

## 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.

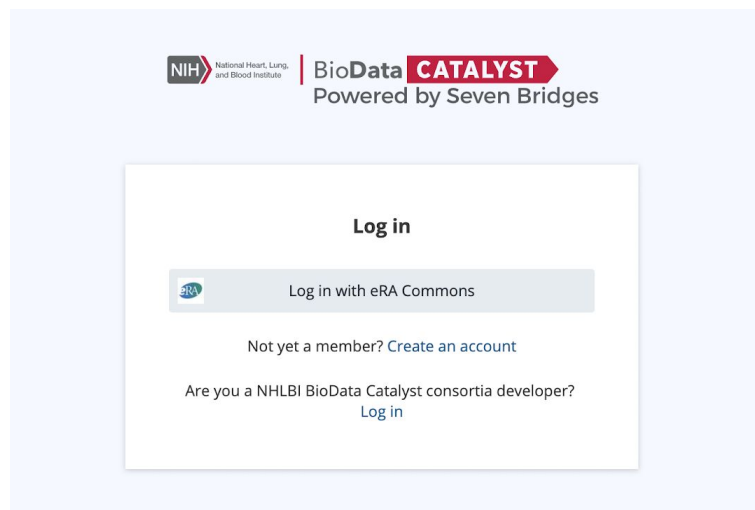
To give lab staff downloader status, please refer to [these instructions](#). **Please note that having other researchers listed on your dbGaP DAR application as internal and external collaborators will not result in these individuals getting access to hosted dataset on BioData Catalyst.** PI's will 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.

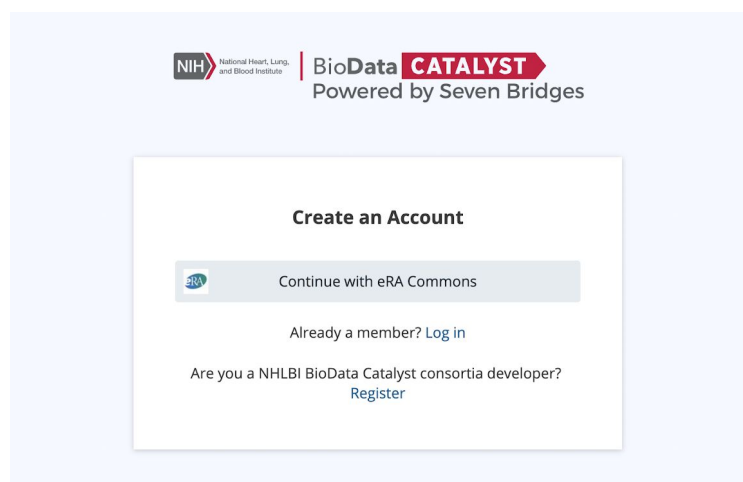
Researchers also have the option to bring their own datasets to the BioData Catalyst platforms. As described in the [BioData Catalyst Data Use Policy](#), users can upload data for which they have the appropriate approval provided that they do not violate the terms of their Data Use Agreements, Limitations, or Institutional Review Board policies and guidelines. To learn more about how to bring your own data to the platform, please refer to the [Documentation](#).

## Account Registration on BioData Catalyst powered by Seven Bridges

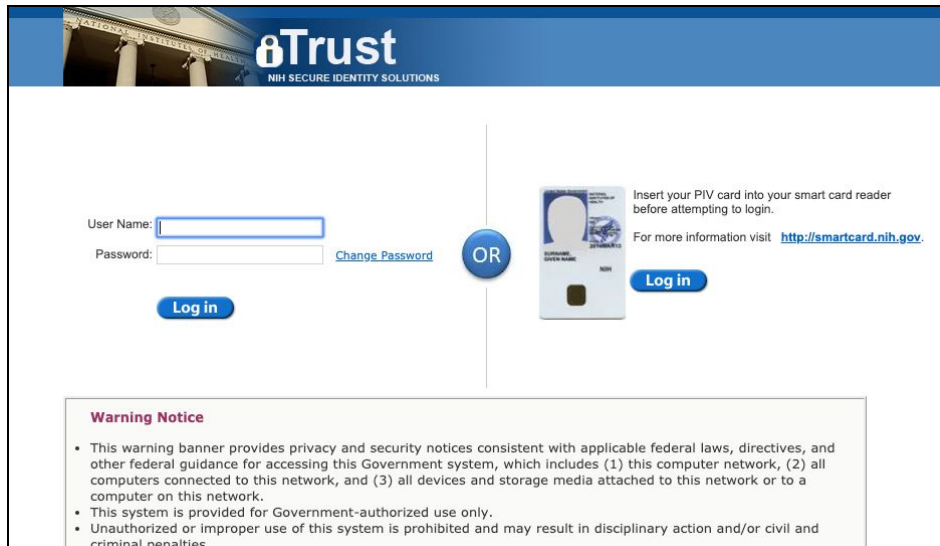
To create an account, please visit the platform login page at <https://platform.sb.biodatacatalyst.nihbi.nih.gov>. Please select “Create an account.”



Accounts should be created with eRA Commons credentials. Select “Continue with eRA Commons.”



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



The iTrust login page features a blue header with the iTrust logo and the text "NIH SECURE IDENTITY SOLUTIONS". Below the header, there are two login options: a standard username and password login, and a PIV card login. The username and password fields are on the left, with a "Log in" button below them. A "Change Password" link is next to the password field. In the center is a blue circle with the word "OR". To the right is a PIV card image with the text "Insert your PIV card into your smart card reader before attempting to login." and a "Log in" button. Below the PIV card image is a link to "http://smartcard.nih.gov". At the bottom, there is a "Warning Notice" box with a red header and a list of three bullet points regarding privacy, security, and unauthorized use.

User Name:   
Password:  [Change Password](#)

[Log in](#)

OR

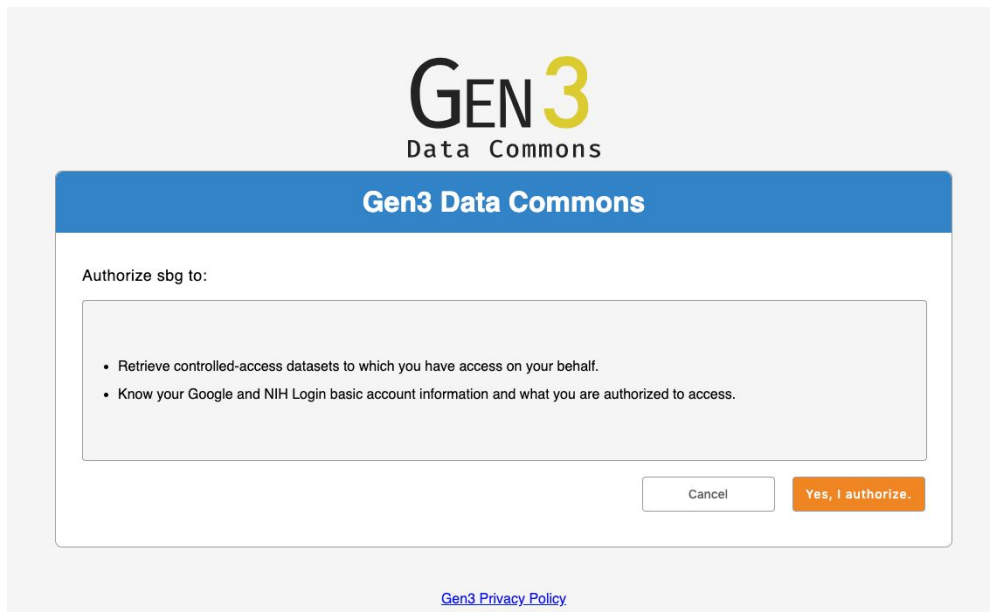
Insert your PIV card into your smart card reader before attempting to login.  
For more information visit <http://smartcard.nih.gov>

[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.

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 Gen3 Data Commons authorization dialog has a blue header with the text "Gen3 Data Commons". Below the header, it says "Authorize sbg to:" followed by a list of two bullet points: "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, there are two buttons: "Cancel" and "Yes, I authorize." Below the dialog, there is a link to "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.

**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 acknowledge the Privacy Act notice below.

Proceed to the platform

This statement is provided pursuant to the Privacy Act of 1974 (5 U.S.C. § 552a): The information requested on this form is authorized to be collected pursuant to 42 U.S.C. 217a, 241, 281, 282, 284; 48 CFR Subpart 15.3; and Executive Order 13478. Completing the form is voluntary, however, declining to provide any or all of the requested information may result in denial of access to controlled data. The principal purpose for which the information will be used is to authenticate users who request access to controlled access data. The information will be used to contact you in response to requests you have specifically made on this Web site. Your personal information may also be used to audit your activity on the system in order to ensure compliance with NIH policies. The information you provide will be included in a Privacy Act system of Records, and will be used and may be disclosed for the purposes and routine uses described and published in the following System of Records Notice (SORN): 09-90-1401, Records About Restricted Dataset Requesters, HHS/OS/Other <https://www.federalregister.gov/documents/2018/03/14/2018-05176/privacy-act-of-1974-system-of-records>

After clicking “Proceed to the platform,” the system will verify your email. You will receive \$500 in pilot credits upon account creation so you can start testing out the platform and running analyses.

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: a platform demo; recommendations on the hosted bioinformatics tools/workflows; tips for accessing the hosted datasets or uploading your own data; and a tailored in-depth training over web call. If you already have a research project planned on the platform, please email the Seven Bridges Program Manager for BioData Catalyst, Alison Leaf ([alison.leaf@sevenbridges.com](mailto:alison.leaf@sevenbridges.com)).

## 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” the first time you 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](mailto: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 contact Alison Leaf ([alison.leaf@sevenbridges.com](mailto:alison.leaf@sevenbridges.com)).

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 create a project to learn more about the options for 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 NIAH logo, 'BioData CATALYST Powered by Seven Bridges', and dropdown menus for 'Projects', 'Data', 'Public Gallery', 'Public projects', 'Developer', a notification bell, and the user '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', followed by the text 'or learn more about projects.' The right panel, titled 'ANALYSES', features a search bar and two tabs: 'Tasks' (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 titled 'Create a project' and has a close button (X) in the top right corner. It contains the following fields and options:

- Name:** A text input field containing 'test project'.
- Project URL:** A text input field containing 'https://f4c.sbgenomics.com/u/danielventre/test-project'.
- Billing Group:** A dropdown menu showing 'Pilot Funds (danielventre)'.
- Location:** A dropdown menu showing 'AWS (us-east-1)'.
- Execution settings:**
  - Spot Instances:** A toggle switch set to 'On'.
  - Memoization:** A toggle switch set to 'Off'.
- Data:** A checkbox labeled 'This project will contain' followed by a red 'CONTROLLED' button and a 'Data' link.

At the bottom of the dialog are 'Cancel' and 'Create' buttons. The background shows the main application interface with a search bar and a 'Create' button.

## 1. Name the project

Following along with the red step indicators in the screenshot above, 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. Each user has a “Pilot Funds” billing group 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](mailto:alison.leaf@sevenbridges.com)) for assistance. For this **example project**, select the Pilot Funds billing group.

### 3. Location

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 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 the cloud location 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 stored on both AWS us-east-1 and 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 trends in supply and demand. 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 may 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 automatically 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 and consists of 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 shows 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 for danielventre. The main content area is divided into three panels:

- DESCRIPTION:** Contains a welcome message, an explanation of projects as core building blocks, and a list of actions users can take within a project (e.g., exploring public datasets, installing tools, uploading data, collaborating). It also includes a note about recording project details and a markdown editor for notes.
- MEMBERS:** Shows the user "danielventre" as the "OWNER" with permissions to "Write, Copy, Execute, Admin". It includes a button to "Invite new members" and a message encouraging collaboration.
- ANALYSES:** Features a search bar and a list of tasks. The "Data Cruncher" task is currently selected.

A large "NHLBI" watermark is visible across the bottom of the interface.

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

fastq 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

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.

NIH | BioData CATALYST Powered by Seven Bridges

Projects Data Public Gallery Public projects Developer danielventre

Dashboard Files Apps Tasks test project Interactive Analysis Settings Notes

Use apps to get results.  
Wrap your own or add existing.

Create + Add app

or learn more about apps

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. Select the “x” in the top right corner to go back to the Apps tab of the project.

The screenshot shows the 'Add apps to "test project"' interface. At the top, there are tabs for 'Public Apps', 'Projects', 'My Apps', and 'Create New App'. Below the tabs is a search bar containing 'fastqc', with filters for 'Category' and 'Toolkit', and a 'Reset search' button. The search results are displayed in a grid of six app cards. The first card, 'FastQC Analysis', is highlighted with a red arrow pointing to its 'Copy' button. Each card includes a description, tags for 'QUALITY-CONTROL', 'FASTQ-PROCESSING', and 'QUANTIFICATION', and 'Copy' and 'Run' buttons. The other cards are 'ChIP-seq FastQC', 'FastQC', 'SBG FastQC Beautifier', 'SBG FastQC Extract', and 'Bismark Analysis'.

Add apps to "test project" ✕

Public Apps Projects My Apps Create New App

fastqc 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/workflow along with helpful information like the required inputs, outputs, and common issues (see below).



The screenshot shows the BioData CATALYST interface, powered by Seven Bridges. The top navigation bar includes links for Projects, Data, Public Gallery, Public projects, Developer, and a user profile for danielventre. The main navigation bar has tabs for Dashboard, Files, Apps (selected), and Tasks. The current view is for the 'test project'.

The 'FastQC Analysis' app page is displayed. It features a 'COPY' button and a 'Revision 0' dropdown. The app is described as a copy of the latest revision by danielventre on Feb. 4, 2020 at 12:39. The description states that the FastQC tool, developed by the Babraham Institute, analyzes sequence data from FASTQ, BAM, or SAM files to identify technical problems. It provides instructions on how to use the pipeline on sequencing data.

The 'Required inputs' section specifies that the app takes FASTQ Reads (ID: FASTQ\_reads) as input. It notes that the number of threads can be set, or it will default to one CPU core per file. It also mentions that the task will fail if the number of files is too large for the instance's CPU cores.

The 'Outputs' section indicates that the app produces a Report ZIP (ID: report\_zip), which is a ZIP archive containing the FastQC HTML report and dependencies.

On the right side, the 'Basic Information' section lists the following details:

- 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
- Links: [Homepage](#), [Documentation](#)

The 'Workflow steps' section shows a single step: 'FastQC'.

To set up the draft Task, select “Run” on the App from the Apps tab. 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 shows the hosted TOPMed studies listed by disease type. The phs numbers are listed for each study. Additional details about each of the studies can be viewed by selecting “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.

BioData Catalyst hosts TOPMed studies from Freeze5 and has CRAM files and single-sample VCF files for all subjects. There are also multi-sample VCFs and phenotype files on a per study per consent group basis. In addition, BioData Catalyst also hosts TOPMed parent studies with phenotype files. Follow the instructions below to see how to find specific file types for a TOPMed study and consent group.

Starting from the dataset selection page, select the study “NHLBI TOPMed: The Jackson Heart Study.” All TOPMed studies (genomic data) have the preface “NHLBI TOPMed” in the name. We will query only one dataset, however users have the option to select several hosted studies at once. Click “Explore 1 selected” to go to the query page of the Data Browser.

The screenshot shows the BioData Catalyst Data Browser interface. The top navigation bar includes 'Projects', 'Data' (selected), 'Public Gallery', 'Public projects', 'Automations', 'Developer', 'Staff', and a user profile 'alisonleaf'. Below the navigation bar is a 'New query' button and a search bar labeled 'Search by ID'. The main content area is titled 'Select dataset(s)' and features a search bar for 'Study name, abbreviated name, or dbGaP accession number'. A list of datasets is shown under the 'Heart Disease' category. The 'JHS NHLBI TOPMed: The Jackson Heart Study' dataset is selected, and its details are expanded, showing consent codes and counts for files, samples, and subjects.

Dataset	Consent codes	Files	Samples	Subjects
FHS Framingham Cohort	phs000007.v30.p11 · Description			
FHS NHLBI TOPMed: Genomic Activities such as Whole...	phs000974.v4.p3 · Description			
JHS The Jackson Heart Study (JHS)	phs000286.v6.p2 · Description			
JHS NHLBI TOPMed: The Jackson Heart Study	phs000964.v4.p1 · Description	6,500	3,406	3,406

From the query page, users can see the list of active datasets on the left-hand side. They can search over several different entities to find files on the platform including Subject, Sample, and File. Each of these entities has a number of properties that you can filter through to search the file metadata. Select the File entity to begin this search. Your screen will look like the second image below.

BioData CATALYST

Powered by Seven Bridges

Projects ▾

Data ▾

Public Gallery ▾

Public projects ▾

Automations

Developer ▾

Staff ▾

🔔 ▾

alisonleaf ▾

JHS

New query

🔍 Search by ID

Explore JHS Dataset

🔍 e.g. CRAM, WGS, Blood, Heart, Male ...

Choose an entity

Subject ⓘ

Sample ⓘ

File ⓘ

Study ⓘ

Saved

Examples

FHS Test

👁

🗑

BioData CATALYST

Powered by Seven Bridges

Projects ▾

Data ▾

Public Gallery ▾

Public projects ▾

Automations

Developer ▾

Staff ▾

🔔 ▾

alisonleaf ▾

JHS

New query Edited

Create new query

Queries ▾

🔍 Search by ID

📄 Copy files to project

Add entity

Subject ⓘ

Sample ⓘ

File ⓘ

Filter by ⓘ x

Properties and values 🔍

File name ⓘ

Access level ⓘ

Assay Type ⓘ

Assembly Name ⓘ

Consent ⓘ

Coverage ⓘ

Add entity

Study ⓘ

🔄

File

6,500 --

Export ▾

📄 Details

📊 Analytics

File	Details for NWD833816.b38.irc.v1.cram	Connections
<div><div>📄</div><div>NWD833816.b38.irc.v1.cram</div></div> <div><div>📄</div><div>NWD159406.freeze5.v1.vcf.gz</div></div> <div><div>📄</div><div>NWD163918.freeze5.v1.vcf.gz</div></div>	<div>JHS</div> <div><div>Access level ⓘ</div><div>Assay Type ⓘ</div><div>Assembly Name ⓘ</div></div> <div><div>Controlled</div><div>WGS</div><div>GRCh38</div></div>	<div>Inbound:</div> <div>No inbound connections</div> <div>Outbound:</div> <div>No outbound connections</div>

Privacy Policy

Data Sharing Policy

Freedom of Information Act (FOIA)

Accessibility

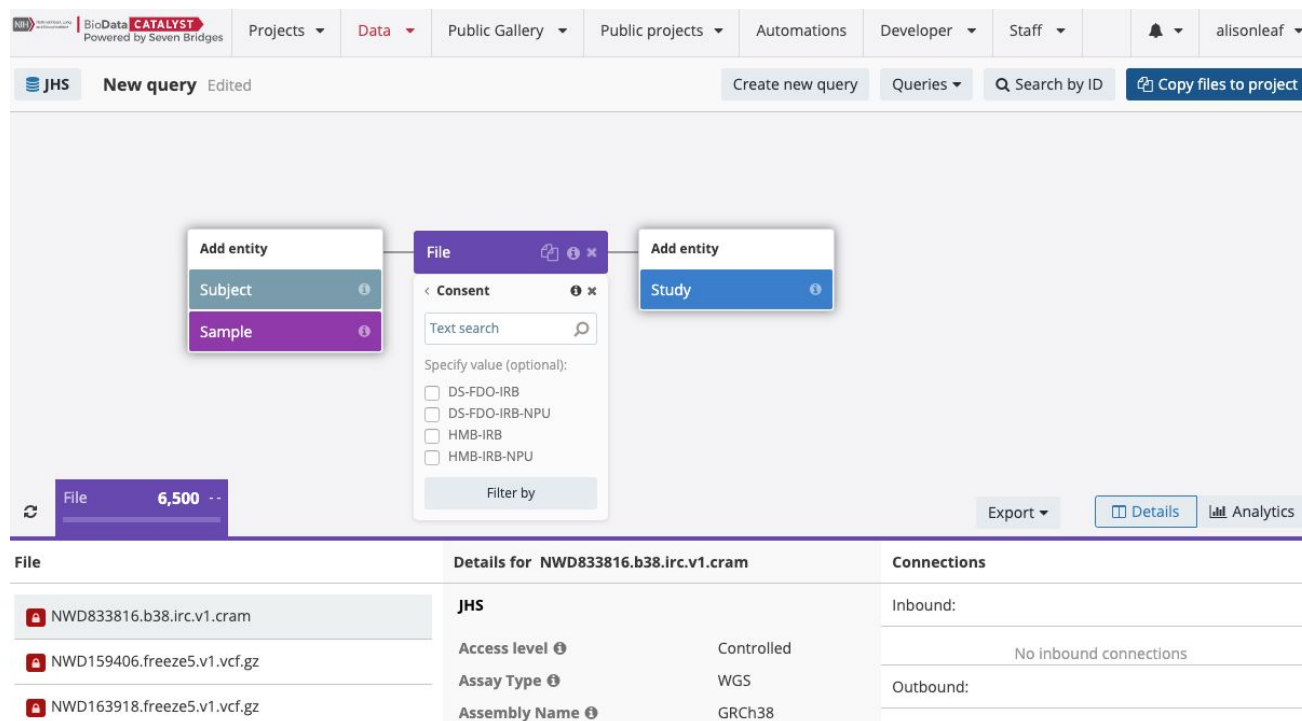
U.S. Department of Health & Human Services

National Institutes of Health

National Heart, Lung, and Blood Institute

19 sevenbridges.com

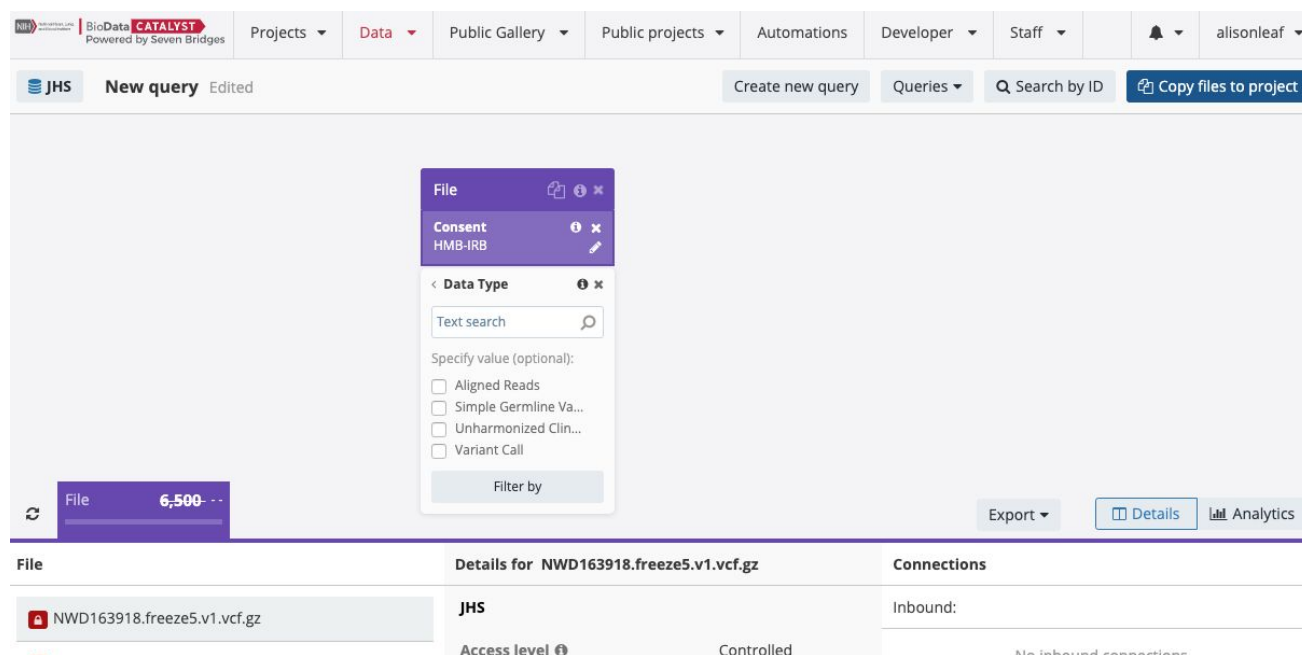
Now you can select the consent group you want to work with. In the File entity, select the “Consent” property. All of the consent groups in the selected study will be listed as shown below:



The screenshot shows the BioData CATALYST interface. A query is being built with a 'File' entity. The 'Consent' property is selected, and a list of consent groups is displayed. The 'File' entity is named 'File' and has 6,500 records. The 'Details' tab for the selected file shows 'JHS' as the study, 'Controlled' as the access level, 'WGS' as the assay type, and 'GRCh38' as the assembly name.

File	Details for NWD833816.b38.irc.v1.cram	Connections
NWD833816.b38.irc.v1.cram	JHS	Inbound:
NWD159406.freeze5.v1.vcf.gz	Access level: Controlled	No inbound connections
NWD163918.freeze5.v1.vcf.gz	Assay Type: WGS	Outbound:
	Assembly Name: GRCh38	

As an example, select HMB-IRB and “Filter by.” Now add the Property “Data Type,” which will show the 4 different types of files for this study: aligned reads (CRAMs), Simple Germline Variation (single-sample VCFs), Unharmonized Clinical Data (phenotype files on a per study and consent group basis), and Variant Call (multi-sample VCFs on a per study and consent group basis).



The screenshot shows the BioData CATALYST interface. A query is being built with a 'File' entity. The 'Data Type' property is selected, and a list of data types is displayed. The 'File' entity is named 'File' and has 6,500 records. The 'Details' tab for the selected file shows 'JHS' as the study, 'Controlled' as the access level, and 'No inbound connections' as the connection status.

File	Details for NWD163918.freeze5.v1.vcf.gz	Connections
NWD163918.freeze5.v1.vcf.gz	JHS	Inbound:
	Access level: Controlled	No inbound connections

Select Aligned Reads and “Filter by” to find the CRAM files for this study and consent group. It is important to refresh the results in the lower left corner next to the file tab. This will show the total number of CRAM files for this study and consent group on BioData Catalyst.

The screenshot shows the BioData Catalyst interface. At the top, there's a navigation bar with 'BioData CATALYST Powered by Seven Bridges' and various tabs like 'Projects', 'Data', 'Public Gallery', etc. Below this, a 'New query' section is visible. A filter panel is open, showing 'File' (2,018), 'Consent' (HMB-IRB), and 'Data Type' (Aligned Reads). The 'Filter by' section is also visible. Below the filter panel, a table lists the results. The first column is 'File', showing three CRAM files with red lock icons. The second column is 'Details for NWD132617.b38.irc.v1.cram', showing 'JHS' as the project, 'Access level' as 'Controlled', 'Assay Type' as 'WGS', and 'Assembly Name' as 'GRCh38'. The third column is 'Connections', showing 'Inbound' and 'Outbound' sections, both indicating 'No inbound/outbound connections'.

File	Details for NWD132617.b38.irc.v1.cram	Connections
NWD132617.b38.irc.v1.cram	JHS	Inbound:
NWD354668.b38.irc.v1.cram	Access level ⓘ Controlled	No inbound connections
NWD654937.b38.irc.v1.cram	Assay Type ⓘ WGS	Outbound:
	Assembly Name ⓘ GRCh38	No outbound connections

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 link these files to a project for analysis, select “Copy files to project” in the upper right-hand part of the page. Please note that users can only link files to a project if they are approved to access those files on dbGaP. Therefore, if you are not approved for the “NHLBI TOPMed: The Jackson Heart Study” consent group HMB-IRB, the platform will prevent you from bringing these controlled files to a project. Please also note that because these are controlled files, they must be linked to a project marked as controlled.

Only open access metadata is available for users to search over in the Data Browser, so all BioData Catalyst users can view all available studies on BioData Catalyst and perform the same searches.