

NHLBI BioData Catalyst Community Hours

A Tour of the Analysis Workspaces
with Seven Bridges and Terra

August 25 at 1 PM EDT



National Heart, Lung,
and Blood Institute

BioData

CATALYST

Statement of Conduct

The BioData Catalyst Consortium is dedicated to **providing a harassment-free experience for everyone**, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, or religion (or lack thereof). We do not tolerate harassment of community members in any form. Sexual language and imagery is generally not appropriate for any venue, including meetings, presentations, or discussions.

<https://bdcatalyst.gitbook.io/biodata-catalyst-documentation/community/statement-of-conduct>

Agenda

- **Welcome!**
- **Topic presentations**
 - Introduction to workspaces (10 minutes)
 - Tour of Terra (10 minutes)
 - Tour of Seven Bridges (10 minutes)
 - Requesting cloud credits and getting started on each platform (5 minutes)
- **Discussion and questions**

Welcome!

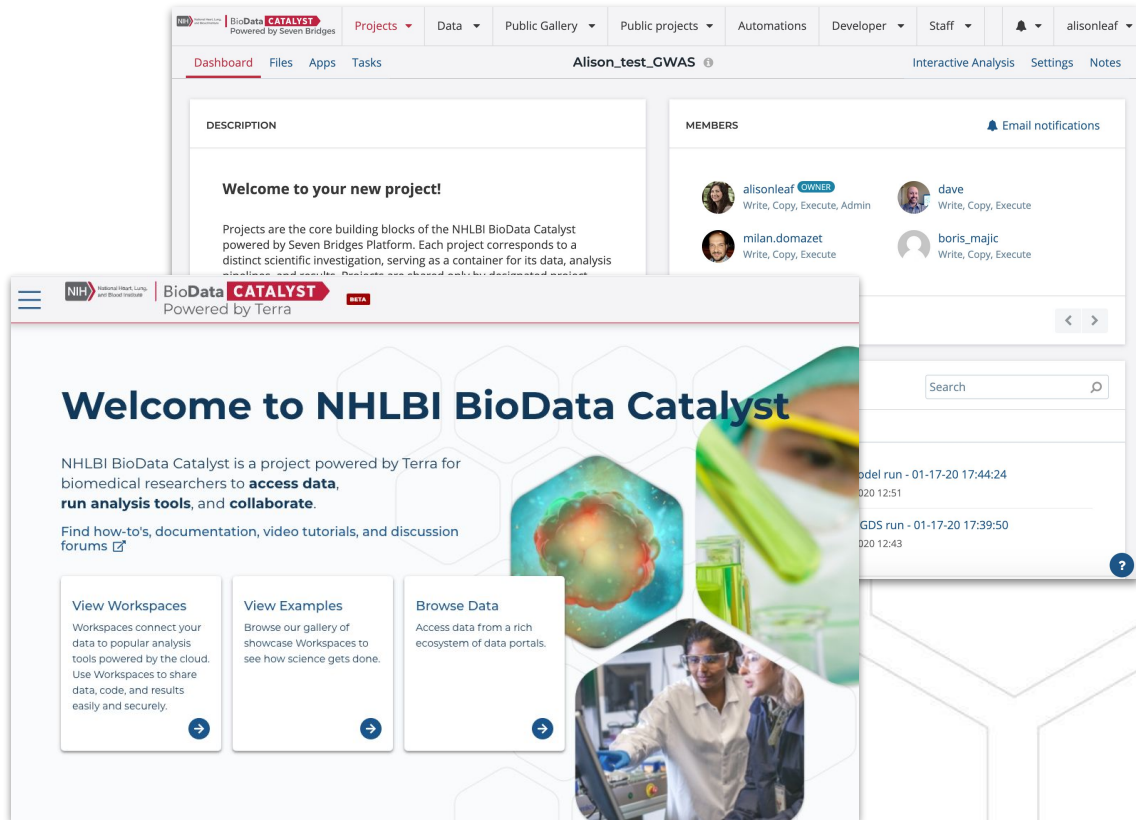
- Host and co-hosts
 - **BDC3:** Amber Voght, Paul Kerr
 - **Seven Bridges:** Dave Roberson, Tony Patelunas
 - **Terra:** Tiffany Miller
- Notes, slides, and video will be available on the [community forum](#)
- Questions: Use chat at any time **or** use audio during discussion time
- We encourage you to submit unanswered questions, no matter how big or small, to our [help desk](#)
- Join the ecosystem: <https://biodatacatalyst.nhlbi.nih.gov/contact/ecosystem>

Introduction to Workspaces

Tiffany Miller

Harness the computing power of the cloud

Stop waiting in line for your university cluster: Set up workspaces on BioData Catalyst and process 1000s of samples at once using instances on AWS or Google Cloud



Organize files, methods, and results in Workspaces/Projects

- Workspaces serve as a place to organize files and tools.
- Set up and kick off analyses.
- Functionality to organize files & metadata.

WORKSPACES terra-outreach/DEMO-Working-with-gnomAD...

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

ABOUT THE WORKSPACE

This workspace demonstrates several options for working with cloud-hosted gnomAD data from within Terra, intended as a companion to [this blog post](#).

Disclaimer: The code examples in this workspace are provided as a proof of concept and may not remain functional over time.

Working with gnomAD in Terra

As announced on the [gnomAD blog](#) in October 2020, the entire gnomAD dataset is now available for direct use or download from Google Cloud as well as Amazon Web Services and Microsoft Azure. (If you're not familiar with gnomAD, see the appendix at the bottom of this page).

There are several ways that you could interact with the [gnomAD data hosted by Google Cloud](#) from within Terra, so we developed this workspace to illustrate some approaches that we thought might be useful:

- Running a workflow on the callset VCFs
- Exploring the callset using Hail in a notebook
- Exploring the callset using BigQuery in a notebook

We focused on the variant callsets included in the v3.1 release, which is the latest available at the time we write this. For a full inventory of the dataset across all releases, see the [Downloads page](#) of the gnomAD website, which provides item-by-item links to the data on each of the participating cloud platforms.

The gnomAD v3.1 callsets: contents and available formats

The main resource that we typically see people derive from gnomAD (outside of the excellent [gnomAD browser](#)) is the **sites-only form of the overall callset**, which contains all variant calls made across all the project samples, but without any individual-level genotypes.

As [described on the gnomAD blog](#), the v3.1 release also includes a **callset of individual-level genotypes** for a

WORKSPACE INFORMATION

CREATION DATE 11/4/2020	LAST UPDATED 11/5/2020
SUBMISSIONS 1	ACCESS LEVEL Reader
GOOGLE PROJECT ID terra-outreach	

OWNERS

[vdauwera@broadinstitute.org](#)
[schaluva@broadinstitute.org](#)

TAGS

demo gatk genomics gnomAD Hail

Google Bucket

Name: fc-15ed1113-c2db-430d-952a...
Location: multi-region: US
[Open in browser](#)

Example from BioData Catalyst Powered by Terra

Work privately or alongside collaborators

If you are the only member of a project, you are the only person who will have access to your data, tools, and analyses.

To collaborate, add members to project. Administrative capabilities allow for granular permissions to limit what project members can see and do.

The screenshot displays the BioData CATALYST web interface. The top navigation bar includes links for Projects, Data, Public Gallery, Public projects, Automations, Developer, Staff, and a user profile dropdown for 'alisonleaf'. The main content area is titled 'Alison_test_CWAS' and features a 'DESCRIPTION' section with a welcome message and a list of actions within the project. A 'MEMBERS' section lists three users: alisonleaf (OWNER), dave, and milan.domazet. A 'Manage members' modal is open, showing a list of current members and an 'Invite new members' section with a search bar and a dropdown menu for permissions.

BioData CATALYST
Powered by Seven Bridges

Projects Data Public Gallery Public projects Automations Developer Staff alisonleaf

Dashboard Files Apps Tasks Alison_test_CWAS Interactive Analysis Settings Notes

DESCRIPTION

Welcome to your new project!

Projects are the core building blocks of the NHLBI BioData Catalyst powered by Seven Bridges Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects are shared only by designated project members.

Within your project, you can:

- Start [exploring public datasets](#) straight away
- [Install your tools on the platform](#) and create workflows
- [Upload your own private data](#) and analyze it along with public datasets
- [Collaborate securely](#) with other researchers

Please record the details of your project here, such as its aims, experimental context, and your project members. Remember you run on the platform are for your own notes.

You can also use markdown

Good luck with your research! [Knowledge Center](#)

MEMBERS [Email notifications](#)

alisonleaf **OWNER** Write, Copy, Execute, Admin
dave Write, Copy, Execute
milan.domazet Write, Copy, Execute
boris_majic Write, Copy, Execute

[Manage members](#)

ANALYSES

[Tasks](#) [Data Cruncher](#)

Manage members

1 member **Permissions** ([Learn more](#))

alisonleaf **OWNER** Joined on July 2, 2020 10:04

Invite new members

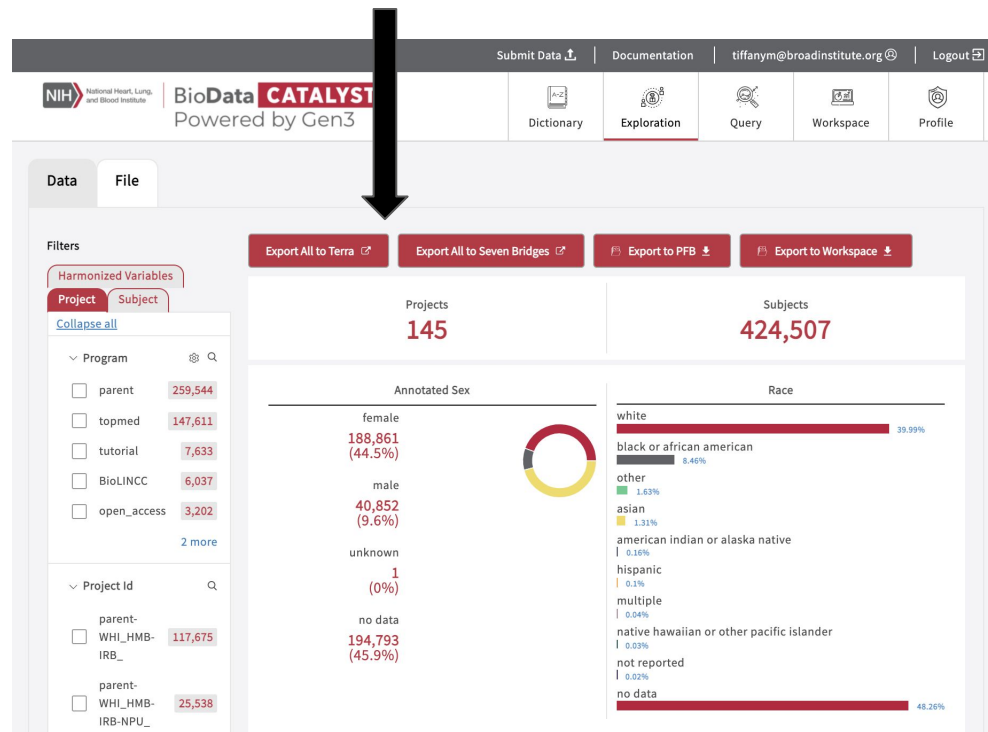
[Write, Copy, Execute](#) [Invite](#)

sara_seepo Sara Seepo

Example from BioData Catalyst Powered by Seven Bridges

Access and analyze 3.4 PB of hosted data on BioData Catalyst

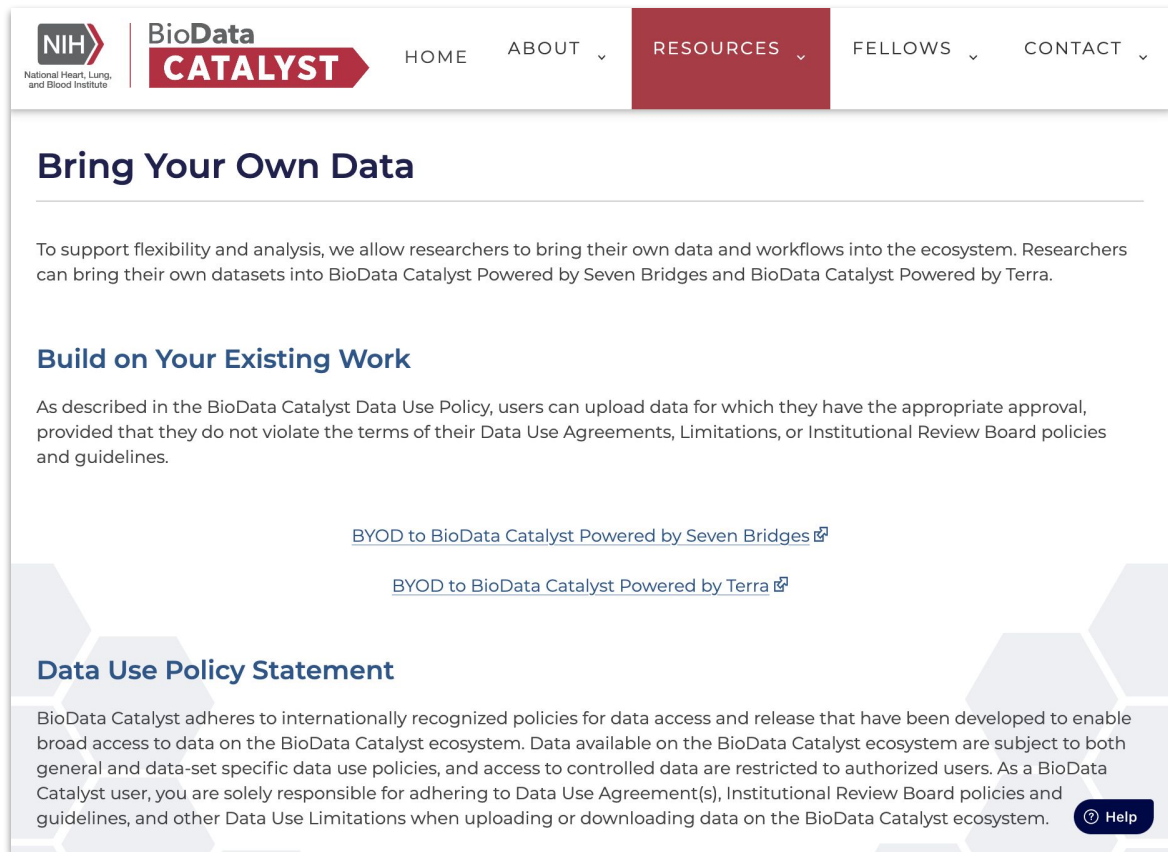
- Access data based on dbGaP permissions across the BioData Catalyst ecosystem.
- Select files and export to workspace **-OR-** form a cohort and work with associated files in workspace.
- Perform analysis on NHLBI's hosted data and avoid moving around large datasets.



Example from BioData Catalyst Powered by Gen3

Bring your own data

- For smaller uploads, drag and drop from computer.
- For larger uploads, Command Line Upload.
- Access private cloud storage bucket from workspace and use as external file repository.



The screenshot shows the BioData Catalyst website interface. At the top is a navigation bar with the NIH logo and 'BioData CATALYST' branding. The 'RESOURCES' menu item is highlighted in a red box. Below the navigation bar, the main heading is 'Bring Your Own Data'. The text explains that researchers can bring their own datasets into the ecosystem. There are two sub-sections: 'Build on Your Existing Work' with a link to the BioData Catalyst Data Use Policy, and 'Data Use Policy Statement' which details the policies for data access and release. A 'Help' button is visible in the bottom right corner.

BioData CATALYST HOME ABOUT RESOURCES FELLOWS CONTACT

Bring Your Own Data

To support flexibility and analysis, we allow researchers to bring their own data and workflows into the ecosystem. Researchers can bring their own datasets into BioData Catalyst Powered by Seven Bridges and BioData Catalyst Powered by Terra.

Build on Your Existing Work

As described in the BioData Catalyst Data Use Policy, users can upload data for which they have the appropriate approval, provided that they do not violate the terms of their Data Use Agreements, Limitations, or Institutional Review Board policies and guidelines.

[BYOD to BioData Catalyst Powered by Seven Bridges](#)

[BYOD to BioData Catalyst Powered by Terra](#)

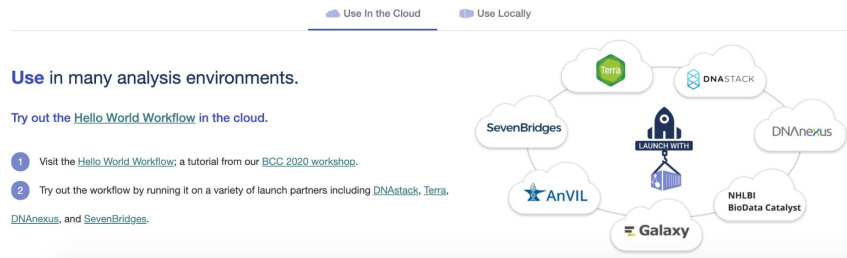
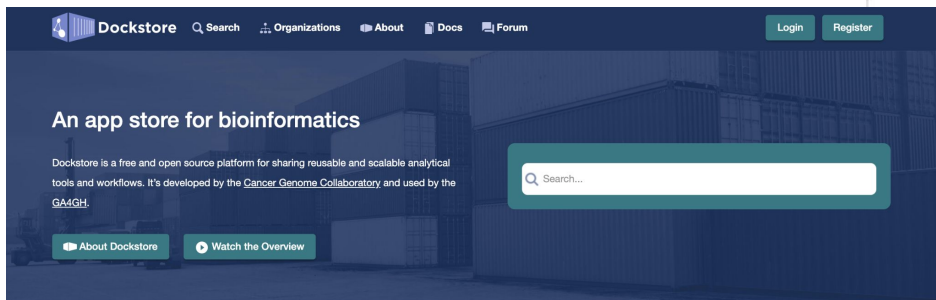
Data Use Policy Statement

BioData Catalyst adheres to internationally recognized policies for data access and release that have been developed to enable broad access to data on the BioData Catalyst ecosystem. Data available on the BioData Catalyst ecosystem are subject to both general and data-set specific data use policies, and access to controlled data are restricted to authorized users. As a BioData Catalyst user, you are solely responsible for adhering to Data Use Agreement(s), Institutional Review Board policies and guidelines, and other Data Use Limitations when uploading or downloading data on the BioData Catalyst ecosystem.

Help

Reproducible analyses using dockerized workflows

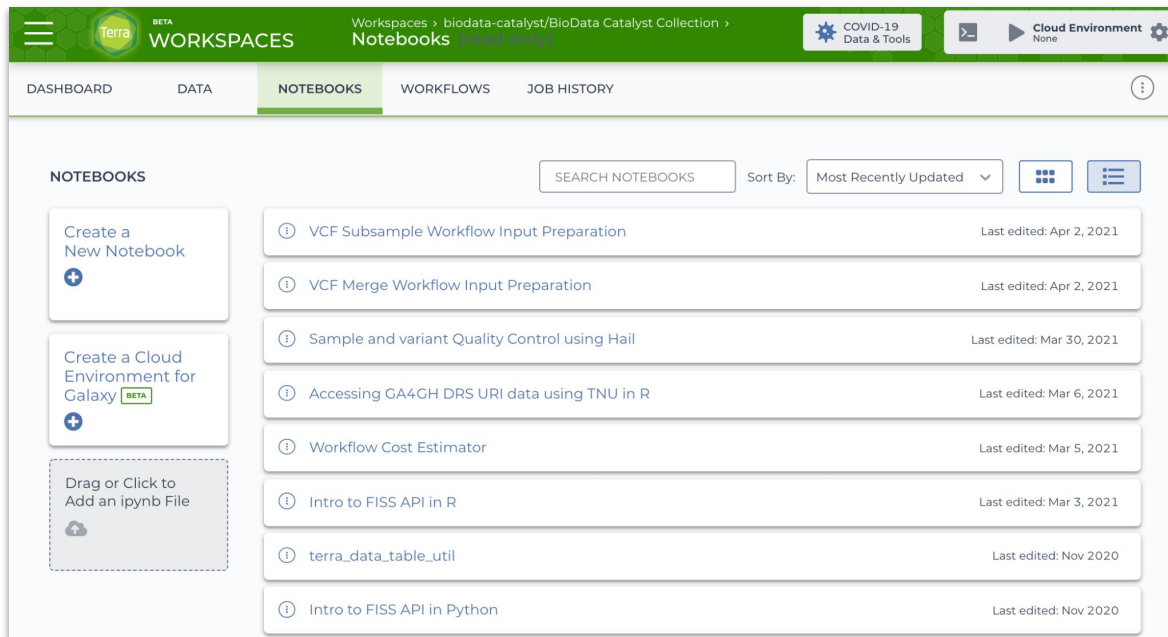
- Find and export publicly available workflows from Dockstore.
- Docker containers allow you to package your required software/tools and easily run your workflow on any computing platform.
- With your workflow and docker, launch large jobs without waiting in a queue.



Example from Dockstore

Interactive analysis

- Spin up an interactive environment to analyze outputs of your workflows or any other data in the cloud.
- Built-in applications include Jupyter Notebooks and RStudio.



Example from BioData Catalyst Powered by Terra

Up next

- Differentiators between the Terra and Seven Bridges platforms
- How to get free Pilot Credits
- Getting started guides

Terra

Tiffany Miller

Navigation in Terra



Notes on navigation

URL: <https://terra.biodatacatalyst.nhlbi.nih.gov/> or app.terra.bio

ERA Commons ID is not required to register (link in your profile page)

The screenshot shows the NHLBI BioData Catalyst homepage. At the top, there is a header with the NIH logo, the text "BioData CATALYST", and "Powered by Terra". The main heading is "Welcome to NHLBI BioData Catalyst". Below this, a paragraph states: "NHLBI BioData Catalyst is a project powered by Terra for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**." Another paragraph says: "Find how-to's, documentation, video tutorials, and discussion forums". There are three main action buttons: "View Workspaces" (describing connecting data to analysis tools), "View Examples" (describing a gallery of showcase Workspaces), and "Browse Data" (describing access to a rich ecosystem of data portals). Each button has a right-pointing arrow icon. The background features hexagonal patterns and images of a cell, a test tube, and two scientists working.

NIH National Heart, Lung, and Blood Institute | BioData CATALYST | Powered by Terra

Welcome to NHLBI BioData Catalyst

NHLBI BioData Catalyst is a 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

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.

Privacy Policy | Freedom of Information Act (FOIA) | Accessibility | U.S. Department of Health & Human Services | National Institutes of Health | USA.gov | National Heart, Lung, and Blood Institute

The screenshot shows the Terra website homepage. At the top, there is a header with the Terra logo and "Powered by Terra". The main heading is "Welcome to Terra". Below this, a paragraph states: "Terra is a cloud-native platform for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**." Another paragraph says: "Find how-to's, documentation, video tutorials, and discussion forums". A third paragraph says: "Learn more about the Terra platform and our co-branded sites". There are three main action buttons: "View Workspaces" (describing connecting data to analysis tools), "View Examples" (describing a gallery of showcase Workspaces), and "Browse Data" (describing access to a rich ecosystem of data portals). Each button has a right-pointing arrow icon. The background features hexagonal patterns and images of a cell, a test tube, and two scientists working.

Terra | Powered by Terra

Welcome to Terra

Terra is a cloud-native platform for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**.

Find how-to's, documentation, video tutorials, and discussion forums

Learn more about the Terra platform and our co-branded sites

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.

Data & Tools for COVID-19/SARS CoV2 analysis

[See this article](#) for a summary of available resources.

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

Notes on navigation

New Featured workspaces Organization

<https://terra.biodatacatalyst.nhlbi.nih.gov/#library/showcase>

Check out the BioData Catalyst Collection workspace

The screenshot displays the BioData Catalyst Library interface. At the top, the header includes the NIH logo, the text "BioData CATALYST Powered by Terra", and a "BETA LIBRARY" badge. Below the header, there are three tabs: "DATASETS", "FEATURED WORKSPACES" (which is active), and "CODE & WORKFLOWS".

The "Featured workspaces" section shows a search bar with the text "Search Name or Description" and a "Sort by" dropdown menu set to "most recent". A "clear" link is also present. On the left side of this section, there is a sidebar with a list of categories and their counts: "Getting Started" (3), "Analysis Tools" (expanded, showing "WDLs" (37), "Jupyter Notebooks" (23), "Hail" (2), "Bioconductor" (1), "GATK" (19), and "Spark" (3)), "Experimental Strategy" (1), "Data Generation Technology" (1), "Scientific Domain" (1), "Datasets" (1), "Utilities" (1), and "Projects" (1).

The main content area displays three featured workspaces, each with a star icon, a title, a date, and a description:

- GATK4-RNA-Germline-VariantCalling** (Jul 14, 2021): A workspace containing GATK Best Practices for Germline Variant Calling in RNAseq. It includes a WDL workflow that calls germline short variants (SNPs/Indels) from RNAseq data using GATK v4.1 and related tools. A detailed description of the workflows is available in [Gat's Best Practices Document] (<https://software.broadinstitute.org/gatk/best-practices/workflow?id=11164>). Scroll down for details on the workflow, including input and output descriptions and requirements, estimated run times and costs.
- Peat-Demo** (May 18, 2021): A workspace demonstrating how to use [Peat (external link)] (<https://broad.io/peat>) to save overhead by grouping jobs into fewer WDL scatter branches. To compare scatter with and without Peat, this workspace has two simple demo workflows using WDL scatter, one with and one without using Peat.
- infercnv** (Apr 6, 2021): A workspace containing a fully reproducible example workflow for inferring copy number from single-cell RNA sequencing data. Complete documentation for InferCNV is available on the [here] (<https://github.com/broadinstitute/inferCNV/wiki>).

Below these, two more workspaces are partially visible:

- CRDC-Dynamic-Queries-for-NIH-Genomic-Data-Commons-Projects** (Mar 29, 2021): A workspace showing how to take a query result from the [NCI Genomic Data Commons] (<https://portal.gdc.cancer.gov/>) (GDC) data portal and use it as the input to a workflow (or Notebook) in FireCloud.

Terra differentiators

Workflow Language	Workflow Description Language (WDL)
Cloud Provider	Google Cloud Platform, Azure (<i>coming</i>)
Applications	<p>Preloaded applications and options to bring-your-own through a user-friendly interface.</p> <p>Galaxy (including full community toolshed), IGV, Seqr</p>
Interactive Analysis Features	<p>Highly customizable machines with persistent disks set up to save your work</p> <p>Bioconductor, Hail, GATK and other popular bioinformatics tools preloaded. Available in "best practices" workspaces maintained/approved by the tool developers</p>



About OpenWDL

The **Workflow Description Language** (WDL) is a way to specify data processing workflows with a human-readable and -writeable syntax. WDL makes it straightforward to define analysis tasks, chain them together in workflows, and parallelize their execution. The language makes common patterns simple to express, while also admitting uncommon or complicated behavior; and strives to achieve portability not only across execution platforms, but also different types of users. Whether one is an analyst, a programmer, an operator of a production system, or any other sort of user, WDL should be accessible and understandable.

WDL was originally developed for genome analysis pipelines by the Broad Institute. As its community grew, both end users as well as other organizations using WDL for their own software, it became clear that there was a need to allow WDL to become a true community driven standard. The **OpenWDL** community has thus been formed to steward the WDL language specification and advocate its adoption.

OpenWDL

*Community driven open-development
workflow language*

About

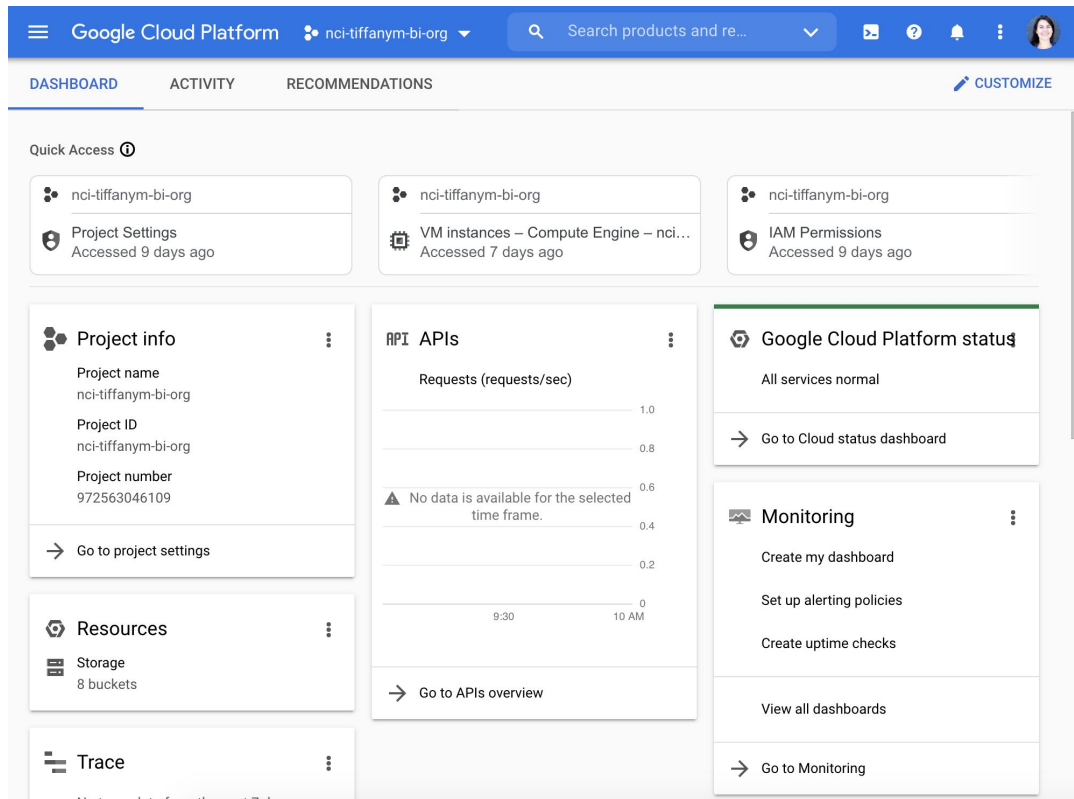
How to participate

More information

Core group

Cloud Providers

- Google Cloud Platform
- Azure (coming)



Applications

- Bring your own Docker image (built off the Terra base image) or initialization script
- Galaxy
<https://galaxyproject.org/>
- IGV
<https://software.broadinstitute.org/software/igv/home>
- Seqr
<https://seqr.broadinstitute.org/>

Interactive Analysis Tutorials

Best practice
workspaces
made available
by tool
developers

Check out the
BioData Catalyst
Collection
workspace

The screenshot displays the Terra WORKSPACES interface. The top navigation bar is green and includes the Terra logo, a 'BETA' badge, and the text 'WORKSPACES'. Below this, a breadcrumb trail shows 'Workspaces > biodata-catalyst/BioData Catalyst Collection > Notebooks (read only)'. On the right of the header, there are buttons for 'COVID-19 Data & Tools' and 'Cloud Environment' (set to 'None'). A secondary navigation bar contains tabs for 'DASHBOARD', 'DATA', 'NOTEBOOKS' (which is active), 'WORKFLOWS', and 'JOB HISTORY'. The main content area is titled 'NOTEBOOKS' and features a search bar, a 'Sort By' dropdown set to 'Most Recently Updated', and two view toggles (grid and list). On the left side of the notebook list, there are three interactive buttons: 'Create a New Notebook', 'Create a Cloud Environment for Galaxy' (with a 'BETA' badge), and a dashed box labeled 'Drag or Click to Add an ipynb File'. The central part of the interface is a list of 12 notebooks, each with a circular icon, a title, a status indicator, and a 'Last edited' timestamp. The notebooks are: 'VCF Subsample Workflow Input Preparation' (Apr 2, 2021), 'VCF Merge Workflow Input Preparation' (Apr 2, 2021), 'Sample and variant Quality Control using Hail' (Mar 30, 2021), 'Accessing GA4GH DRS URI data using TNU in R' (Mar 6, 2021), 'Workflow Cost Estimator' (Mar 5, 2021), 'Intro to FISS API in R' (Mar 3, 2021), 'terra_data_table_util' (Nov 2020), 'Intro to FISS API in Python' (Nov 2020), 'Accessing GA4GH DRS URI data using TNU CLI' (Nov 2020), 'Accessing GA4GH DRS URI data using TNU in Python' (Nov 2020), 'Bring Your Own Data Tutorial' (Oct 2020), and 'Prepare Gen3 data for input into the Integrative Genomics Viewer (IGV) in Terra' (Jul 2020).

Notebook Title	Last edited
VCF Subsample Workflow Input Preparation	Apr 2, 2021
VCF Merge Workflow Input Preparation	Apr 2, 2021
Sample and variant Quality Control using Hail	Mar 30, 2021
Accessing GA4GH DRS URI data using TNU in R	Mar 6, 2021
Workflow Cost Estimator	Mar 5, 2021
Intro to FISS API in R	Mar 3, 2021
terra_data_table_util	Nov 2020
Intro to FISS API in Python	Nov 2020
Accessing GA4GH DRS URI data using TNU CLI	Nov 2020
Accessing GA4GH DRS URI data using TNU in Python	Nov 2020
Bring Your Own Data Tutorial	Oct 2020
Prepare Gen3 data for input into the Integrative Genomics Viewer (IGV) in Terra	Jul 2020

Seven Bridges

Dave Roberson and Tony Patelunas

Live Demo

Seven Bridges differentiators

Workflow Language	Common Workflow Language (CWL)
Cloud Provider	Amazon Web Services, Google Cloud Platform; capability to connect private cloud bucket directly to platform
Applications	<p>Annotation Explorer: Query annotations for all SNVs and dbSNP INDELS (+8 billion variants and ~700 annotations)</p> <p>Genomic Data Overview: Query TOPMed variants and see aggregated results. Compliant with Genomic Data Sharing Agreements.</p>
Interactive Analysis Features	<p>SAS</p> <p>Commonly used libraries like Bioconductor are preloaded on the Docker image for spinning up the RStudio or Jupyterlab environments.</p>

Requesting Pilot Credits

Paul Kerr from BDC3

What are Cloud Credits?

Users are not charged for the storage of hosted datasets; however, if hosted data is used in analyses, users incur costs for computation and storage of derived results.

BioData Catalyst users who upload/import their own data to the system incur storage costs for these uploaded files as well.

Web resource: [Cloud Costs and Credits](#)

Try out both platforms with Pilot Credits

If you don't already have CWL tools or WDL tools and are flexible about which BioData Catalyst workspace to use, **we recommend trying both** to make an informed decision about which platform is the best fit for you.

BioData Catalyst users may request one of the following: *

\$500 in initial pilot cloud credits to begin a project or explore the ecosystem

Select your preferred analysis platform * (or choose to explore both)

✓ Select One

\$500 on Seven Bridges

\$500 on Terra

\$250 each on both Seven Bridges and Terra

Cloud Credits Workflow

1

Sign up for the community

Sign up at
biodatacatalyst.nhlbi.nih.gov/contact/ecosystem

2

Sign up for a workspace

Seven Bridges and/or
Terra

3

Apply for Pilot Credits

Fill out the [Cloud Credits Request form](#).

Use all credits on a single platform, or split.

4

Apply for additional credits or pay yourself

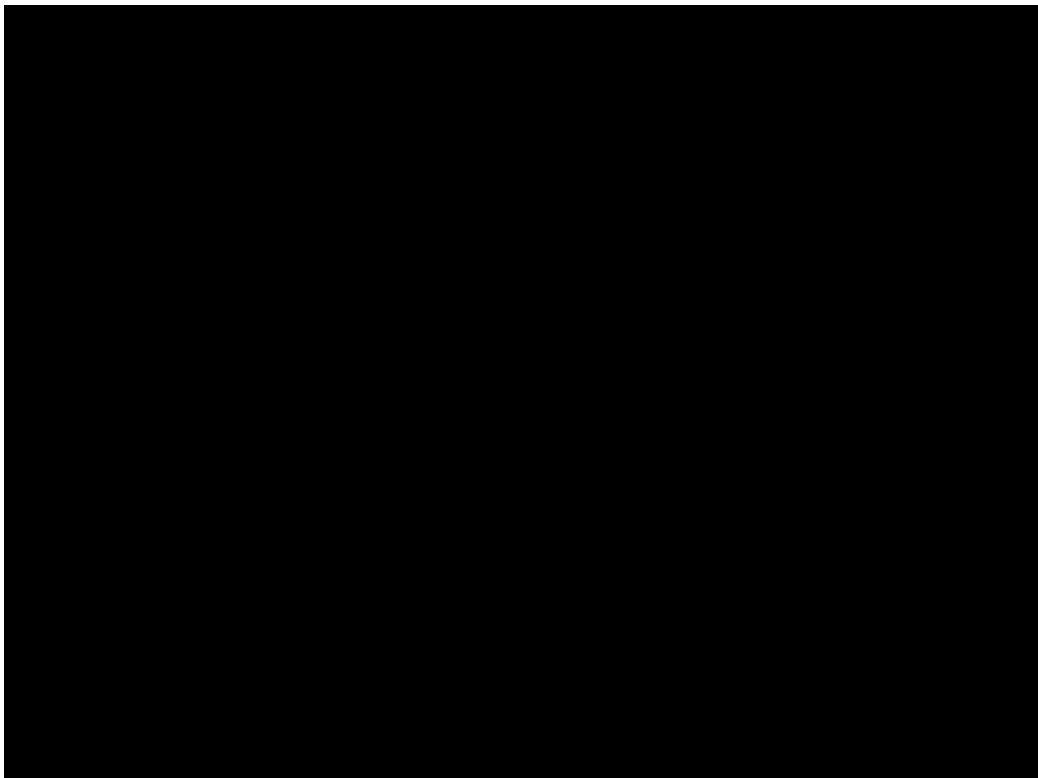
Cover costs after pilot funding has been exceeded.

Potential Exception: Research in the heart, lung, blood, and sleep fields

Web Form

After signing up for a workspace, fill out the **Cloud Credits request form** for free cloud credits:

<https://biodatacatalyst.nhlbi.nih.gov/resources/cloud-credits/>



What are my next steps for Pilot Credits?

- Follow one or both of the platform **Getting Started Guides**
 - [Terra Getting Started](#)
 - [Seven Bridges Getting Started](#)
- Join our next **Community Hours** on Wednesday, September 15th
- Help Desk
<https://biodatacatalyst.nhlbi.nih.gov/contact>

Discussion and Questions

We encourage you to submit unanswered questions to our [help desk](https://biodatacatalyst.nhlbi.nih.gov/contact/):

<https://biodatacatalyst.nhlbi.nih.gov/contact/>

Join the ecosystem

Join the NHLBI BioData
Catalyst Community

<https://biodatacatalyst.nhlbi.nih.gov/contact/ecosystem>

Thank you!

The next Community Hours is

Wednesday, September 15 at 1 PM EDT

(<https://bit.ly/3yeQddp>)

Topic: Cloud Credits and Costs