflows

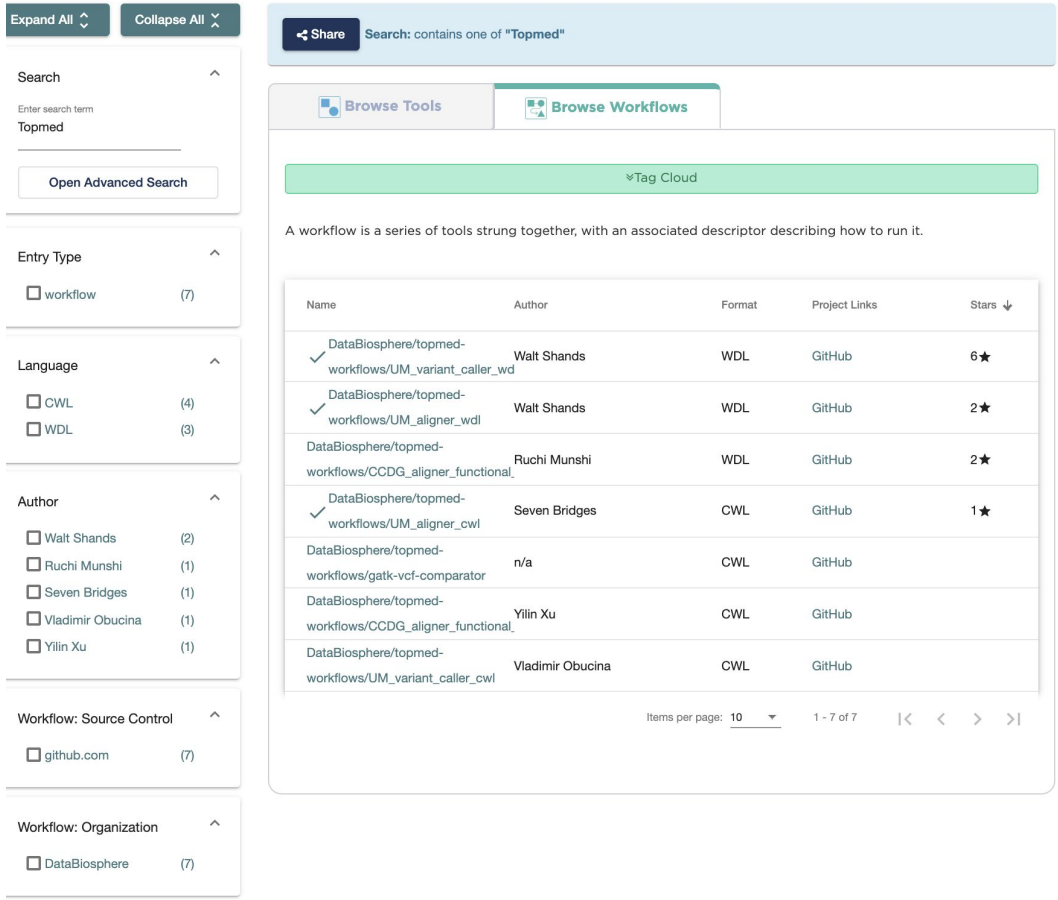
- Many ways to get tools and workflows into Dockstore!
  - External Hosting: Github, Bitbucket, Gitlab, Quay.io, Docker Hub
  - Direct Hosting (descriptors)



Search Functionality

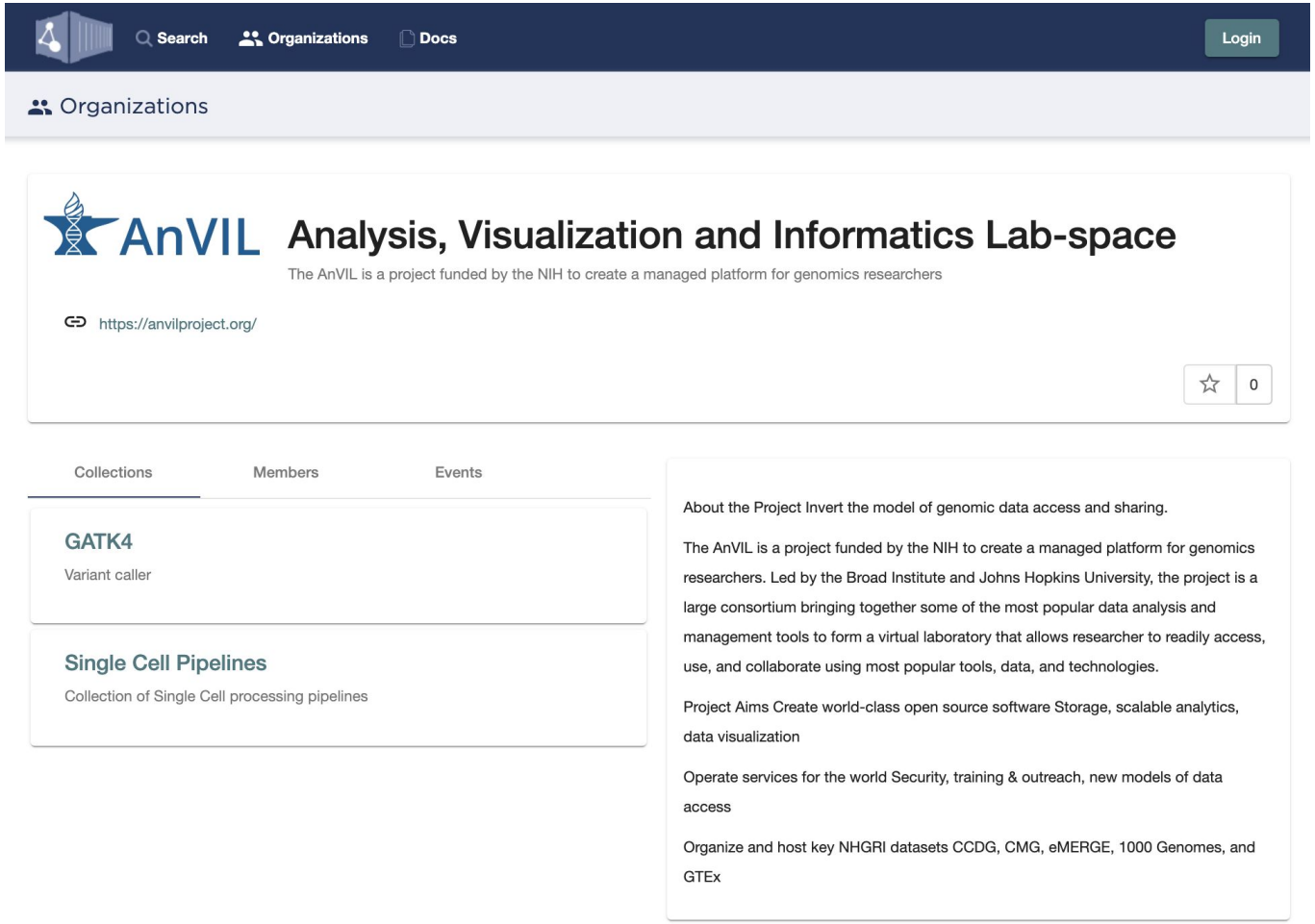
- Search by name, author, organization, descriptor, etc
- Filter results using facets
  - Descriptor Language
  - Labels





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



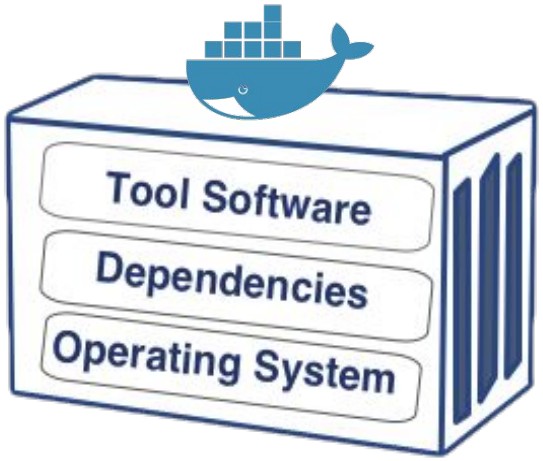
Common Use Case: Simplifying Installation Process

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



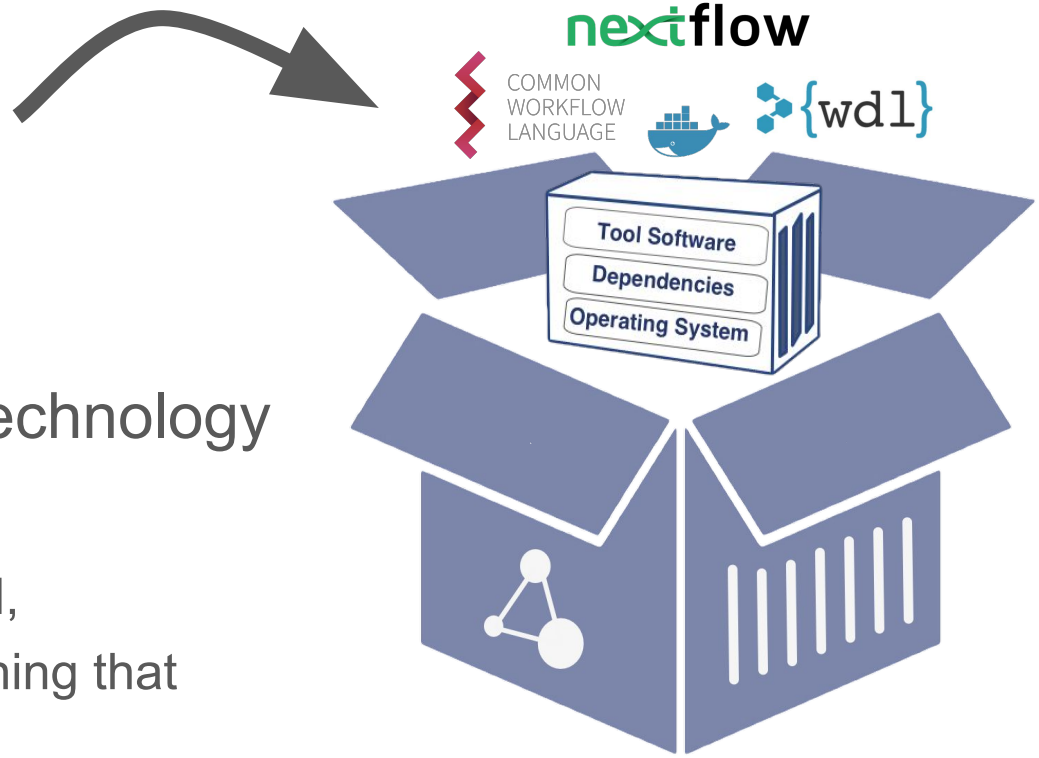
\+

Descriptor

COMMON WORKFLOW LANGUAGE

{wdl}

nextflow

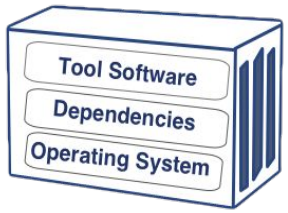


Software is “packaged” using containerization technology and described using descriptor language

- 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

Tool



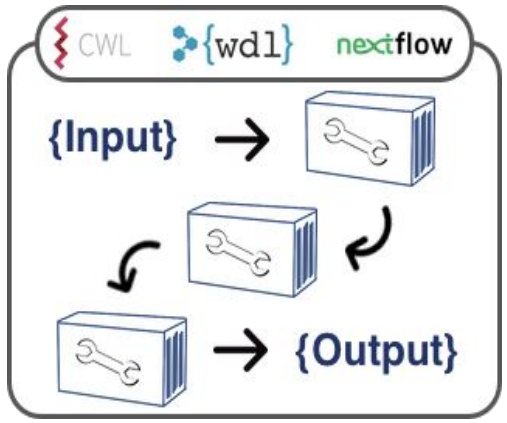
COMMON WORKFLOW LANGUAGE {wdl}

Container + Descriptor

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

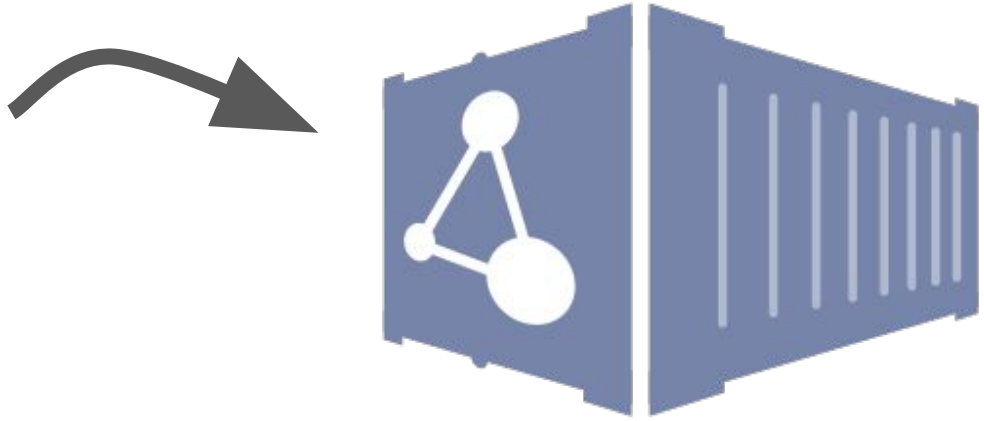
OR

Workflow



Workflow: Tools + Descriptor

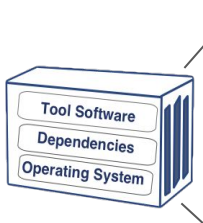
A workflow can use multiple containers and executes multiple actions or steps, still outlined by a descriptor



Content: tool vs workflow (example)

Tool

BWA-MEM



Debian OS
binaries
C
GNU Make
GCC
zlibs

COMMON WORKFLOW LANGUAGE

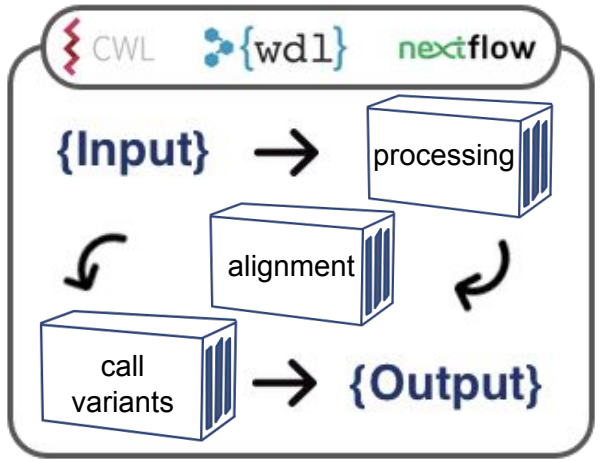
{wdl}

A tool is one.

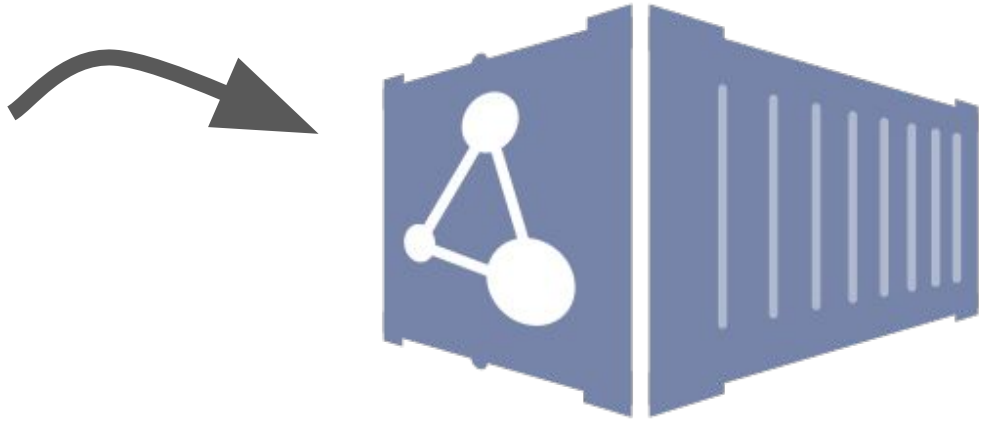
OR

Workflow

Variant Calling Pipeline



A workflow is many.



*note: you *can* register a single tool as a workflow on Dockstore, but a multi-step workflow cannot be registered as a tool.

ex: BWA-MEM can technically be registered as both

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

COMMON WORKFLOW LANGUAGE

{wdl}

nextflow

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

COMMON WORKFLOW LANGUAGE

{wdl}

nextflow

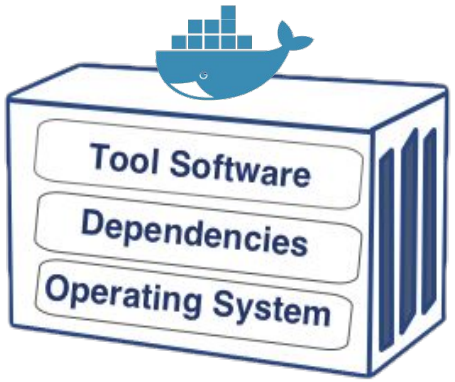
Parameter File (wdl, cwl):

- 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

```
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

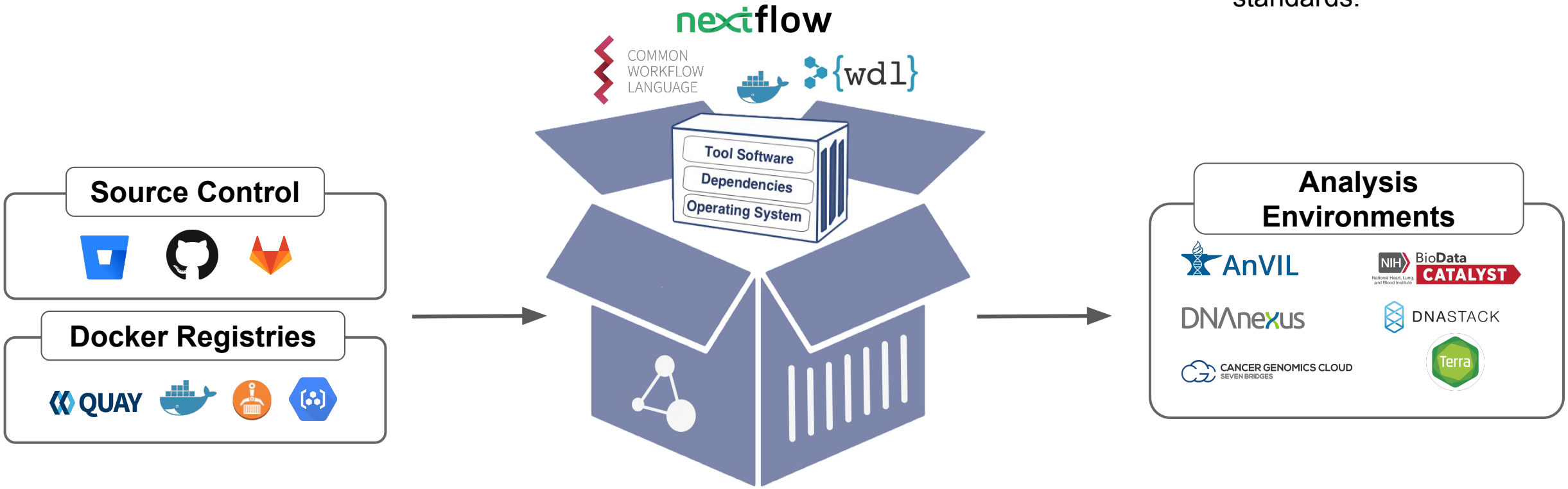
metrics.wdl

```
version 1.0

workflow metrics {
   call flagstat
   output { File align_metrics = Flagstat.metrics }
}

task flagstat {
    input {
      File input_bam
    }

    String stats = basename(input_bam) + ".metrics"
    command {
       samtools flagstat ${input_bam} > ${stats}
    }
    output { File metrics = "${stats}" }

    # set some parameterized runtime parameters
    runtime {
    docker: "quay.io/ldcabansay/samtools:latest"
    }
}
meta {
    author: "Foo Bar"
    email: "foobar@institute.edu"
}
```

Dockstore Ecosystem

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

This powerful interoperability is possible through GA4GH API standards.

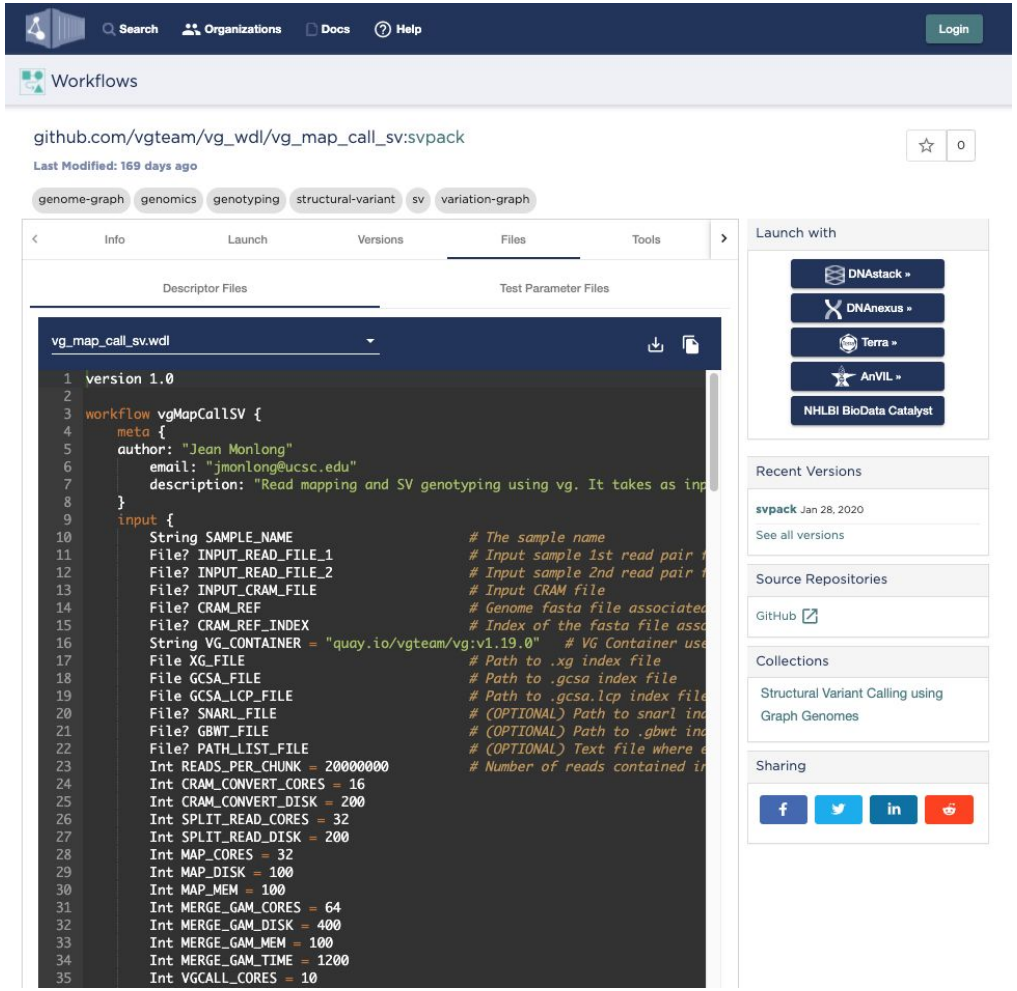


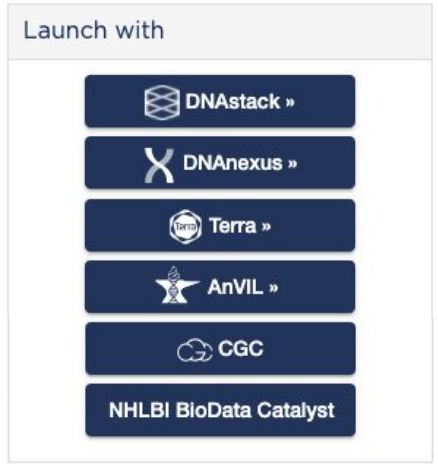
Store your descriptors and containers and descriptors on your preferred sites

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





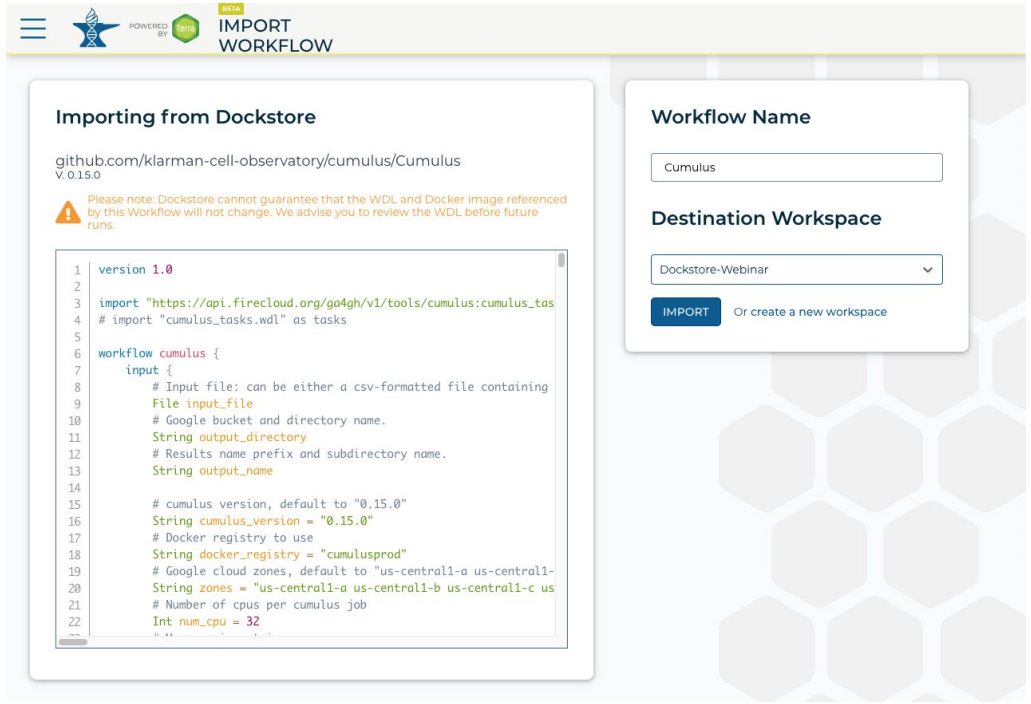
Partner Platforms

- DNAstack
- DNAnexus
- 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



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

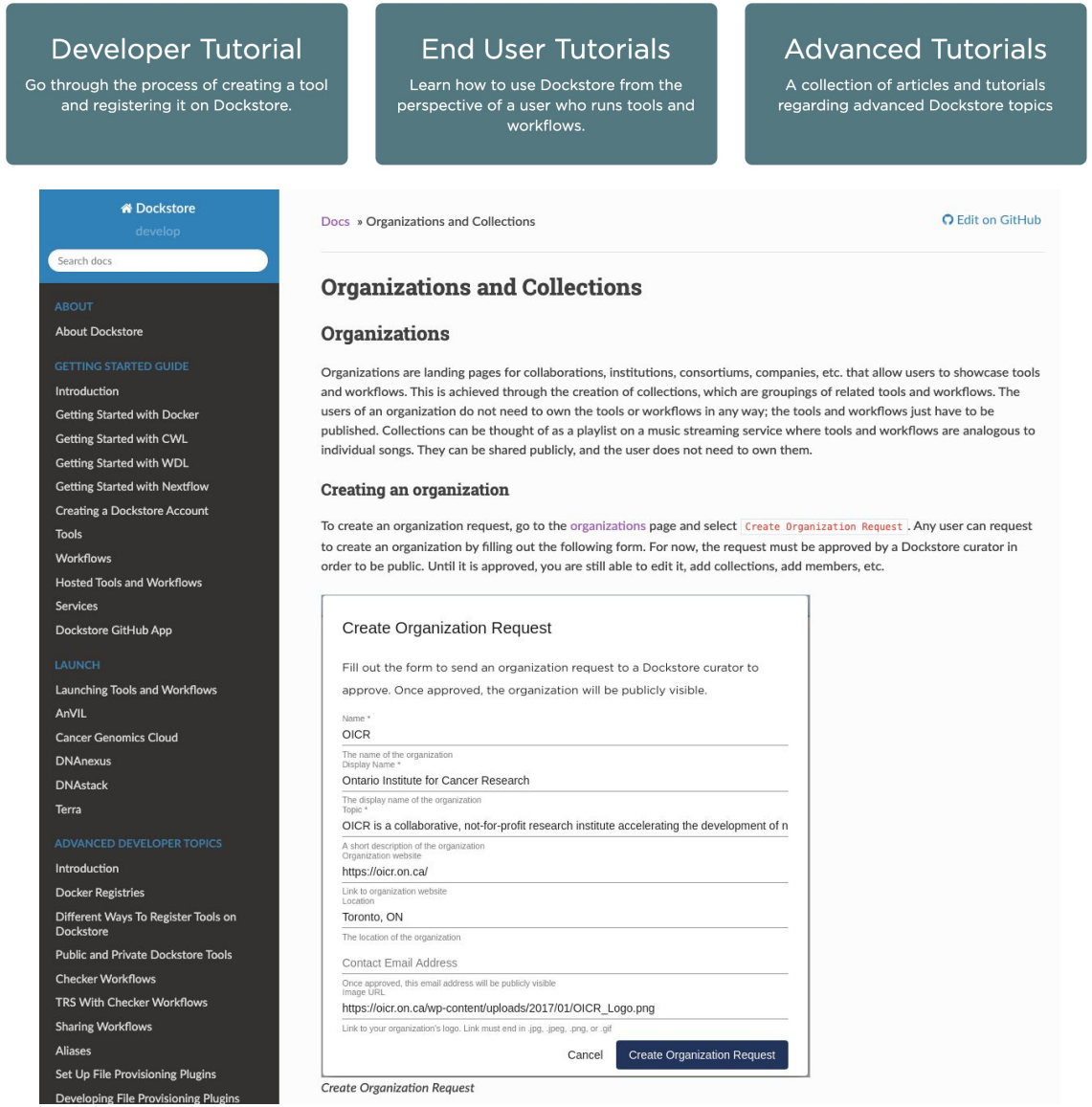
Ex: A team of collaborators from X different institutions are working together on a study that analyzes the genomes of thousands of individuals.

- 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 Identifers (DOIs)
  - Creating Organizations
  - Writing checker workflows
  - And many more!

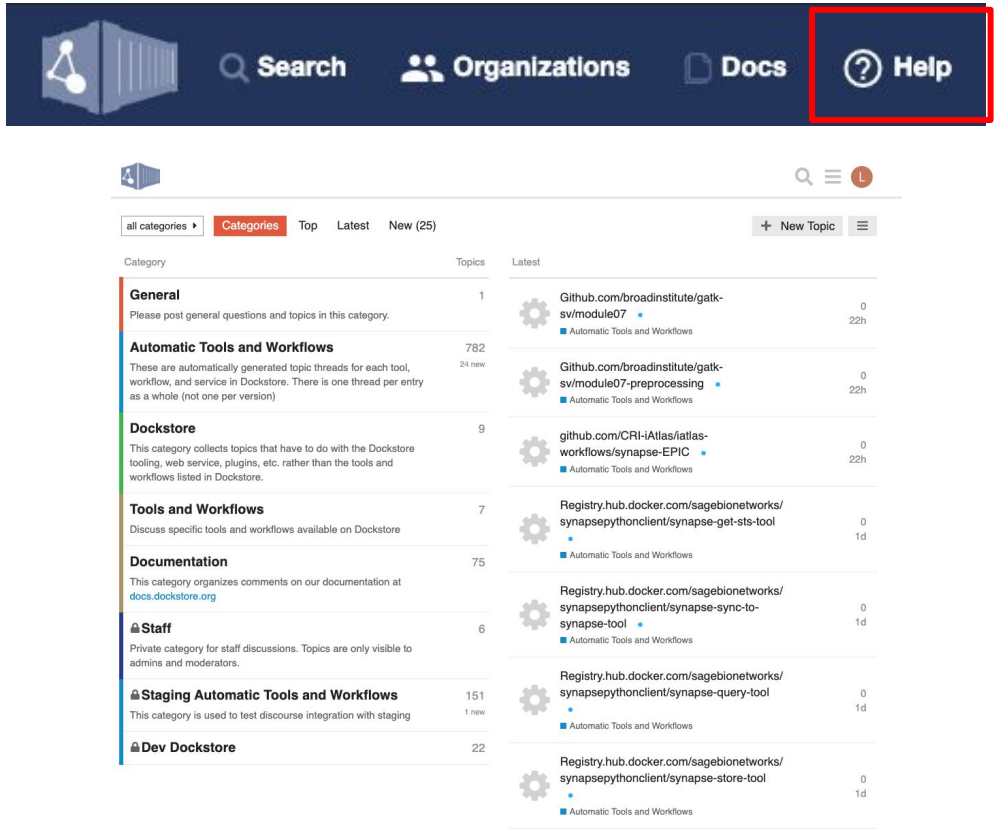
https://docs.dockstore.org/



Getting Help on Dockstore

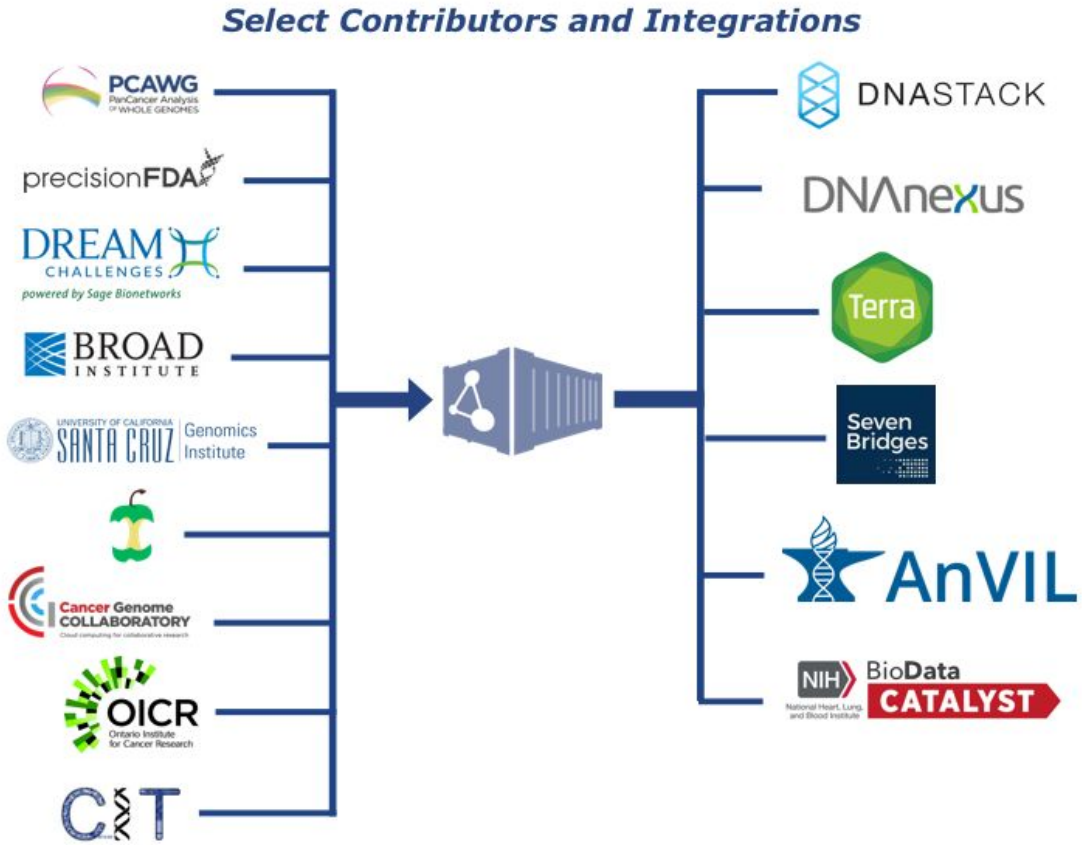
User forum at https://discuss.dockstore.org/

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



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 occuring in this space.



The Dockstore Team

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

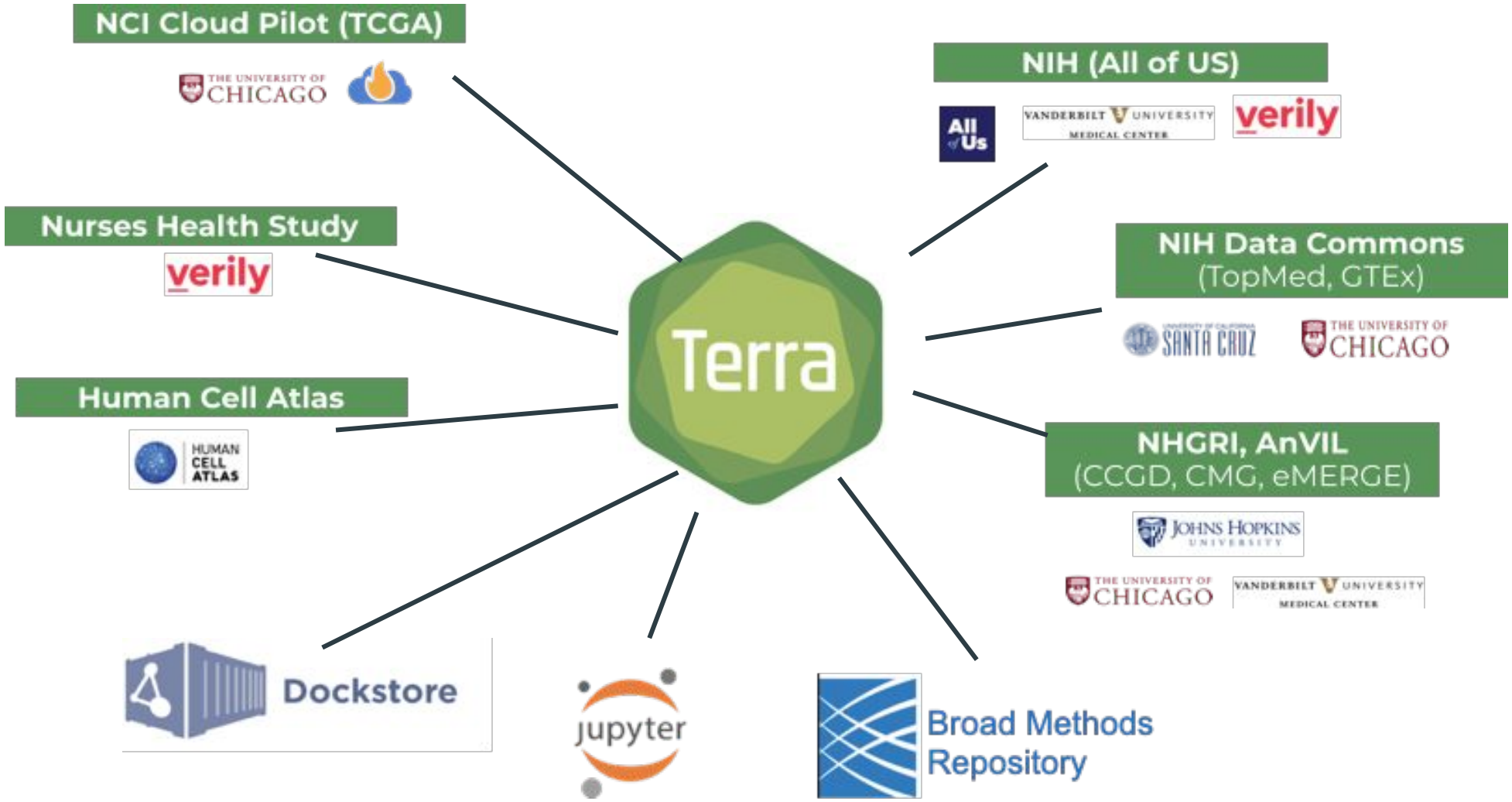
Data Sciences Platform, Broad Institute of MIT and Harvard

BROAD INSTITUTE

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



BROAD INSTITUTE

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

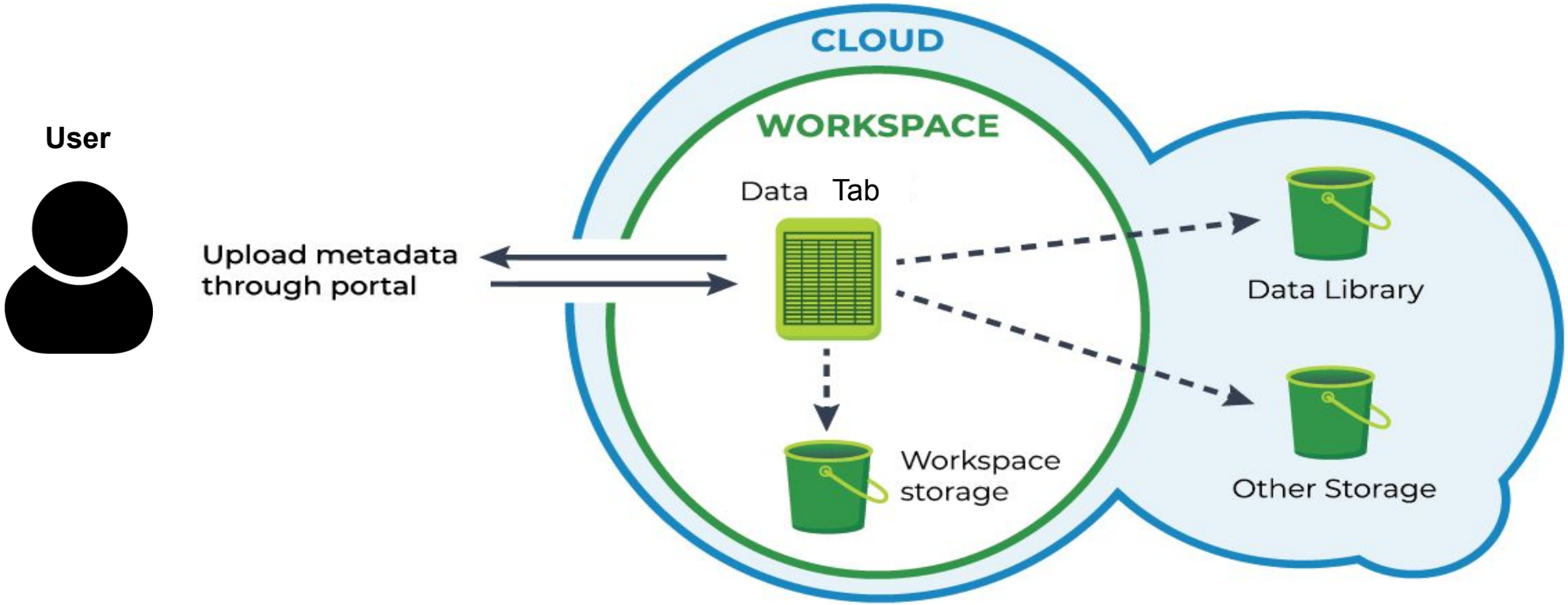
https://terra.bio

BROAD INSTITUTE

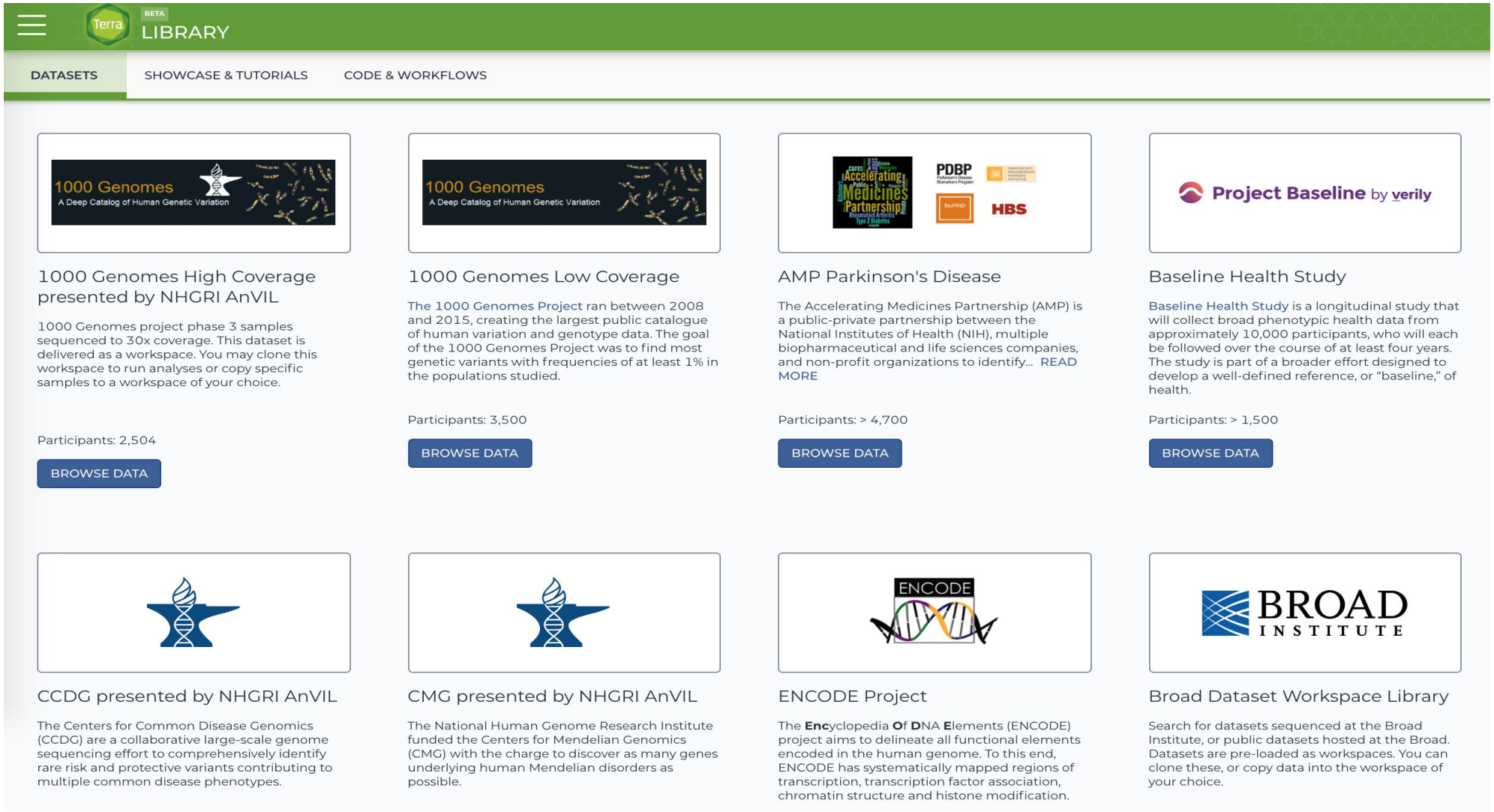
The Workspace - The fundamental unit in Terra

BROAD INSTITUTE

Organize data and tools in the Terra workspace



Find and access data in the Terra Data Library



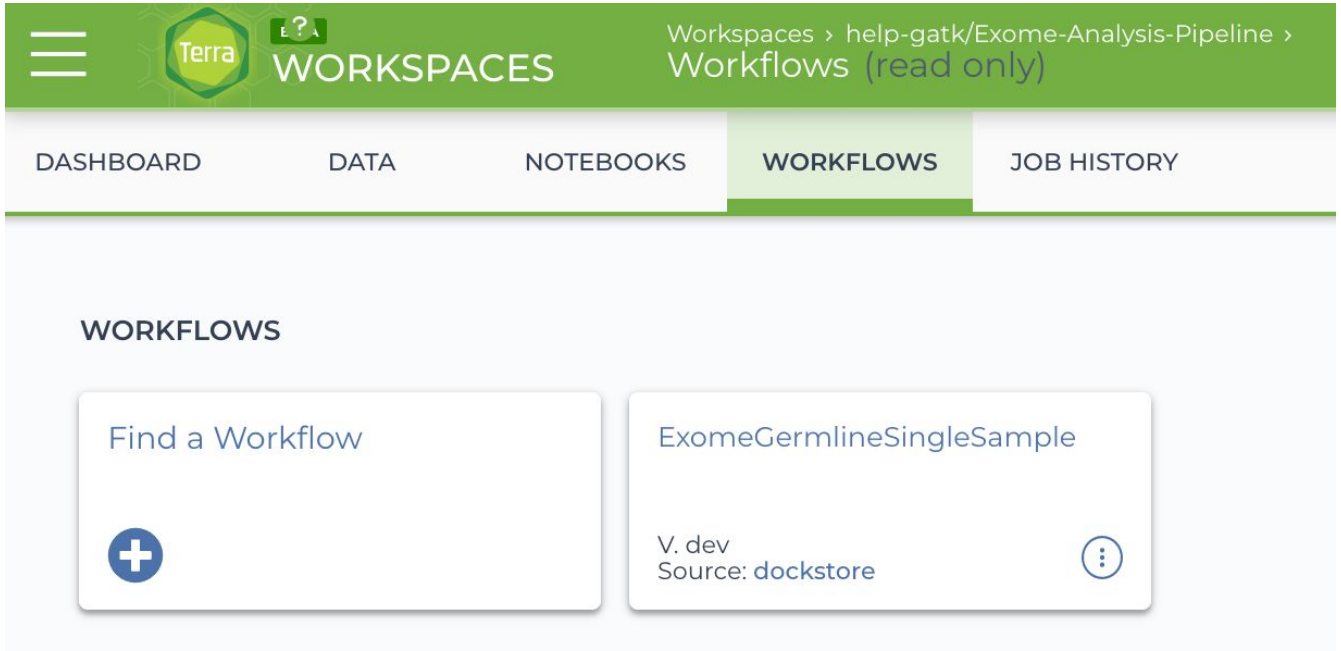
BROAD INSTITUTE

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



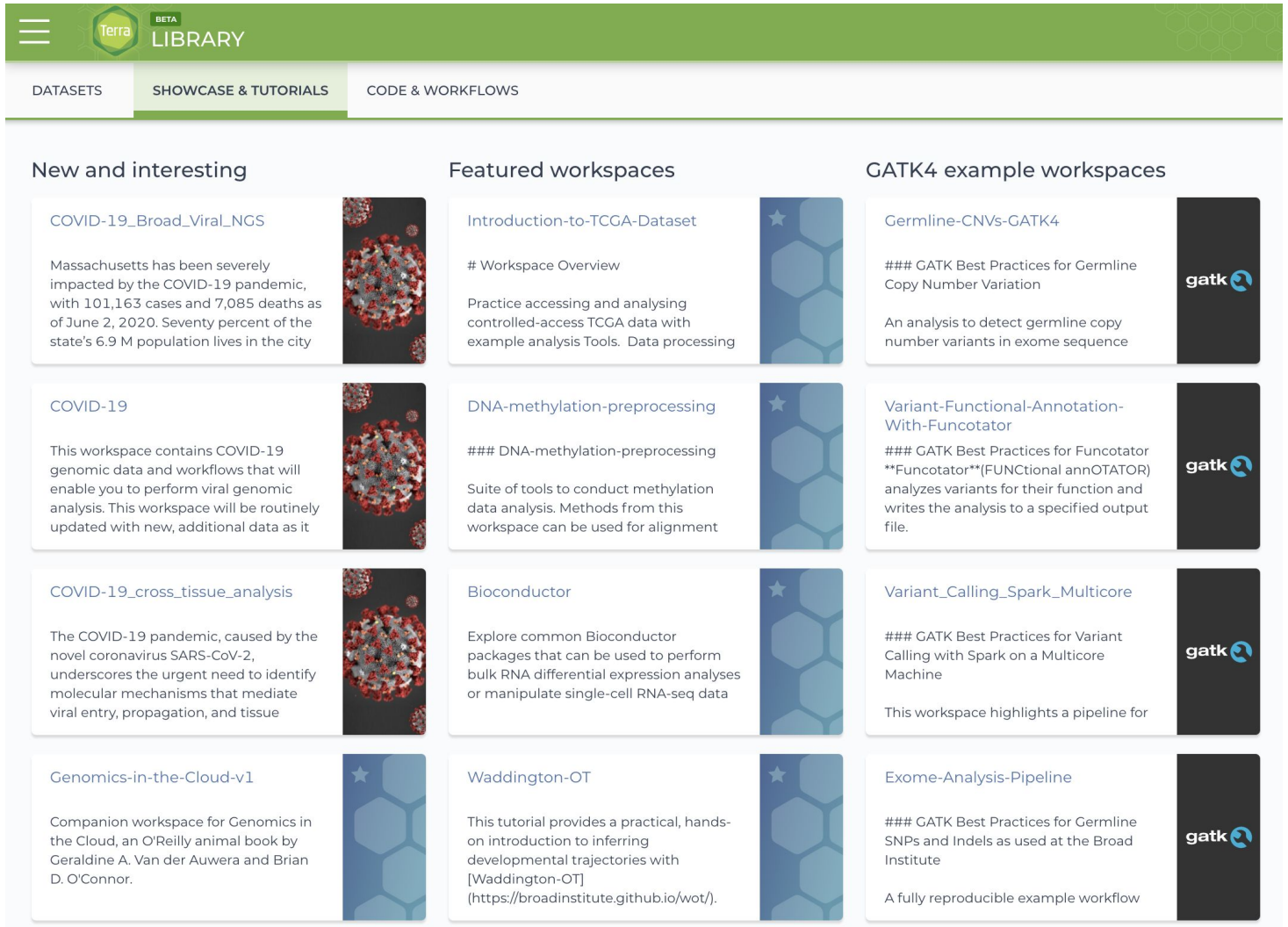
Dockstore

Broad Methods Repository

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

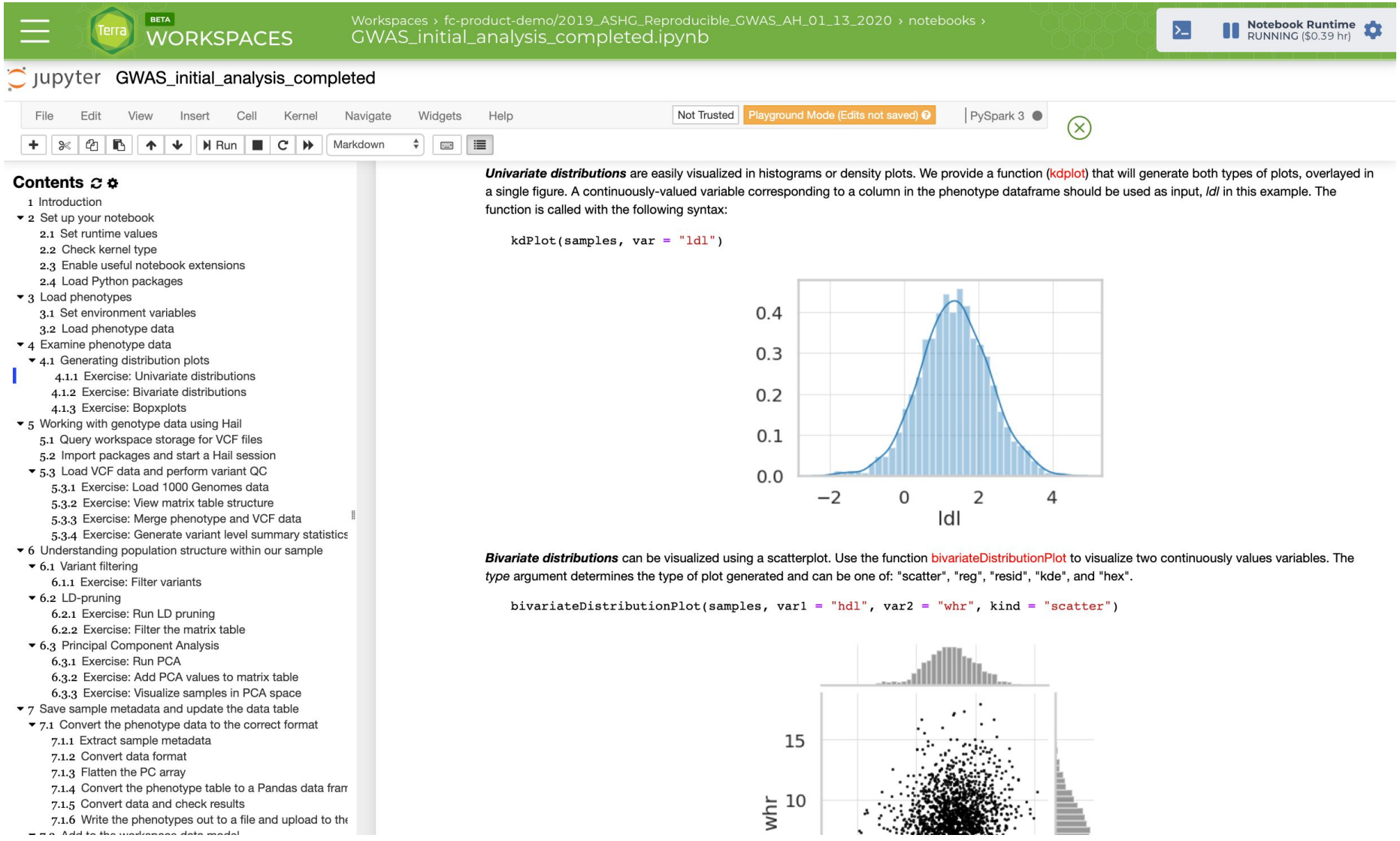
BROAD INSTITUTE

Bulk analysis: GATK4 workflows preloaded in showcase workspaces

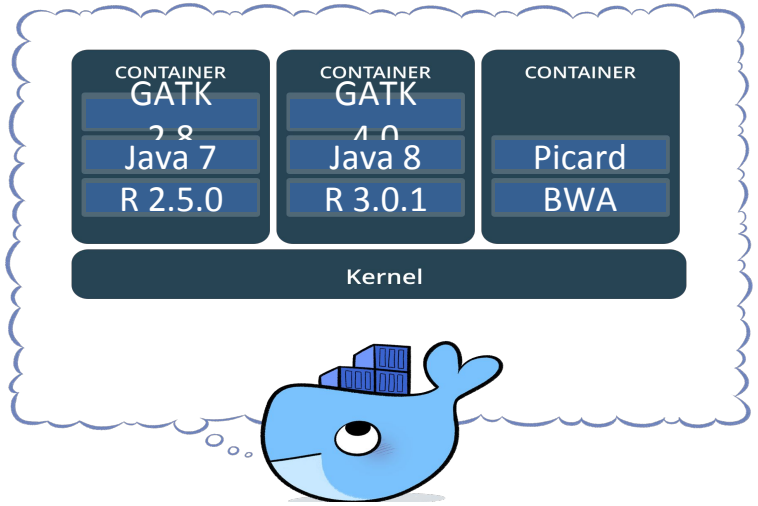


BROAD INSTITUTE

Interactive analysis with Jupyter Notebooks



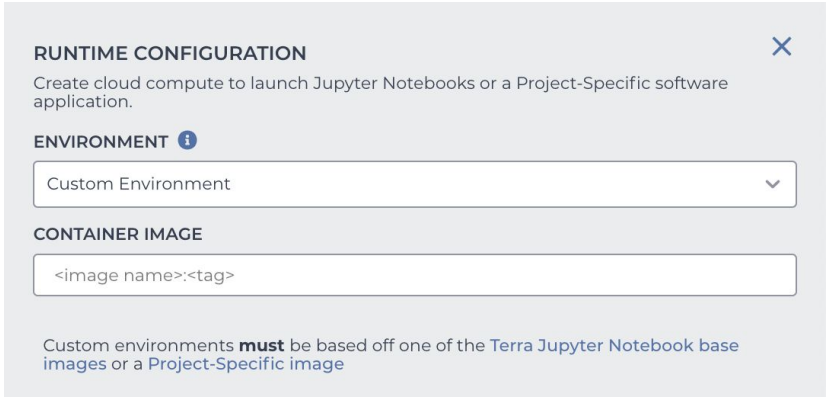
Containers for portability and reproducibility



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.

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



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

BROAD INSTITUTE

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

BROAD INSTITUTE

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/

BROAD INSTITUTE

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