

Onboarding to NHLBI BioData Catalyst powered by Seven Bridges

The tutorial below aims to help you create an account on the BioData Catalyst Platform powered by Seven Bridges and learn the basics of creating a workspace (project), running an analysis, and searching the hosted data. For more information, please refer to the platform documentation at <https://sb-biodatacatalyst.readme.io/docs>.

Accessing hosted TOPMed datasets on BioData Catalyst

BioData Catalyst hosts a number of controlled datasets from the Trans-omics for Precision Medicine (TOPMed) initiative. These datasets are stored in Amazon Web Services (AWS) and Google Cloud storage buckets operated by NHLBI such that the BioData Catalyst ecosystem enables users to access the same copy of the data. Access to these hosted datasets is controlled programmatically by services within the *BioData Catalyst* ecosystem for user authentication and authorization. **Users log into BioData Catalyst platforms using their eRA Commons credentials and authentication is performed by iTrust.**

The BioData Catalyst ecosystem manages user access to the hosted controlled data using data access approval from the NIH Database of Genotypes and Phenotypes (dbGaP). Therefore, users who want to access one or more of the hosted controlled studies on the ecosystem must be approved for access to that study in dbGaP. Principal Investigators who have approved Data Access Requests (DARs) on dbGaP for the BioData Catalyst datasets will be able to programmatically access those data on the platforms and services within the BioData Catalyst ecosystem.

Principal Investigators with an approved DAR can enable their lab staff to access the hosted datasets on the BioData Catalyst ecosystem by giving the lab staff “designated downloader status” on dbGaP. These individuals must:

- Have an eRA commons account or an NIH username and password. [Please see these instructions.](#)
- Log in to dbGaP at least once

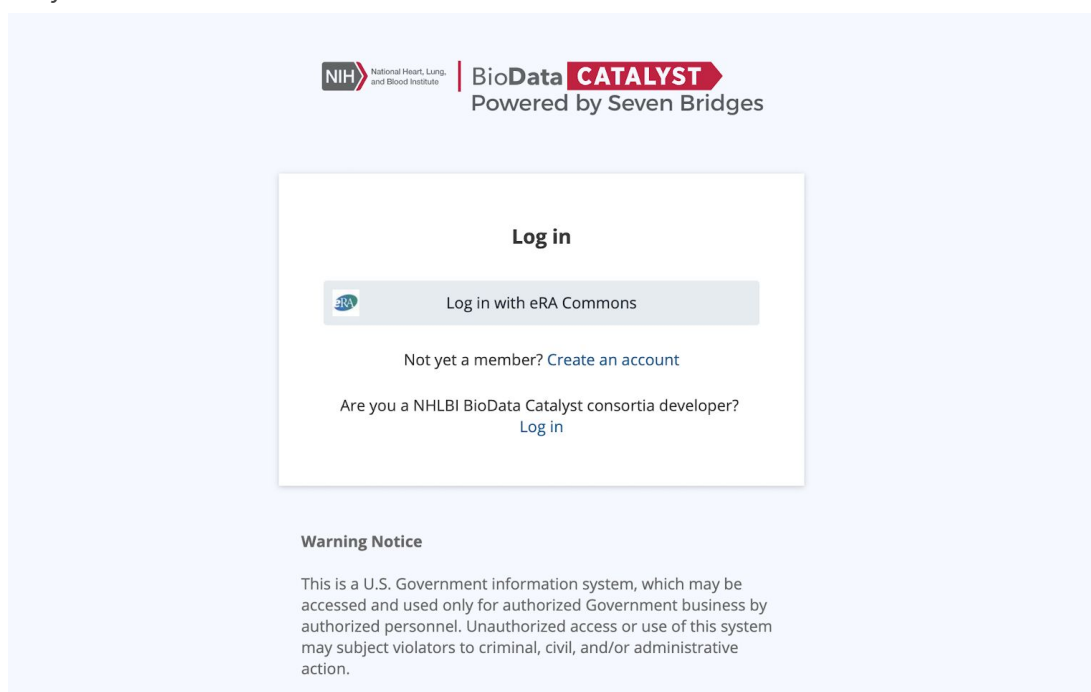
To give lab staff downloader status, please refer to [these instructions](#).

Please note that having other researchers listed on your dbGaP DAR application as internal and external collaborators will not result in these individuals getting access to hosted dataset on BioData Catalyst. PI's will need to add internal collaborators from their dbGaP application to the list of designated downloaders as described above. In addition, external collaborators will need to go through this same process for those at their own institution.

Researchers also have the option to bring their own datasets to the BioData Catalyst platforms. 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. To learn more about how to bring your own data to BioData Catalyst powered by Seven Bridges, please refer to the [Documentation](#).


Account Registration on BioData Catalyst powered by Seven Bridges

To create an account, please visit the platform login page at <https://platform.sb.biodatacatalyst.nhlbi.nih.gov>. All researchers will use eRA Commons credentials to login to the system. Please select "Create an account."



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

Log in

 Log in with eRA Commons


Not yet a member? [Create an account](#)

Are you a NHLBI BioData Catalyst consortia developer?
[Log in](#)


Warning Notice

This is a U.S. Government information system, which may be accessed and used only for authorized Government business by authorized personnel. Unauthorized access or use of this system may subject violators to criminal, civil, and/or administrative action.

Then select "Continue with eRA Commons."

 National Heart, Lung, and Blood Institute | **BioData CATALYST**
Powered by Seven Bridges

Create an Account

 Continue with eRA Commons


Already a member? [Log in](#)

Are you a NHLBI BioData Catalyst consortia developer?
[Register](#)

Warning Notice

This is a U.S. Government information system, which may be accessed and used only for authorized Government business by authorized personnel. Unauthorized access or use of this system may subject violators to criminal, civil, and/or administrative action.

The platform will redirect you to iTrust where you can enter in your eRA commons credentials.




User Name:

Password: [Change Password](#)

[Log in](#)

OR



Insert your PIV card into your smart card reader before attempting to login.

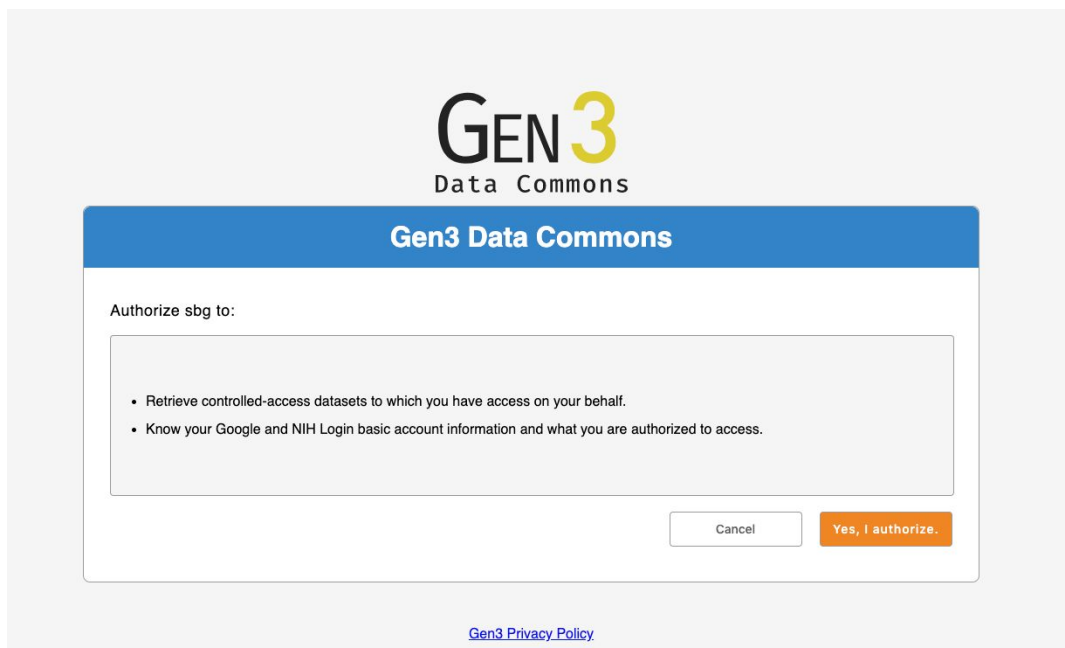
For more information visit <http://smartcard.nih.gov>.

[Log in](#)

Warning Notice

- This warning banner provides privacy and security notices consistent with applicable federal laws, directives, and other federal guidance for accessing this Government system, which includes (1) this computer network, (2) all computers connected to this network, and (3) all devices and storage media attached to this network or to a computer on this network.
- This system is provided for Government-authorized use only.
- Unauthorized or improper use of this system is prohibited and may result in disciplinary action and/or civil and criminal penalties.
- Personal use of social media and networking sites on this system is limited as to not interfere with official work duties and is subject to monitoring.
- By using this system, you understand and consent to the following:
 - The Government may monitor, record, and audit your system usage, including usage of personal devices and email systems for official duties or to conduct HHS business. Therefore, you have no reasonable expectation of privacy regarding any communication or data transiting or stored on this system. At any time, and for any lawful Government purpose, the government may monitor, intercept, and search and seize any communication or data transiting or stored on this system.
 - Any communication or data transiting or stored on this system may be disclosed or used for any lawful Government purpose.

After you submit your eRA Commons credentials, the system will redirect you to the BioData Catalyst Gen3 service which manages user authentication and authorization. Please select “Yes, I authorize.”



The image shows a web interface for 'Gen3 Data Commons'. At the top, the logo 'GEN3 Data Commons' is displayed, with 'GEN3' in a large, bold, black font and 'Data Commons' in a smaller, black font below it. Below the logo is a blue header bar with the text 'Gen3 Data Commons' in white. The main content area is white and contains the text 'Authorize sbg to:' followed by a list of permissions: 'Retrieve controlled-access datasets to which you have access on your behalf.' and 'Know your Google and NIH Login basic account information and what you are authorized to access.' At the bottom right of the main content area are two buttons: 'Cancel' and 'Yes, I authorize.' Below the main content area is a link to the 'Gen3 Privacy Policy'.

GEN3
Data Commons

Gen3 Data Commons

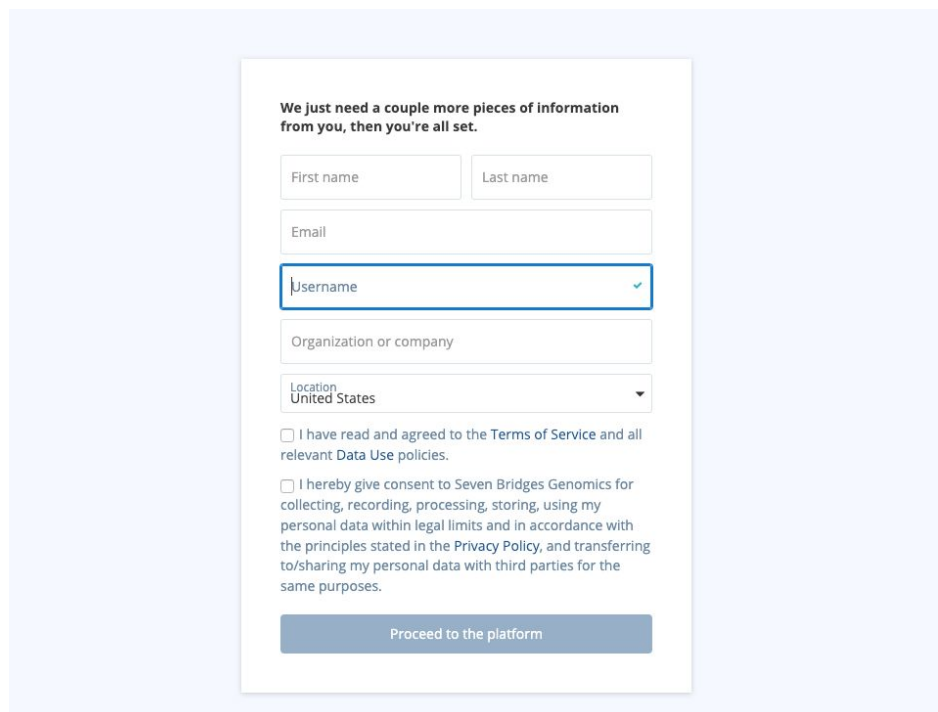
Authorize sbg to:

- Retrieve controlled-access datasets to which you have access on your behalf.
- Know your Google and NIH Login basic account information and what you are authorized to access.

Cancel Yes, I authorize.

[Gen3 Privacy Policy](#)

Then you will be directed to the account registration page. For platform username, we suggest using “firstnamelastname” or “firstname.lastname” since this will make it easier for other platform members to search for you in the collaboration feature.



The image shows a web form for account registration. The form is titled 'We just need a couple more pieces of information from you, then you're all set.' and contains several input fields: 'First name', 'Last name', 'Email', 'Username' (with a dropdown arrow), 'Organization or company', and 'Location' (with a dropdown arrow). Below the input fields are two checkboxes for terms of service and consent. At the bottom of the form is a button labeled 'Proceed to the platform'.

We just need a couple more pieces of information from you, then you're all set.

First name Last name

Email

Username

Organization or company

Location
United States

☐ I have read and agreed to the [Terms of Service](#) and all relevant [Data Use](#) policies.

☐ I hereby give consent to Seven Bridges Genomics for collecting, recording, processing, storing, using my personal data within legal limits and in accordance with the principles stated in the [Privacy Policy](#), and transferring to/sharing my personal data with third parties for the same purposes.

Proceed to the platform

After clicking “Proceed to the platform,” the system will verify your email. New users have \$500 of pilot credits that they receive upon account creation. This enables users to get started with testing out the platform and running some analyses. If you already have a project planned that you would like to perform on the platform, please email the Seven Bridges Program Manager for BioData Catalyst, Alison Leaf (alison.leaf@sevenbridges.com).

The Seven Bridges team is eager to help you get the most out of using the platform. We would like to hear about your analysis plans so that we can support your research to the best of our abilities. This includes but is not limited to: an initial platform demo over web call; recommendations on which of the hosted bioinformatics tools/workflows to use; tips for accessing the hosted datasets or uploading your own data; and a tailored in-depth training over web call.

Controlled data questionnaire

For users who are approved to access controlled hosted data on BioData Catalyst, like the TOPMed studies, you will be prompted to take a “Controlled data questionnaire” upon first login. Please refer to the explanations below for filling out the questionnaire. If you have any questions, don’t hesitate to reach out to the Seven Bridges Program Manager Alison Leaf (alison.leaf@sevenbridges.com).

Controlled Data Questionnaire

The DataSTAGE powered by Seven Bridges allows you to view and analyze controlled-access data, according to the access granted to you. Before you can begin, you will need to answer a few questions about controlled-access data usage. You will only need to complete this questionnaire once. [Learn more](#)

I'm studying the influence of lifestyle changes on disease progression, but some of the variables I'm interested in are not captured. Am I allowed to try to identify participants and ask them follow-up questions?

☐ Yes, absolutely. The participants may decide if they want to ignore this request.

☐ No, absolutely not. All data access agreements specifically prohibit attempts to re-identify or re-contact research participants.

☐ Maybe. It depends on the follow-up questions that I wish to ask.

As a Certified User I can:

☐ Browse open- and controlled-access data from the Data Browser

☐ Create a controlled project

☐ Add other Certified Users to a controlled project

☐ Access raw files according to my data access approvals

☐ All of the above

A controlled project:

☐ A. Serves as a workspace for analyzing controlled-access data

☐ B. Allows all members to access all raw data

☐ C. Allows all members to access all derived data

☐ A, B and C

☐ A and C

☐ B and C

It is my responsibility to ensure that all members of a project have appropriate data access approvals for the study being conducted.

☐ True

☐ False

If I download a controlled-access file or a file derived from a controlled-access file, I may share it with any other researcher at my institution.

☐ True

☐ False

Check answers

Projects: Projects are workspaces that serve as containers for files, bioinformatics tools and workflows, and analyses. Users can create projects and add collaborators to those projects with specific permissions for what those collaborators are able to see and do within the project.

Raw data: This refers to the hosted datasets on the platform. Access to these raw files is controlled programmatically by the platform such that users are only able to access files they have approval for. The hosted data is available for search via the Data Browser feature. The Data Browser feature has both

open and controlled datasets available for search. Users can select files from the hosted datasets to add to projects and then use them in analyses. The platform only lets users add files to projects if they are approved to access those files. In addition, once a raw file has been added to a project, users are only able to use those raw files in an analysis if they are approved for access.

Derived data: Derived data are the output files from running a bioinformatics tool or workflow. These files are stored in the projects where the analyses were run. All members of a project are able to access the derived data files within a project. The platform does not control a user's ability to access derived data apart from who is a member of a given project.

Certified user: A user on the platform who is approved to access hosted controlled data.

Controlled project: When users create new projects, they have the option to set the project as a "Controlled project." The purpose of Controlled projects is to help users protect Controlled data by restricting access to users who have the necessary approvals to work with that data. If users want to work with hosted controlled data from the Data Browser feature, then the user must add those controlled data files to a Controlled project. In addition, the project owner (and other admins) can only add new members to the project if those members also have access to Controlled data on the platform. Members of Controlled projects can see a list of all the files in that project, however they are only able to access raw files (use the file as an input in a tool or workflow) if they have appropriate access approval. All project members can access derived data.

Check data access permissions on BioData Catalyst

BioData Catalyst programmatically controls user permissions on the hosted controlled datasets like the TOPMed studies. To see which of the hosted datasets you are approved to access on BioData Catalyst, click on your username in the top right-hand corner of the platform. Select the tab for Data Access and verify that there are green check marks next to the datasets that you expect to have access to. The datasets are referred to using their dbGaP phs numbers and consent codes. If you don't see what you expect, please refer to the BioData Catalyst web pages on troubleshooting data access.

The example user shown below has access to all of the hosted TOPMed studies, however please note that your data access should match up with what you have access to in dbGaP.

NIH | BioData CATALYST
Powered by Seven Bridges

Account Settings | **Dataset Access**

NHLBI BIODATA CATALYST POWERED BY SEVEN BRIDGES

You have access to the following datasets:

BHEAVNER

Controlled Datasets

- ADMIN controlled data
- PHS000920.C2 controlled data
- PHS000921.C2 controlled data
- PHS000946.C1 controlled data
- PHS000951.C1 controlled data
- PHS000951.C2 controlled data
- PHS000954.C1 controlled data
- PHS000956.C2 controlled data
- PHS000964.C1 controlled data
- PHS000964.C2 controlled data
- PHS000964.C3 controlled data

Create a project

Projects are workspaces that serve as the core building blocks of BioData Catalyst powered by Seven Bridges. Each project corresponds to a distinct scientific investigation and serves as a container for data, analysis workflows, and results. Multiple analyses can be carried out within a project.

Projects are secure and private. The project creator has the option to add collaborators to the project as project members. Each project has at least one administrator, who controls the project members' permissions to execute analyses. You can be a member of multiple projects each with different teams of researchers.

Let's go through the steps of creating a project to learn more about the options of how they can be configured. On the main dashboard, select "Create a project."

The screenshot shows the BioData Catalyst dashboard. The top navigation bar includes the NCI logo, 'BioData CATALYST Powered by Seven Bridges', and several dropdown menus: 'Projects', 'Data', 'Public Gallery', 'Public projects', 'Developer', a notification bell, and a user profile 'danielventre'. The main content area is divided into two panels. The left panel, titled 'PROJECTS', contains the text 'All analyses on the Platform take place inside projects.' and a prominent blue button labeled '+ Create a project'. Below the button is a link 'or learn more about projects.' The right panel, titled 'ANALYSES', features a search bar and two tabs: 'Tasks' (which is selected) and 'Data Cruncher'. Below the tabs, it says 'Your executions will appear here. Before you start, [learn more about them.](#)' The footer contains links for 'Terms', 'Privacy', 'Data Use', and 'Copyright', along with a help icon (question mark) in a blue circle.

The screenshot shows the 'Create a project' dialog in the Seven Bridges BioData CATALYST interface. The dialog is overlaid on a blurred background of the main application. Five red circles with white numbers 1 through 5 highlight specific fields:

1. Name (text input with 'test project')
2. Billing Group (dropdown menu with 'Pilot Funds (danielventre)')
3. Location (dropdown menu with 'AWS (us-east-1)')
4. Spot Instances (checkbox, currently checked)
5. This project will contain (checkbox, currently unchecked) followed by a 'CONTROLLED' label and a 'Data' link

The dialog also includes a 'Project URL' field, 'Execution settings' for 'Memoization' (OFF), and 'Cancel' and 'Create' buttons at the bottom.

1. Name the project

Following along with the red step indicators in the screenshot above, we begin by picking a name for your project. Your project will be assigned a short name based on the name that you give it, which is used as an ID to refer to the project when using the API.

2. Billing group

Billing groups are used to track the costs associated with cloud storage and computation on the platform. Each project must be assigned a billing group. New users are automatically assigned a “Pilot Funds” billing group upon account registration with \$500 in cloud credits to get started with analyses. The Seven Bridges Support Team will reach out when a user is close to using up their Pilot Funds and offer to create a new billing group for further work. If you already have a project planned, please reach out to the Seven Bridges Program Manager Alison Leaf (alison.leaf@sevenbridges.com) for assistance.

For this **example project**, select the Pilot Funds billing group.

3. Location

In an effort to enable users to compute on data where it lives, the platform offers the choice to perform computation on two different cloud locations (cloud provider and region): AWS us-east-1 and Google us-west-1. When users create a new project, they will select one of these two cloud locations as the location for the project. All computation within the project will take place on this cloud location. In addition, any resulting files from analyses will be stored on this cloud location. Analyses can use input files from any cloud location. However, if the input files are stored on a different cloud location than the one set for the project, data egress will occur. Therefore, it is typically most efficient to select a cloud location for the project based on where a majority of the input files are stored.

For users who will analyze the hosted TOPMed CRAM or VCF files, either cloud location can be used because the datasets are duplicated with a copy stored on AWS us-east-1 and another copy on Google us-west-1.

If users have their own private data in a cloud bucket (AWS or Google), it can be connected to the platform. For these analyses, users will likely want to select the cloud location based on the location of the private cloud storage bucket.

For this **example project**, select AWS-us-east-1 as the cloud location.

4. Execution Settings

The first selection in the Execution Settings is whether to use a discounted type of computation instance on AWS or Google Cloud, which uses the cloud provider's spare capacity. For AWS, the platform supports EC2 Spot instances. With Spot instances, you pay the Spot price that is in effect for the time period your instances are running. Spot instance prices are set by Amazon EC2 and adjust gradually based on long-term trends in supply and demand for Spot instance capacity. Spot Instances are available at up to a 90% discount compared to On-Demand prices. For Google, the platform supports Preemptible instances. They offer the same machine types and options as regular compute instances and last for up to 24 hours. Pricing is fixed so you will always get low cost and financial predictability, without taking the risk of gambling on variable market pricing. Preemptible instances are up to 80% cheaper than regular instances.

Both AWS and Google might terminate these instances at any time if they require access to those resources due to high demand. The job(s) running on the instance at the time of termination will be interrupted and have to be run again from the beginning. The jobs will be restarted on an equivalent regular On-Demand instance to minimize time wasted in completing your analysis. Restarting jobs on another instance will inevitably prolong execution time and add to the cost. **Therefore, these instances are not recommended for running long, time-critical jobs.**

For this **example project**, turn on AWS Spot instances.

The next selection for Execution Settings is whether to use memoization when running analyses. Memoization allows researchers and bioinformaticians to restart from a point of failure in a workflow by enabling the reuse of existing outputs. This memoization feature is currently in beta stage and you can read more about it on the [Seven Bridges blog](#).

For this **example project**, leave memoization in the default “off” setting.

5. Controlled Projects

Projects can be set as either “Open” or “Controlled.”

Open Data Projects are designed to host both **Open Data** and **your private data**. Open Data is available to all the users on the Platform upon sign up. Open Data contains data which is not unique to an individual, such as de-identified clinical data, gene expression data, copy number alterations in regions of the genome, epigenetic data, and summaries of data compiled across individuals. Note that you cannot copy Controlled Data inside an Open Data Project.

Controlled projects help users protect hosted **Controlled Data** (like the TOPMed studies) by restricting access to other users who also have approval to work with one or more of the hosted controlled datasets. If users want to work with hosted controlled data from the Data Browser feature, then the user must add those controlled data files to a Controlled project. In addition, the project owner (and other admins) can only add new members to the project if those members also have access to Controlled data on the platform. Members of Controlled projects can see a list of all the files in that project, however they are only able to access raw files (use the file as an input in a tool or workflow) if they have appropriate access approval. All project members can access derived data.

For this **example project**, leave the box unchecked so that the project will be “Open.”

Running analyses

Single executions of bioinformatics tools and workflows on the platform are called “Tasks.” All Tasks are run from within projects. We will set up an example Task using the **FastQC workflow**.

The screenshot displays the Seven Bridges BioData CATALYST interface for a project named "test project". The top navigation bar includes links for Projects, Data, Public Gallery, Public projects, Developer, and a user profile dropdown for "danielventre". Below this, a sub-navigation bar shows Dashboard, Files, Apps, and Tasks. The main content area is split into two panels. The left panel, titled "DESCRIPTION", features a "Welcome to your new project!" message, an explanation of projects as core building blocks, and a list of actions users can take within a project. The right panel, titled "MEMBERS", shows the user "danielventre" as the owner and includes a button to "Invite new members". Below the members section is an "ANALYSES" section with a search bar and tabs for "Tasks" and "Data Cruncher".

Let’s start by adding files to the project. Go to the **Files tab** at the top and then select “Add Files.” This will bring you to the page shown below where you can select from “Public Files” that Seven Bridges makes available or files that are in your other projects. In addition, researchers can upload/import files to the platform using “FTP/HTTP,” the “Data Tools” or the “Volumes” feature which enables connecting private cloud storage to the platform.

For this example Task, please select the first two files shown under Public Reference Files and “Copy to Project” (red arrow).

Add files to "test project"

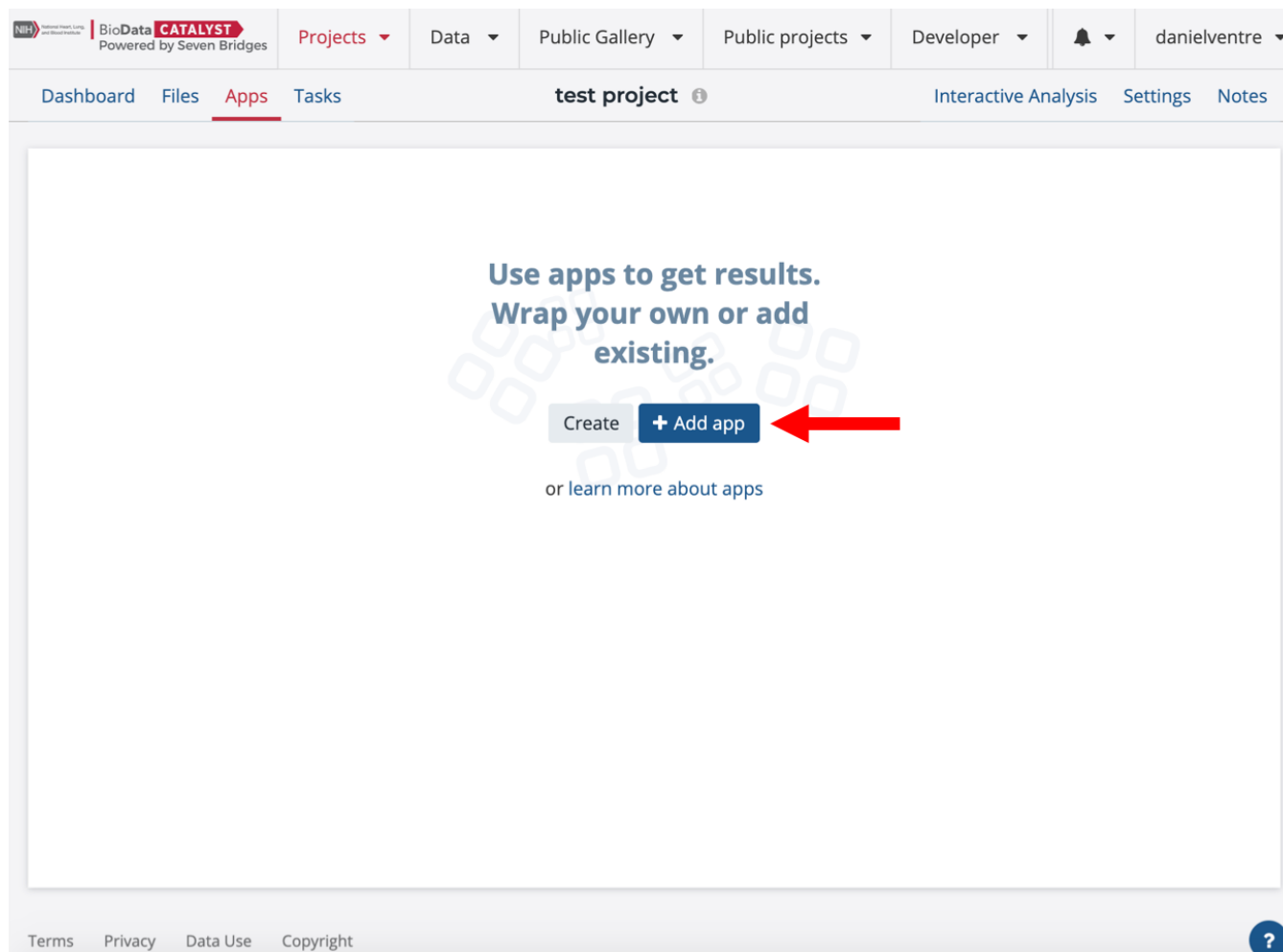
Public Files
Projects
FTP / HTTP
Data Tools
Volumes

Files
Category: All
Type: All
Sample ID: All
Tags: All
+
Copy to Project

	Name	Size	Path	Type
<input checked="" type="checkbox"/>	C835.HCC1143.2.converted.pe_1.fastq <small>WES TUMOR SAMPLE</small>	7.1 GiB	Files	FASTQ
<input checked="" type="checkbox"/>	C835.HCC1143.2.converted.pe_2.fastq <small>WES TUMOR SAMPLE</small>	7.1 GiB	Files	FASTQ
<input type="checkbox"/>	C835.HCC1143_BL.4.converted.pe_1.fastq <small>WES NORMAL SAMPLE</small>	6.2 GiB	Files	FASTQ
<input type="checkbox"/>	C835.HCC1143_BL.4.converted.pe_2.fastq <small>WES NORMAL SAMPLE</small>	6.2 GiB	Files	FASTQ
<input type="checkbox"/>	CCLE-HCC1143-DNA-10_Illumina.converted.pe_1.fastq <small>WGS TUMOR SAMPLE</small>	199.9 GiB	Files	FASTQ
<input type="checkbox"/>	CCLE-HCC1143-DNA-10_Illumina.converted.pe_2.fastq <small>WGS TUMOR SAMPLE</small>	199.9 GiB	Files	FASTQ
<input type="checkbox"/>	CCLE-HCC1143BL-DNA-10_Illumina.converted.pe_1.fastq <small>WGS NORMAL SAMPLE</small>	209.8 GiB	Files	FASTQ
<input type="checkbox"/>	CCLE-HCC1143BL-DNA-10_Illumina.converted.pe_2.fastq <small>WGS NORMAL SAMPLE</small>	209.8 GiB	Files	FASTQ
<input type="checkbox"/>	G20479.HCC1143.2.converted.pe_1.fastq <small>RNA-SEQ</small>	17.4 GiB	Files	FASTQ

Refresh
Showing 1-44 of 44

After adding Files to the project, the user can go to the Apps tab and select “Add app” (red arrow). Apps is the platform term for tools and workflows.



Users have the option to run hundreds of hosted tools and workflows that can be found under the “Public Apps” tab. Tools are denoted with a purple “T” and workflows are denoted with a yellow “W.” All tools and workflows are in the Common Workflow Language (CWL) which is both human and machine-readable and has all the necessary information to run the tool in a reproducible way. Users also have the option to bring their own tools and workflows to the platform using Docker and our SDK. Please reach out to the Seven Bridges team if this is of interest to you.

Search for the **FastQC workflow** in the Public Apps and then select to “Copy” this workflow to your project. **Please note:** when you select the “copy” button shown above by the red arrow, the App URL is displayed, and a second row of buttons appears prompting you to “cancel” or “copy.” Hitting “copy” again then copies your app to your project, and a banner notification appears to confirm.

Add apps to "test project"

Public Apps
Projects
My Apps
Create New App

Category
Toolkit
Reset search

FastQC Analysis

SBGTools 1
The FastQC tool, developed by the Babraham Institute, analyzes sequence data from FASTQ, BAM, or SAM files. It produce...

QUALITY-CONTROL
FASTQ-PROCESSING

Copy
Run

ChIP-seq FastQC

FastQC 0.11.4
FastQC reads a set of sequence files and produces a quality control (QC) report from each one. These reports consist o...

FASTQ-PROCESSING
QUALITY-CONTROL
QUANTIFICATION

Copy
Run

FastQC

FastQC 0.11.4
FastQC reads a set of sequence files and produces a quality control (QC) report from each one. These reports consist o...

FASTQ-PROCESSING
QUALITY-CONTROL
QUANTIFICATION

Copy
Run

SBG FastQC Beautifier

SBGTools
FastQC Beautifier is a simple re-packer for FastQC reports. Adds additional files in order for a nicer HTML render to ...

SBG FastQC Extract

SBGTools
Extract FASTQC BQ extracts the base quality by read position values from multiple files and places them into a single ...

Bismark Analysis

Bismark 0.19.0
Bismark Analysis 0.19.0 is a workflow for analyzing DNA methylation, a type of epigenetic modification, by process...

Please note that for each of the hosted tools and workflows in the Public Apps, users can see a description of the tool/workflows along with helpful information like the required inputs, outputs, and common issues (see below).

The screenshot shows the BioData CATALYST interface. The top navigation bar includes links for Projects, Data, Public Gallery, Public projects, Developer, and a user profile for danielventre. The main header shows 'test project' and links for Interactive Analysis, Settings, and Notes. The 'Apps' tab is selected, displaying the 'FastQC Analysis' app. The app page includes a 'COPY' button, a 'Revision 0' dropdown, and 'Edit', 'Run', and a menu icon. The description states it's a copy of the latest revision by danielventre on Feb. 4, 2020. The 'Description' section explains that FastQC analyzes sequence data from FASTQ, BAM, or SAM files. The 'Required inputs' section specifies FASTQ Reads (ID: FASTQ_reads) and provides instructions on setting the number of threads. The 'Outputs' section mentions a Report ZIP file. On the right, 'Basic Information' lists CWL Version (sbg:draft-2), Contributors (danielventre), Toolkit (SBGTools 1), License (Apache License 2.0), Category (Quality-Control, FASTQ-Processing), App Id (danielventre/te... project/fastqc-analysis), and Links (Homepage, Documentation). 'Workflow steps' shows 'FastQC' as the first step. A help icon (?) is in the bottom right corner.

To set up the draft Task, select “Run” on the App. This will bring you the screen where you will select your input files. When selecting input files for this run, only the FASTQ read files are required. Then click “Run” and your Task will queue up to Run. The completed Task and output file links are shown in the second image.

BioData **CATALYST**
 Powered by Seven Bridges

Projects ▾
 Data ▾
 Public Gallery ▾
 Public projects ▾
 Developer ▾

danielventre ▾

Dashboard
 Files
 Apps
 Tasks

test project ⓘ

Interactive Analysis Settings Notes

QUEUED FastQC Analysis run - 02-04-20 17:42:00 ⓘ
 [Get support](#)
[View stats & logs](#)
[Abort](#)
[Edit and rerun](#)

Queued on Feb. 4, 2020 12:44 by danielventre

Spot Instances: **On** ⓘ Memoization: **Off** ⓘ Progress: **The task has been submitted and is awaiting execution.** Duration: **Less than a minute** ⓘ

App: FastQC Analysis - Revision: 0

Inputs ⓘ

FASTQ Reads ⓘ ⓘ
 C835.HCC1143.2.converted.pe_1.fastq
 C835.HCC1143.2.converted.pe_2.fastq

adapters_file ⓘ
 No files selected

contaminants_file ⓘ
 No files selected

limits_file ⓘ
 No files selected

App Settings

Show all ▾

FastQC (#FastQC)
 Amount of memory allocated per job execution.
 ⓘ Determined by the number of input files

Casava ⓘ	No value
Format ⓘ	FASTQ
Kmers ⓘ	7
Nano ⓘ	No value
Nogroup ⓘ	No value
Number of CPUs. ⓘ	Determined by the number of input files
Quiet ⓘ	No value
Threads ⓘ	1

Outputs

FastQC Charts ⓘ No value
 Report ZIP ⓘ No value

BioData **CATALYST**
 Powered by Seven Bridges

Projects ▾
 Data ▾
 Public Gallery ▾
 Public projects ▾
 Developer ▾

danielventre ▾

Dashboard
 Files
 Apps
 Tasks

test project ⓘ

Interactive Analysis Settings Notes

COMPLETED FastQC Analysis run - 02-04-20 17:42:00 ⓘ
 [Get support](#)
[View stats & logs](#)
[Edit and rerun](#)

Executed on Feb. 4, 2020 12:44 by danielventre

Spot Instances: **On** ⓘ Memoization: **Off** ⓘ Price: **\$0.04** ⓘ Duration: **8 minutes** ⓘ

App: FastQC Analysis - Revision: 0

Inputs ⓘ

FASTQ Reads ⓘ ⓘ
 C835.HCC1143.2.converted.pe_1.fastq
 C835.HCC1143.2.converted.pe_2.fastq

adapters_file ⓘ
 No files selected

contaminants_file ⓘ
 No files selected

limits_file ⓘ
 No files selected

App Settings

Show all ▾

FastQC (#FastQC)
 Amount of memory allocated per job execution.
 ⓘ Determined by the number of input files

Casava ⓘ	No value
Format ⓘ	FASTQ
Kmers ⓘ	7
Nano ⓘ	No value
Nogroup ⓘ	No value
Number of CPUs. ⓘ	Determined by the number of input files
Quiet ⓘ	No value
Threads ⓘ	1

Outputs ⓘ

FastQC Charts ⓘ ⓘ
 C835.HCC1143.2.converted.pe_2_fastqc.b64...
 C835.HCC1143.2.converted.pe_1_fastqc.b64...

Report ZIP ⓘ ⓘ
 C835.HCC1143.2.converted.pe_2_fastqc.zip
 C835.HCC1143.2.converted.pe_1_fastqc.zip

Search and access hosted TOPMed studies

If you would like to work with the hosted TOPMed studies (or see the list of available studies), navigate to “Data” on the top menu bar and then select the **“Data Browser.”** This will bring you to a page that has all the TOPMed studies listed by disease type. The phs numbers are listed for each study. If the user wants to see additional details about each of the studies, they can click “Details” to expand the box. This will show which consent groups are included for each study as well as the total number of subjects and files.

Users can select one or more of the hosted studies to bring to the query page of the Data Browser by clicking “Explore selected.”

The screenshot shows the SevenBridges Data Browser interface. At the top, there is a navigation bar with links: Projects, Data (selected), Public Gallery, Public projects, Developer, a notification bell, and a user profile (danielventre). Below the navigation bar is a search bar labeled "New query" and a "Search by ID" button. The main content area is titled "Select dataset(s)" and includes a search input field. Under the "Heart Disease" category, there is a list of datasets. The first dataset, "SAFS NHLBI TOPMed: San Antonio Family Heart Study ...", is selected with a checkbox. A red arrow points to the "Details" link for this dataset. The details for this dataset are expanded, showing consent codes (DS-DHD-IRB-PUB-MDS-RD) and counts for Files (3,308), Samples (1,829), and Subjects (1,787). Other datasets are listed below, including "CHS NHLBI TOPMed: Trans-Omics for Precision Medicine ...".

From the query page, users can see the list of active datasets on the left-hand side. They can search for Files or Subjects from a particular consent group (those that they have approval for on dbGaP), by selecting the “Consent” property under the File or Subject entity. Then click “Add Property” and refresh the number of Files identified in the lower left-hand corner.

SAFS

New query

Edited

Create new query

Queries

Search by ID

Copy files to project

Add entity

Subject

Sample

File

< Consent

Text search

Specify value (optional):

☐ DS-DHD-IRB-PUB-M...

Add property

File

3,308

Export

Details

Analytics

File	Details for NWD576625.b38.irc.v1.cram	Connections
<div><div></div><div>NWD576625.b38.irc.v1.cram</div></div>	<div><div>SAFS</div><div><div>Access level</div><div>Assay Type</div><div>Assembly Name</div><div>Consent</div><div>Coverage</div><div>File Type</div></div><div><div>Controlled</div><div>WGS</div><div>GRCh38</div><div>DS-DHD-IRB-PUB-MDS-RD</div><div>37.64</div><div>CRAM</div></div></div>	<div><div>Inbound:</div><div>No inbound connections</div><div>Outbound:</div><div>No outbound connections</div></div>

Terms

Privacy

Data Use

Copyright

?

The screenshot shows the SevenBridges BioData CATALYST interface. At the top, there's a navigation bar with 'Projects', 'Data', 'Public Gallery', 'Public projects', and 'Developer' tabs. A user profile 'danielventre' is logged in. Below the navigation bar, there's a 'New query' section with a 'Create new query' button and a 'Search by ID' field. A 'Copy files to project' button is also visible. In the center, a modal window titled 'Add entity' is open, showing 'File', 'Subject', and 'Sample' options. The 'File' option is selected, and a 'File Type' sub-modal is open, showing a search bar and checkboxes for 'CRAM' and 'VCF'. The 'VCF' checkbox is checked. Below the modal, a 'File' entity is shown with a count of 3,308. At the bottom, there's a table with three columns: 'File', 'Details for NWD576625.b38.irc.v1.cram', and 'Connections'. The 'File' column lists five files with red lock icons. The 'Details' column shows properties like 'Access level', 'Assay Type', 'Assembly Name', 'Consent', 'Coverage', and 'File Type'. The 'Connections' column shows 'Inbound' and 'Outbound' connection status.

File	Details for NWD576625.b38.irc.v1.cram	Connections
NWD576625.b38.irc.v1.cram	SAFS	Inbound:
NWD429769.b38.irc.v1.cram	Access level ⓘ Controlled	No inbound connections
NWD975468.b38.irc.v1.cram	Assay Type ⓘ WGS	
NWD974028.freeze5.v1.vcf.gz	Assembly Name ⓘ GRCh38	Outbound:
NWD169790.freeze5.v1.vcf.gz	Consent ⓘ DS-DHD-IRB-PUB-MDS-RD	No outbound connections
	Coverage 37.64	
	File Type ⓘ CRAM	

The user can specifically look for VCF files or CRAM files by adding the “File Type” property to the File entity of the Data Browser. Select to Add the Property and then refresh the number of Files identified in the lower left-hand corner. The file names identified are listed in the bottom part of the page with red lock symbols next to them to indicate that they are controlled files.

To bring these files to a project for analysis, select “Copy files to project” in the upper right-hand part of the page.

NIH

BioData CATALYST

Powered by Seven Bridges

Projects ▾

Data ▾

Public Gallery ▾

Public projects ▾

Developer ▾

danielventre ▾

SAFS

New query Edited

Create new query

Queries ▾

Search by ID

Copy files to project

Add entity

Subject

Sample

File

Consent

File Type

Add property

Properties and values

Find specific

Access level

File

1,479 --

Export ▾

Details

Analytics

File	Details for NWD974028.freeze5.v1.vcf.gz	Connections
<div><div></div><div>NWD974028.freeze5.v1.vcf.gz</div></div>	<div>SAFS</div> <div>Access level ⓘ</div> <div>Assay Type ⓘ</div> <div>Assembly Name ⓘ</div> <div>Consent ⓘ</div> <div>Coverage</div> <div>File Type ⓘ</div>	<div>Inbound:</div> <div>No inbound connections</div> <div>Outbound:</div> <div>No outbound connections</div>

Terms

Privacy

Data Use

Copyright

?

Users can search for known TOPMed files of interest using the “Search by ID” function. Paste dbGaP identifiers for the subject or file in the box. Users can search for IDs of files or subjects spanning multiple TOPMed studies at once.

NIH

BioData CATALYST

Powered by Seven Bridges

Projects ▾

Data ▾

Public Gallery ▾

Public projects ▾

Developer ▾

Search by ID

New query

Search by ID

Type your UUIDs, IDs (barcodes), or file names.

UUIDs, IDs (barcodes), or file names, separated by comma or a new line.

500 items max.

Next

AMISH

NHLBI TOPMed: Genetics of Cardiometabolic H...

Details ▾

phs000956.v4.p1 · Description

HVH

NHLBI TOPMed: Heart and Vascular Health Study ...

Details ▾

phs000993.v4.p2 · Description

Terms

Privacy

Data Use

Copyright