

# BioData Catalyst Half-day Workshop

Wednesday, April 27th at 1 p.m. ET

**We will get started shortly.**



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

Interact with us on our forum during today's workshop: <https://bit.ly/3kg5LJk>

# BioData Catalyst Half-day Workshop

Wednesday, April 27th at 1 p.m. ET

**Welcome! Let's get started.**



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.

Resource: [Statement of Conduct](#)

# Agenda

Topic	Time
<a href="#">Introductions and Housekeeping</a>	5 min
<a href="#">What is BioData Catalyst?</a>	10 min
<a href="#">Interactive Demo: Finding and Using NHLBI Hosted Data</a>	30 min
<a href="#">Bring Your Own Data</a>	5 min
<a href="#">Tools, Workflows, and Interactive Analysis</a>	10 min
<a href="#">Understanding, Estimating, and Managing Cloud Costs</a>	5 min
<a href="#">Interoperability</a>	10 min
<a href="#">Researcher Presentation and Q&amp;A: Use of TOPMed WGS as Public Controls, Ravi Mathur</a>	30 min
<b>COFFEE AND SNACK BREAK - 30 MINUTES</b>	
<a href="#">GENESIS Workflows</a>	1h 45min



# Introductions and Housekeeping



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Meet Your Hosts



**Alisa Manning**

*BioData Catalyst Powered by Terra  
Broad Institute*



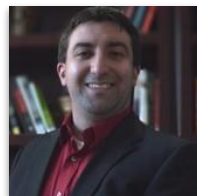
**Emily Hughes**

*BioData Catalyst Powered by PIC-SURE  
Harvard Medical School*



**Rebecca Boyles**

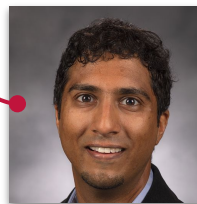
*RTI International*



**Tony Patelunas**

*BioData Catalyst Powered by Seven Bridges  
Seven Bridges*

**Thank you to our guest researcher**



**Ravi Mathur**

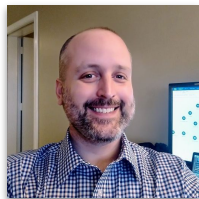
*Statistician, RTI International  
Use of TOPMed WGS as Public  
Controls on BioData Catalyst*

# Have a question during the workshop?

Ask questions **at any time** for live support:

<https://bdcatalyst.freshdesk.com/support/discussions> → CHARGE Workshop

Slides and recording will be posted to the forum, so make sure to **Follow** !



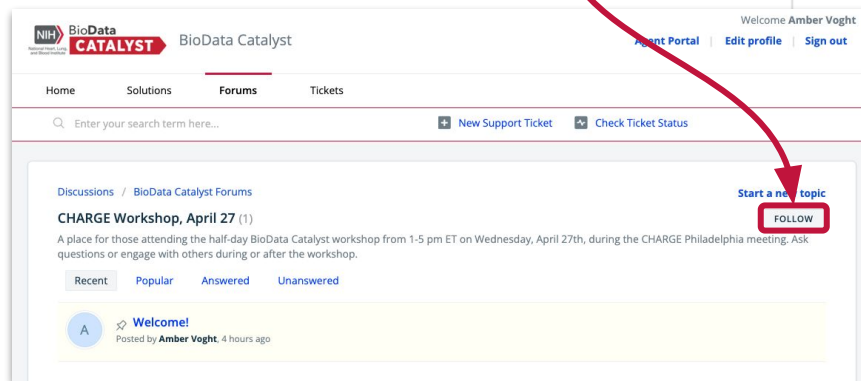
**Dave Roberson**

Community Engagement Specialist,  
Seven Bridges



**Amber Voght**

User Engagement Specialist,  
BioData Catalyst Coordinating Center



# Questions before we begin?

**Next up:** What is BioData Catalyst?

# What is BioData Catalyst?

Rebecca Boyles, RTI International



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

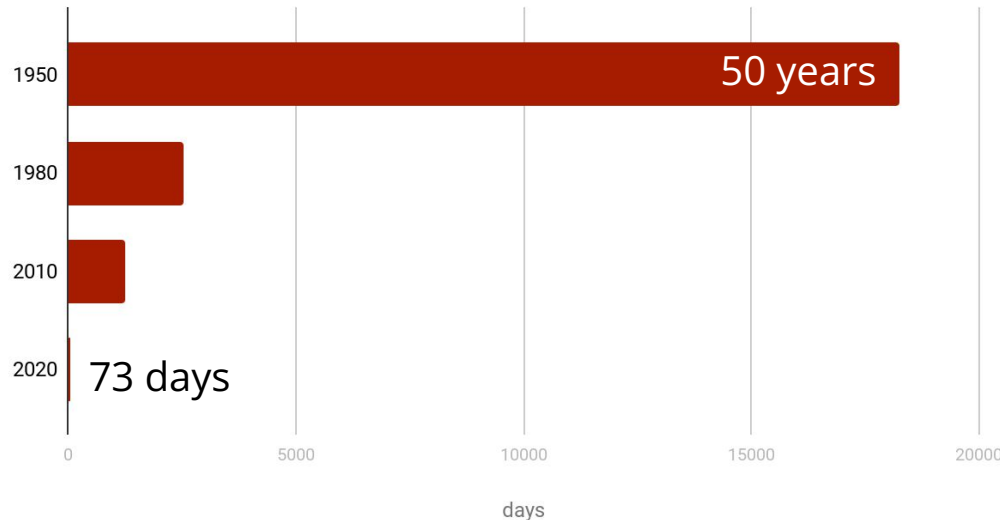
# Let's talk about:

- **Intro to BioData Catalyst**
  - Data growth
  - Mission and vision
  - Platform overview
- Where to find more information
- How and why to get involved in the community



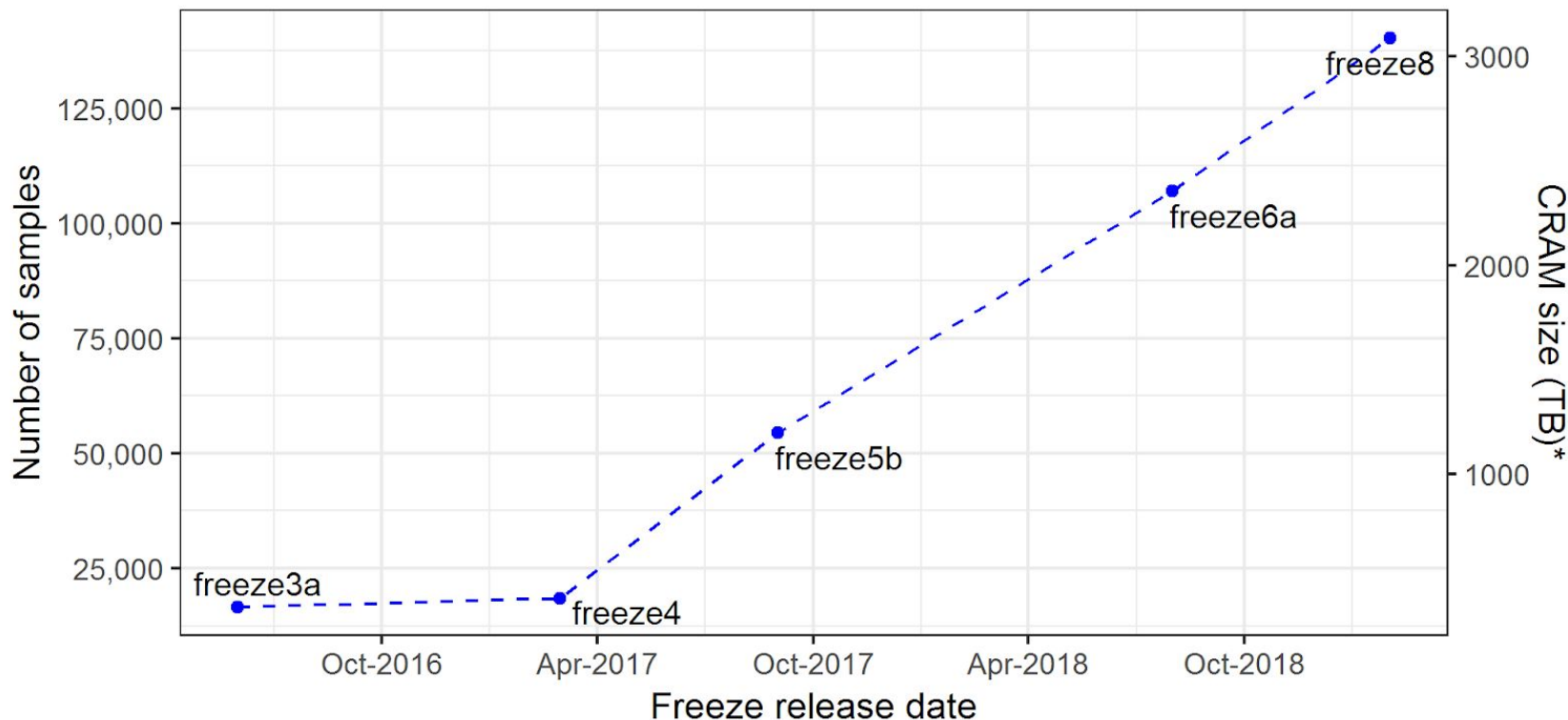
# The rate of data generation is accelerating rapidly

Doubling Time of Health Knowledge



- More biomedical data will be generated this year than all previous years **combined**
- Diverse data modalities including Health data, Survey, Sequencing, Imaging, Metabolomics, Proteomics, Sensor, E-Phys, Flow Cytometry, and so on

## Growth in TOPMed Genome Sequence Data



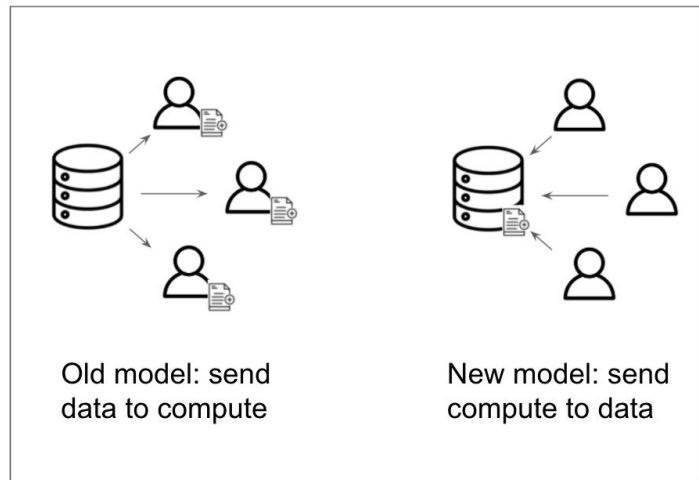
\*Based on average size of 22 GB for 1 DNA sample



# Using the Cloud to store and analyze growing health data

Data Growth | Mission & Vision | Overview

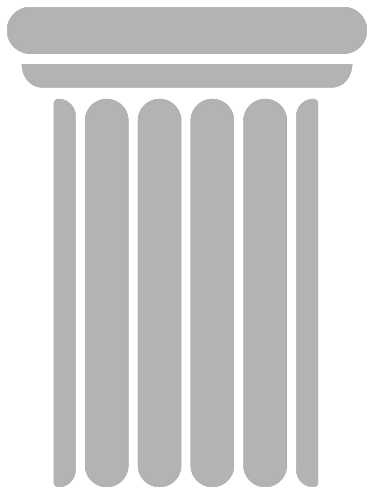
- Immediate scaling -- no need to wait to purchase and install hardware.
- Levels the playing field -- even researchers at institutions without large compute infrastructure investments can access powerful data and compute resources.
- Many researchers can access data without needing to physically copy it.
- Data and methods in a single place streamlines reproducibility.



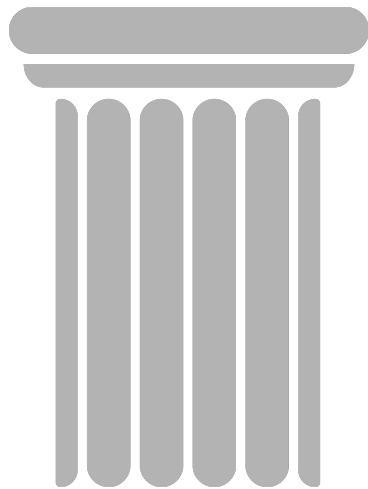
# NHLBI BioData Catalyst

Data Growth | [Mission & Vision](#) | Overview

## Mission



## Vision



The *mission* is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.

The *vision* is to be a community-driven ecosystem implementing data science solutions to democratize data and computational access to advance Heart, Lung, Blood, and Sleep science.

**WHO?**



**WHAT?**



Genomics



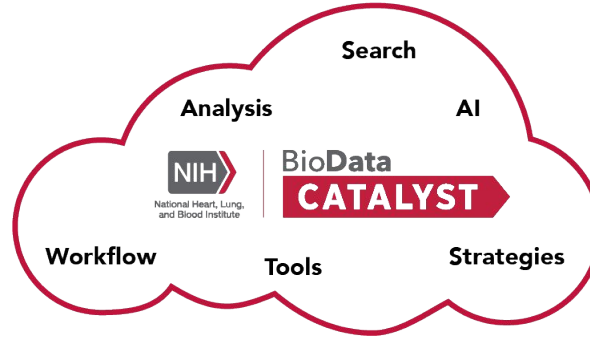
Clinical



Imagery

DATA  
HARMONIZATION

**WHERE?**



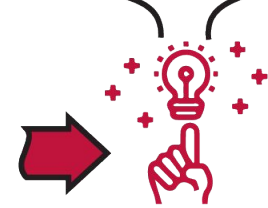
- UNDERSTAND
- OPEN SCIENCE
- CROSS-LINK

- COLLABORATE
- SCALE
- SHARE
- INTEROPERATE

**HOW?**

**SCIENCE!**

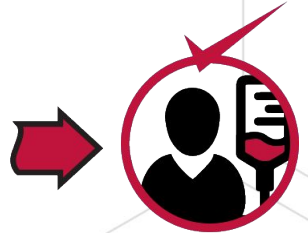
Diagnostics Tools      Therapeutic Options



DISCOVERY

Prevention Strategies

**WHY?**



PATIENTS!

# What BioData Catalyst offers

Data Growth | Mission & Vision | [Overview](#)



## Managing the Computing Environment

Elastic Computing



## Easier Access to many High Value Datasets



## Tooling

Data Discovery

Statistical Analysis  
Tools (R, SAS)

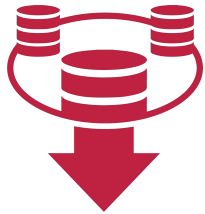
Other Specialized  
Workflows



## Community and Peer Interactions

# The Computing Environment

Data Growth | Mission & Vision | [Overview](#)



No need to  
**download** and  
**manage**  
(multiple) large  
datasets



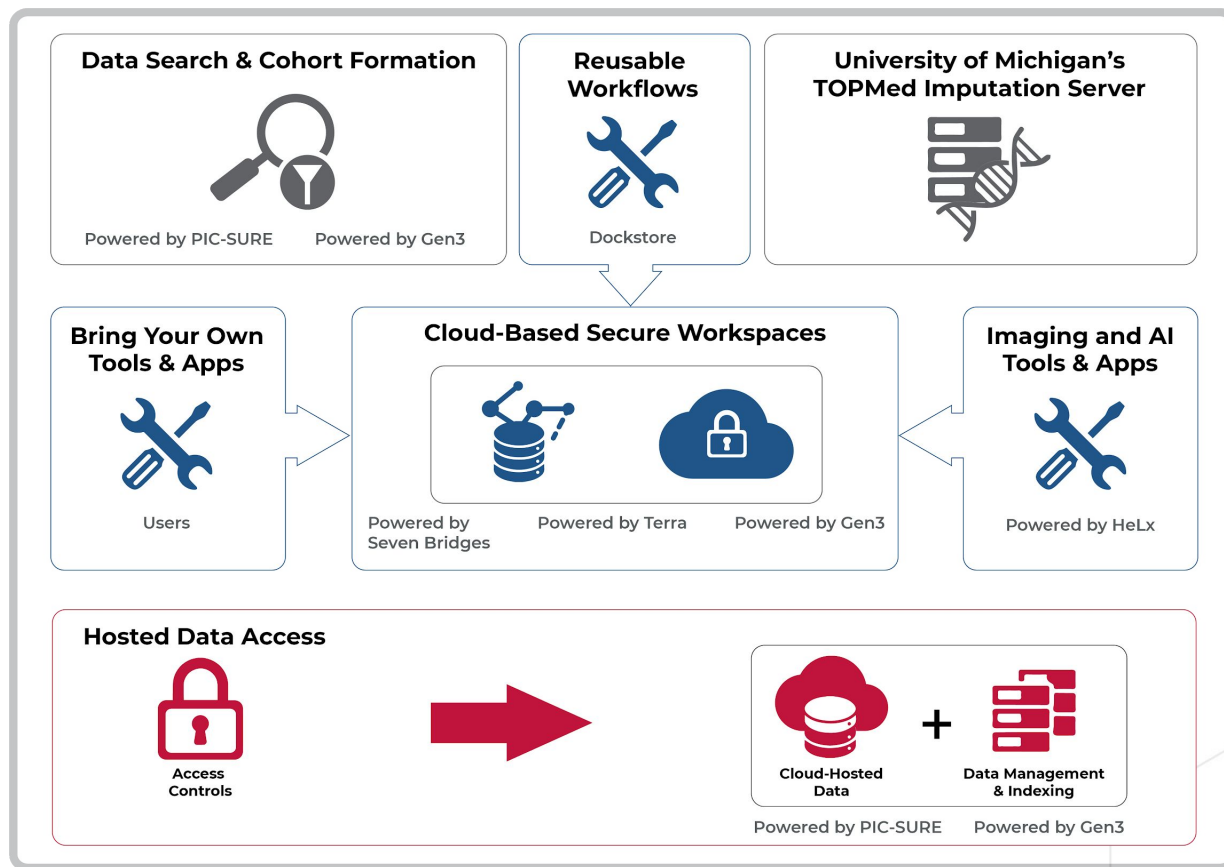
No **computer**  
**system** to  
**manage**



Pay **only** for what  
you **use**



**Help desk** and  
**documentation**



# Let's talk about:

- Intro to BioData Catalyst
- **Where to find more information**
  - Platforms and Services
  - Learning resources
- How and why to get involved in the community



# Platforms and Services

## Explore Data

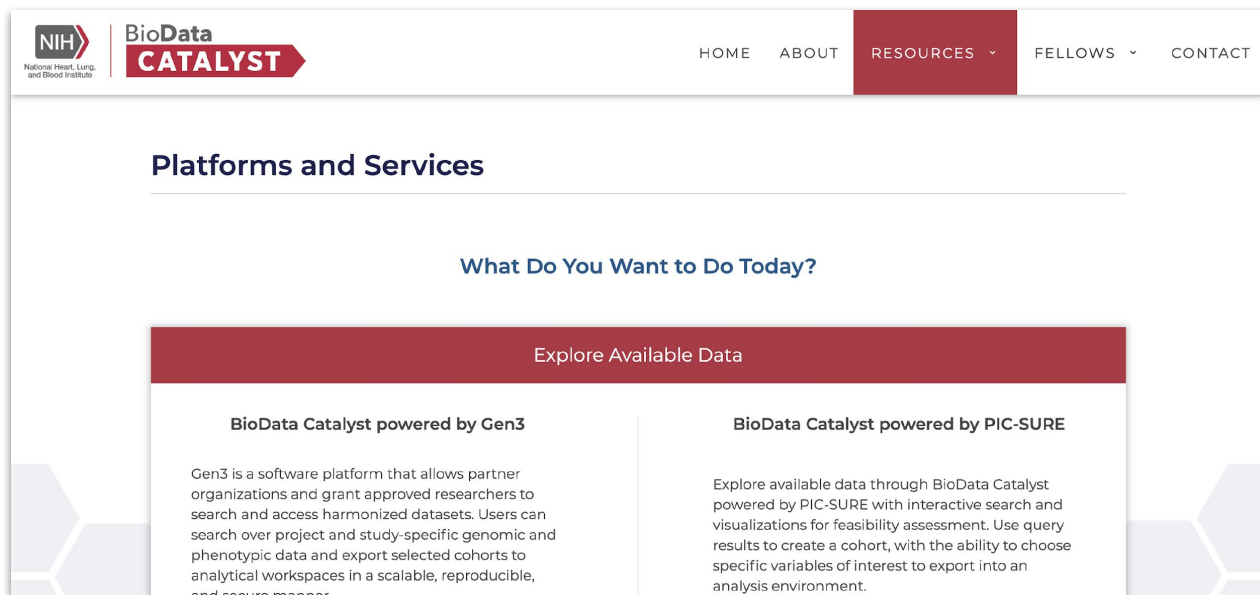
- PIC-SURE
- Gen3

## Analyze Data

- Seven Bridges
- Terra

## Community Tools

- Dockstore



Web resource: [Services](#)



# Learning Resources

Many of the questions you have as a new user may already be answered on either the BioData Catalyst Gitbook or one of the Platform websites.

Our Gitbook documentation includes:

- Instructions on approvals and accounts needed to access BioData Catalyst and how to check your data access
- User Guides for PIC-SURE, Gen3, Seven Bridges, Terra, and Dockstore

**Website resource:** [Learn](#)

**Documentation Resource:** [BioData Catalyst Documentation](#)



You can also find **videos** on our [YouTube channel](#)

# How do I find more information on learning about BioData Catalyst?

- Connect to and learn about the **Platforms and Services** available on the [Services page](#)
- Get started on the ecosystem with [our collection](#) of **learning materials**
- Find our **documentation** in [GitBook](#)
- Subscribe to our **channel** on [YouTube](#)

# Let's talk about:

- Intro to BioData Catalyst
- Where to find more information
- **How and why to get involved in the community**



# Community engagement and support

*Though the primary goal of the NHLBI BioData Catalyst project is to build a data science platform, at its core, this is a people-centric endeavor. BioData Catalyst is also building a community of practice working to collaboratively solve technical and scientific challenges.*



- User-driven, vibrant community
- Peer-to-peer mentoring
- Support available via platforms
- Community Forum
- Community Hours & Showcases

# Community Hours

## BioData Catalyst 101

Wednesday, May 18 at 1 pm ET

<https://bit.ly/38aQ6bs>

**Sign up now!**

View [past materials](#) on our forum

- Curated notes, slides, and recordings on a **variety** of topics, including:
  - Exploring and Accessing Data
  - Interactive Analysis
  - Cloud Costs
  - Reproducible Research Methods
  - Researcher showcases
  - [And more](#) !



You can also find **recordings** on our [YouTube channel](#)

**If you haven't already...**

**Join the NHLBI BioData  
Catalyst Community**

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

# Questions?

**Next up:** Interactive Demo: Finding and Using NHLBI Hosted Data

# Interactive Demo: Finding and Using NHLBI Hosted Data

Emily Hughes, PIC-SURE



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**



# Data available in BioData Catalyst

- The BioData Catalyst ecosystem currently hosts a number of controlled and open datasets:
  - [Trans-omics for Precision Medicine \(TOPMed\)](#) - includes CRAM files, multi-sample VCF files (Freeze8 and Freeze5), study phenotypes, and harmonized phenotypes, with WGS for over 140,000 individuals (Freeze 9 will expand to WGS for over 158,000 individuals, Freeze 10 - >180,000)
  - 1000 Genomes Project
  - PETALNet ORCHID Hydroxychloroquine Trial Data (COVID-19)
  - PETALNet RED CORAL Repository of Electronic Data (COVID-19)
  - BioLINCC Teaching Datasets (Framingham and CAMP)
  - Sickle Cell Disease Datasets (HCT for SCD, BabyHug, Walk-PhaSST, MSH, CSSCD, STOP-II)
- Coming soon:
  - Additional BioLINCC Teaching and Clinical Trials Datasets
  - Additional studies curated by the Cure Sickle Cell Initiative (clinical trials and cohorts)
  - Additional TOPMed data (rolling basis)
  - COVID-19 data (PETALNet Trials, MIS-C, C3PO, ACTIVE4a, etc.)
  - Pediatric Cardiac Genomics Consortium (PCGC) data

# Data available in BioData Catalyst

3.42  
Petabytes of  
data



490,000+  
Data files



280,000+  
Participants



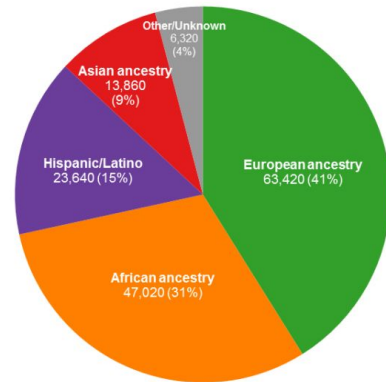
150,000+  
Whole genomes



## TOPMed Data

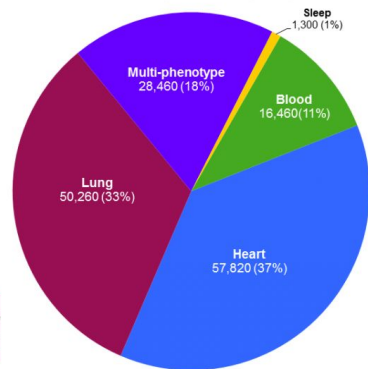
### Ancestry & Ethnicity

Phases 1-6 (~155K Participants)



### Phenotype Focus

Phases 1-6 (~155K Participants)



Hemophilia  
Sickle Cell Disease  
Platelets  
Lipids

Hypertension  
Myocardial Infarction  
Coronary Artery Disease  
Stroke  
Small Vessel Disease  
Venous Thromboembolism  
Congenital Heart Disease  
Atrial Fibrillation  
Coronary Artery Calcification  
Adiposity  
Congestive Heart Failure

Asthma  
Chronic Obstructive Pulmonary Disease  
Idiopathic Pulmonary Fibrosis  
Sarcoidosis  
Interstitial Lung Disease

# Check Access to Data

Three main ways to check your access to data:

1. BioData Catalyst website

- **Demo:** About BioData Catalyst Dataset, <https://biodatacatalyst.nhlbi.nih.gov/resources/data>

2. BioData Catalyst Powered by Gen3

- **Demo:** Exploring files on Gen3, <https://gen3.biodatacatalyst.nhlbi.nih.gov/explorer>

3. BioData Catalyst Powered by PIC-SURE Data Access Dashboard

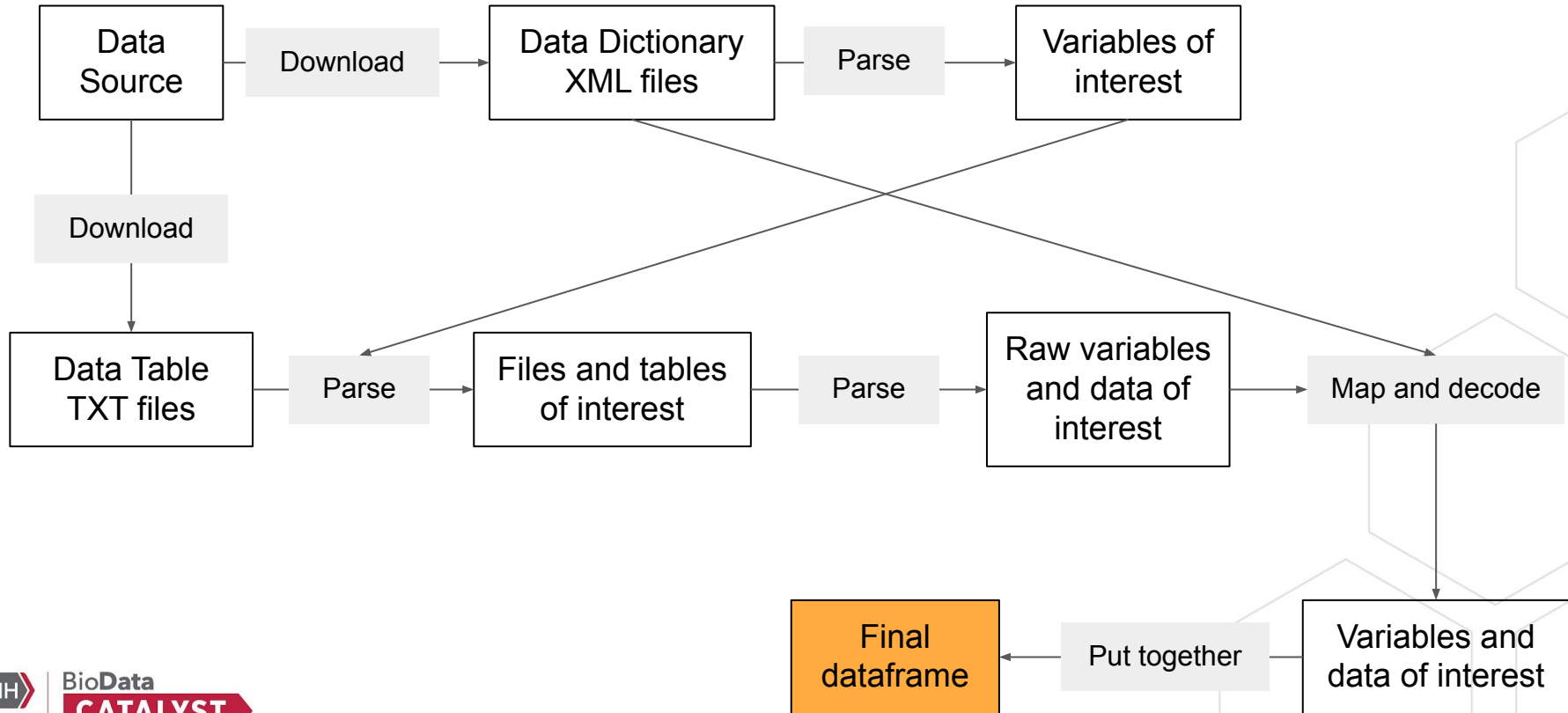
# *BioData Catalyst Powered by PIC-SURE*

Patient  
Information  
Commons  
-  
Standard  
Unification of  
Research  
Elements

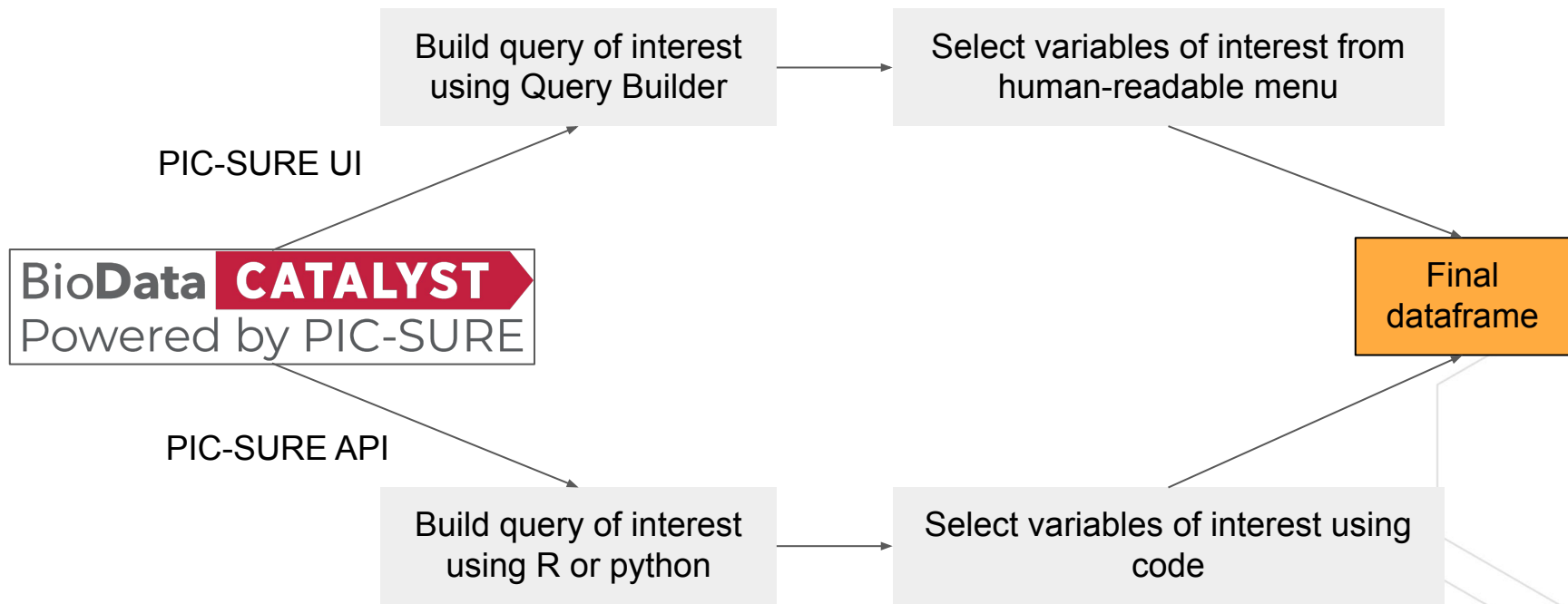
- Allows for searching and exporting data at the **variable** and **variant** level
- Integrates clinical and genomic datasets across BioData Catalyst
- UI allows users to search available data using queries to build cohorts
- Results can be exported via the API for analysis

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

# Traditional Complex Workflow



# PIC-SURE workflow - 2 options



# Open vs Authorized Access

	PIC-SURE Open Access	PIC-SURE Authorized Access
<b>Overview</b>	Allows any user with eRA Commons ID to search any clinical variable in PIC-SURE	Allows users with dbGaP authorization to access data and export to analysis platforms
<b>Access authorization</b>	No approval required, just eRA Commons ID	dbGaP authorization required
<b>Data types</b>	Destigmatized clinical variables	All phenotypic and genomic data
<b>Results</b>	Aggregate counts based on queries	Participant-level data
<b>Use case</b>	Explore datasets to request access to based on query of interest	Filter datasets to cohort of interest to run analyses

# PIC-SURE Data Access Dashboard

**Data Access** tab of PIC-SURE provides summary of Authorized and Open Access and a table view of the available studies.

## Demo

<https://picsure.biodatacatalyst.nhlbi.nih.gov/picsureui/dataAccess>

**BioData CATALYST**  
Powered by PIC-SURE

Authorized Access | Open Access | **Data Access** | User Profile | Help | Log Out

### Authorized Access

[Explore Now](#)

1 Studies  
479 Participants

- dbGap Approval Required
- Authorized Phenotypic and Genomic Datasets
- Aggregate Counts
- Patient Level Data
- Download Authorized Datasets
- R and Python API Access

### Open Access

[Explore Now](#)

61 Studies  
285,864 Participants

- No Authorization Required
- All Phenotypic Datasets Available in PIC-SURE
- Aggregate Counts Only

(Current TOPMed data is Freeze5b)  
P = Phenotype, G = Genomic, P/G = Phenotype/Genomic, n/a = Not Applicable

Identifier	Code	Data Type	Name	Consents	Clinical Variables	Participants with Phenotypes	Samples Sequenced	Version	Participant Number	Access
phs002299	ORCHID	P	COVID19-ORCHID	Health/Medical/Biomedical (HMB)	47	479	n/a	v1	p1	<a href="#">Granted</a>
phs000956	AMISH	P	NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish	Health/Medical/Biomedical (IRB, MDS) (HMB-IRB-MDS)	76	1,123	n/a	v4	p1	<a href="#">Request</a>
phs000280	ARIC	P	Atherosclerosis Risk in Communities (ARIC) Cohort	Health/Medical/Biomedical (IRB) (HMB-IRB)	18,665	15,610	n/a	v7	p1	<a href="#">Request</a>



# PIC-SURE Open Access

**Open Access** provides an intuitive, “Google-like” experience to search variables of interest and retrieve aggregate counts for each study.

## Demo

<https://picsure.biodatacatalyst.nhlbi.nih.gov/picsureui/openAccess>

The screenshot displays the BioData CATALYST Open Access interface. At the top, the NIH logo and "BioData CATALYST Powered by PIC-SURE" are visible. Below the header, there are navigation tabs: "Authorized Access", "Open Access" (selected), "Data Access", "User Profile", "Help", and "Log Out".

The main section is titled "QUERY BUILDER" and contains three query blocks, each with a "back", "delete", and "edit" button:

- Block 1: "DCC\_Harmonized\_data\_set" (highlighted in red), "Body height at baseline., Any Value".
- Block 2: "DCC\_Harmonized\_data\_set" (highlighted in red), "age at measurement of height\_baseline\_1, Greater than or equal to 18".
- Block 3: "DCC\_Harmonized\_data\_set" (highlighted in red), "Subject sex as recorded by the study., Restrict By Value Female".

On the right side, a large blue box displays the total number of participants: "189163 ±3". Below this, the text "Total Participants" is followed by a list of studies with their participant counts and a "Request Access" button for each:

- AMISH (phs000956) 555 participants
- ARIC (phs000280) 8149 ±3 participants
- ARIC (phs001211) 7064 ±3 participants
- CARDIA (phs000285) 2014 participants
- CFS (phs000284) 594 participants
- CFS (phs000954) 524 participants
- CHS (phs000287) 3147 ±3 participants
- CHS (phs001368) 2069 ±3 participants
- COPDGENE (phs000179) 4847 participants
- COPDGENE (phs000951) 4803 participants
- CRA (phs000988) 205 participants
- FHS (phs000007) 8028 participants
- FHS (phs000974) 2230 participants

# PIC-SURE Authorized Access

**Authorized Access** allows users to query studies they are authorized to access and export selected variables to a workspace.

## Demo

<https://picsure.biodatacatalyst.nih.gov/picsureui/queryBuilder#>

The screenshot displays the BioData CATALYST interface, powered by PIC-SURE. The top navigation bar includes links for Authorized Access, Open Access, Data Access, User Profile, Help, and Log Out. The main section is titled "QUERY BUILDER" and contains three query entries, each with "back", "delete", and "edit" buttons:

- Query 1: DCC\_Harmonized\_data\_set, Body height at baseline., Any Value
- Query 2: DCC\_Harmonized\_data\_set, Subject sex as recorded by the study., Restrict By Value Female
- Query 3: DCC\_Harmonized\_data\_set, age at measurement of height\_baseline\_1, Greater than or equal to 18

On the right side, a large blue box displays "188601 Total Participants". Below this, a tree view shows a list of studies and their associated concepts:

- Atherosclerosis Risk in Communities (ARIC) Cohort ( phs000280 ) (14061 concepts)
- CATHeterization GENetics (CATHGEN) ( phs000703 ) (6 concepts)
- COVID19-ORCHID ( phs002299 ) (754 concepts)
- Cardiovascular Health Study (CHS) Cohort: an NHLBI-funded observational study of
- Cooperative Study of Sickle Cell Disease (CSSCD) ( phs002362 ) (5730 concepts)
- Coronary Artery Risk Development in Young Adults (CARDIA) ( phs000285 ) (6605 concepts)
- DCC Harmonized data set (80 concepts)
- 06 - Lipids (12 concepts)
- 02 - Atherosclerosis (12 concepts)
- 04 - Blood cell count (30 concepts)
- 01 - Demographics (6 concepts)
- 07 - Venous Thromboembolism Event (4 concepts)
- 05 - Blood pressure (6 concepts)
- 03 - Baseline common covariates (10 concepts)

# PIC-SURE Application Programming Interface (API)

PIC-SURE API allows researchers to use python or R to search and query at the variable and variant level and export data into a workspace.

Examples available on public GitHub repository  
(<https://github.com/hms-dbmi/Access-to-Data-using-PIC-SURE-API>)

## PIC-SURE API use-case: quick analysis on COPDGene data

This is a tutorial notebook aimed to get the user quickly up and running with the python PIC-SURE API. It covers the main functionalities of the API.

## PIC-SURE python API

### What is PIC-SURE?

As part of the BioData Catalyst initiative, the Patient Information Commons Standard Unification of Research Elements (PIC-SURE) platform has been integrating clinical and genomic datasets from multiple TOPMed and TOPMed related studies funded by the National Heart Lung and Blood Institute (NHLBI).

Original data exposed through the PIC-SURE API encompasses a large heterogeneity of data organization underneath. PIC-SURE hides this complexity and exposes the different study datasets in a single tabular format. By simplifying the process of data extraction, it allows investigators to focus on downstream analysis and to facilitate reproducible science.

## Connecting to a PIC-SURE resource

The following is required to get access to data through the PIC-SURE API:

- a network URL
- a resource id, and
- a user-specific security token.

If you have not already retrieved your user-specific token, please refer to the "Get your security token" section of the [README.md](#) file.

```
In [ ]: PICSURE_network_URL = "https://picsure.biodatacatalyst.nih.gov/picsure"
resource_id = "02e23f52-f354-4e8b-992c-d37c8b9ba140"
token_file = "token.txt"
```

```
In [ ]: with open(token_file, "r") as f:
my_token = f.read()
```

```
In [ ]: client = PicSureClient.Client()
connection = client.connect(PICSURE_network_URL, my_token, True)
adapter = PicSureBdcAdapter.Adapter(connection)
resource = adapter.useResource(resource_id)
```

Two objects are created here: a `connection` and a `resource` object.

Since we will only be using a single resource, the `resource` object is actually the only one we will need to proceed with data analysis hereafter.

It is connected to the specific data source ID we specified and enables us to query and retrieve data from this database.

# PIC-SURE API

## BioData Catalyst Powered by Terra

The screenshot shows the BioData Catalyst workspace interface. The top navigation bar includes the NIH logo, 'BioData CATALYST Powered by Terra', and 'WORKSPACES'. The workspace name is 'biodata-catalyst/BioD...'. The main content area is divided into two columns. The left column contains 'ABOUT THE WORKSPACE' and 'BioData Catalyst Python PIC-SURE API examples'. The right column contains 'WORKSPACE INFORMATION' with a table of metadata and 'OWNERS' with a list of email addresses. Below this is a 'TAGS' section with a dropdown menu and a 'Google Bucket' section with a text input and a button to 'Open in browser'.

**ABOUT THE WORKSPACE**

### BioData Catalyst Python PIC-SURE API examples

This workspace contains Jupyter Notebook examples of PIC-SURE API use cases, using BioData Catalyst studies. PIC-SURE API is available in two languages: R and python. This workspace features the python PIC-SURE API example notebooks and requires python 3.6 or later.

### PIC-SURE API Overview

The main goal of the PICSURE API is to provide a simple and reliable way to work with restricted-access data from TOPMed and TOPMed related studies that are part of BioData Catalyst. Each individual study is accessible in a unique, easy to use, tabular format directly in an R or python environment. The API allows also to query studies subset, based on patients matching specified criteria, as well as to retrieve a cohort that has been created using the [PIC-SURE interface](#). Finally, 43 specific phenotype variables that have been harmonized across multiple TOPMed studies are also accessible directly through the PIC-SURE API.

### Workspace information

- Requirement : python 3.6 or higher. To select the appropriate runtime environment for your Terra Workspace, click on the gear wheel beside "Cloud Environment" in the top right corner, and under Application Configuration select "Default: (GATK 4.1.4.1, Python 3.7.10, R 4.0.5)" or another appropriate configuration.
- Notebooks update information: the central repository for these notebooks is available on the [Access to Data using PIC-SURE API GitHub](#). Currently under active development, the repository is updated on a regular basis. Although the Terra public Workspace will be kept up-to-date as much as possible, there might be a difference between the version of the notebook you're using and the most recent one. So if you run into an unexpected issue when running one of these example notebooks, it may be worth checking for a potential more up-to-date version available on GitHub.

WORKSPACE INFORMATION	
CREATION DATE	4/9/2020
LAST UPDATED	9/22/2021
SUBMISSIONS	0
ACCESS LEVEL	Owner
EST. MONTHLY \$0.00	GOOGLE PROJECT ID biodata-catal...

**OWNERS**

simran\_makwana@hms.harvard.edu  
mbaumann@broadinstitute.org  
cartik.saravani@gmail.com  
emily\_hughes@hms.harvard.edu  
schaalva@broadinstitute.org  
esheets@ucsc.edu  
arnaud.serretarmande@gmail.com  
jmcampen@gmail.com  
avillach@gmail.com

**TAGS**

Add a tag

No tags yet

**Google Bucket**

Name: fc-617b067a-8e41-481d-a817...  
Location: US (multi-region)  
[Open in browser](#)

## BioData Catalyst Powered by Seven Bridges

The screenshot shows the BioData Catalyst workspace interface powered by Seven Bridges. The top navigation bar includes the NIH logo, 'BioData CATALYST Powered by Seven Bridges', and tabs for 'Projects', 'Data', 'Public Gallery', 'Public projects', 'Developer', and a user profile 'ernhughes1'. The main content area is divided into two columns. The left column contains 'DESCRIPTION' and 'Important notes'. The right column contains 'ANALYSES' with a search bar and a list of saved notebooks.

**DESCRIPTION**

This project contains JupyterLab and RStudio example notebooks for accessing PIC-SURE API. They can be located in Interactive Analysis > Data Cruncher. You can access them quickly by clicking on one of the links below:

- PIC-SURE JupyterLab examples
- PIC-SURE RStudio examples

Examples are provided by Dr. Paul Avillach's team at Harvard Medical School Department of Biomedical Informatics and are reflecting their [GitHub repository](#). The files in this project are kept up to date with the contents of the PIC-SURE API repository.

### Important notes

- If you would like to work with the PIC-SURE public project, make a copy of the project by selecting the "i" next to the project name. Select to copy the project. This will bring up the project creation menu. The network access will be set to "Block network access" by default, however you will need to change the setting to "Allow network access" in order to use the PIC-SURE API from the platform. If you have any questions, please contact [support@sevenbridges.com](mailto:support@sevenbridges.com).
- In order to use these notebooks, you will need to provide your PIC-SURE security token in the API request. To keep your security token private, it is best to work with this notebook in a project where you are the sole member. If you run this notebook in a project with collaborators, the token.txt file would be visible to other members of the project.

**ANALYSES**

Search

Tasks Data Cruncher

**SAVED** PIC-SURE JupyterLab examples  
Created by biodatacatalyst - Sept. 3, 2021 10:20

**SAVED** PIC-SURE RStudio examples  
Created by biodatacatalyst - Sept. 3, 2021 10:18

# Export into workspace

**Dataset ID** can be used to export selected data into a workspace. This data is saved as a dataframe, which can then be used for further analysis.

**Brief demo:** Export data into Seven Bridges workspace

Export Data
Download
Copy query ID to clipboard
DataSetID: 5c6c6396-6b65-48fa-b7af-41d3bf9eedac Status: AVAILABLE

# Questions?

**Next up:** Bring Your Own Data

# Bring Your Own Data

Tony Patelunas, Program Manager at Seven Bridges



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Bring-Your-Own Data

- To support **flexibility and analysis**, we allow researchers to bring their own data and workflows into the ecosystem.
- 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 IRB policies and guidelines.

Web resource: [Bring Your Own Data](#)



# Seven Bridges workspace environment

Private, secure workspaces with the option to collaborate

Set up analyses with visual user interface or API

Jupyterlab Notebooks and RStudio

Compute on AWS or Google

Hundreds of hosted CWL pipelines

The screenshot displays the Seven Bridges workspace environment interface for a project named "Alison\_test\_GWAS". The top navigation bar includes the NIH BioData CATALYST logo, a "Projects" dropdown menu, and various other navigation options like "Data", "Public Gallery", "Public projects", "Automations", "Developer", "Staff", and a user profile "alisonleaf". Below the navigation bar, the interface is divided into two main sections: "DESCRIPTION" and "MEMBERS".

**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 any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the platform are logged on the task page. This notepad is just for your own notes.

You can also [use markdown](#) here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the [Knowledge Center](#)

**MEMBERS**

**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**

**COMPLETED** GENESIS Null Model run - 01-17-20 17:44:24  
Submitted by alisonleaf · Jan. 17, 2020 12:51

**COMPLETED** GENESIS VCF to GDS run - 01-17-20 17:39:50  
Submitted by alisonleaf · Jan. 17, 2020 12:43

# Work alone in a private project

When you upload data, it is linked to a specific project.

If you are the only member of the project, then you are the only user who can access the uploaded data.

The screenshot displays the BioData CATALYST web interface. At the top, a navigation bar includes the NIH logo, 'BioData CATALYST Powered by Seven Bridges', and a series of dropdown menus: 'Projects', 'Data', 'Public Gallery', 'Public projects', 'Developer', and 'Staff'. A user profile 'alisonleaf' with a notification bell is on the right. Below the navigation bar, a secondary bar shows 'test project' with an information icon, and links for 'Interactive Analysis', 'Settings', and 'Notes'. The main content area is divided into two columns. The left column, titled 'DESCRIPTION', contains a 'Welcome to your new project!' message, an explanation of projects as building blocks, and a list of actions users can take within the project. The right column, titled 'MEMBERS', shows the user 'alisonleaf' as the 'OWNER' with permissions to 'Write, Copy, Execute, Admin'. It also features a message about teamwork and an 'Invite new members' button. At the bottom right, an 'ANALYSES' section includes a search bar and tabs for 'Tasks' and 'Data Cruncher'. A footer area at the very bottom contains a question mark icon and the text 'Your executions will appear here'.

**DESCRIPTION** Tags

**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 any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the platform are logged on the task page. This notepad is just for your own notes.

You can also [use markdown](#) here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the

**MEMBERS** Email notifications

**alisonleaf** OWNER  
Write, Copy, Execute, Admin

Don't work alone.  
The best research happens in teams.

[+ Invite new members](#)

Share your tools, data, and ideas with collaborators

**ANALYSES** Search

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

Your executions will appear here

# Collaborate in shared projects by adding members

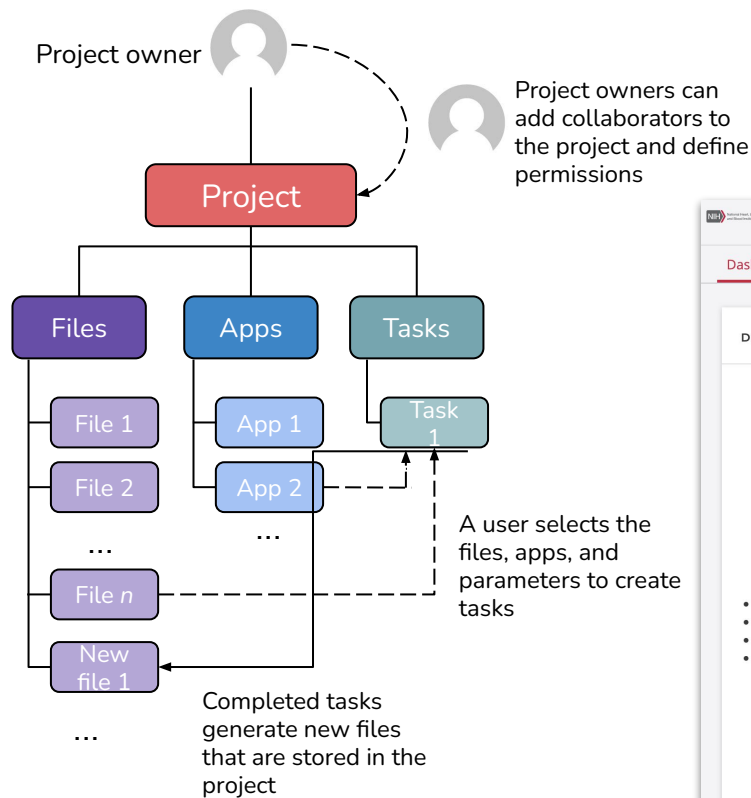
Project owner has administrative capabilities and can choose to collaborate with other platform users

Users can be added/deleted via GUI and public API

Set granular permissions to limit what project members can see/do

The screenshot displays the BioData CATALYST web application interface. The top navigation bar includes the NIH logo, 'BioData CATALYST Powered by Seven Bridges', and various menu items like 'Projects', 'Data', 'Public Gallery', 'Public projects', 'Automations', 'Developer', 'Staff', and a user profile 'alisonleaf'. Below this, a secondary navigation bar shows 'Dashboard', 'Files', 'Apps', and 'Tasks'. The main content area is titled 'Alison\_test\_GWAS' and includes a 'DESCRIPTION' section with a welcome message and instructions on how to use the project. A 'MEMBERS' section on the right lists four users: 'alisonleaf' (OWNER), 'dave', 'milan.domazet', and 'boris\_majic', each with their role and a 'Manage members' button. A 'Manage members' modal is open in the foreground, showing a list of current members (alisonleaf) and an 'Invite new members' section with a search bar and an 'Invite' button. The modal also displays a 'Permissions' section with a message: 'You cannot edit your own permissions.'

# Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions

Dashboard Files Apps Tasks

Alison\_test\_GWAS

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 any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the platform are logged on the task page. This notepad is just for your own notes.

You can also use markdown here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the [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

Search

Tasks Data Cruncher

COMPLETED GENESIS Null Model run - 01-17-20 17:44:24  
Submitted by alisonleaf · Jan. 17, 2020 12:51

COMPLETED GENESIS VCF to GDS run - 01-17-20 17:39:50  
Submitted by alisonleaf · Jan. 17, 2020 12:43

# Conveniently bring in your own data

## Data Tools

Manage your data using any of the following tools to suit your various requirements

### Seven Bridges Command Line Interface (SB CLI)

Upload your data using our fast and secure upload client, taking advantage of parallelization where possible.  
[Learn more](#)

Download ▾

### Seven Bridges File System (SBFS) BETA

Mount your projects and use files locally or download the executable.  
[Learn more](#)

```
curl
https://igor.sbgenomics.com/downloads/sbfs/install.sh -sSf |
sudo sh
```

Download ▾

### Upload files via the API

Upload files using the Seven Bridges Python library.  
[Learn more](#)

```
files = [
    '/foo/bar/baz.bam'
    '/foo/bar/qux.fastq'
]
for file in files:
    api.files.upload(project=
        'my-project', path=file)
```



Drag & drop files from your computer or

[Browse files](#)

This upload method is primarily intended for small-scale uploads. To upload a [larger volume of files](#), please use our [Data Tools](#). [Learn more about uploading from your computer.](#)

### Import from an FTP or HTTP(S) server ⓘ

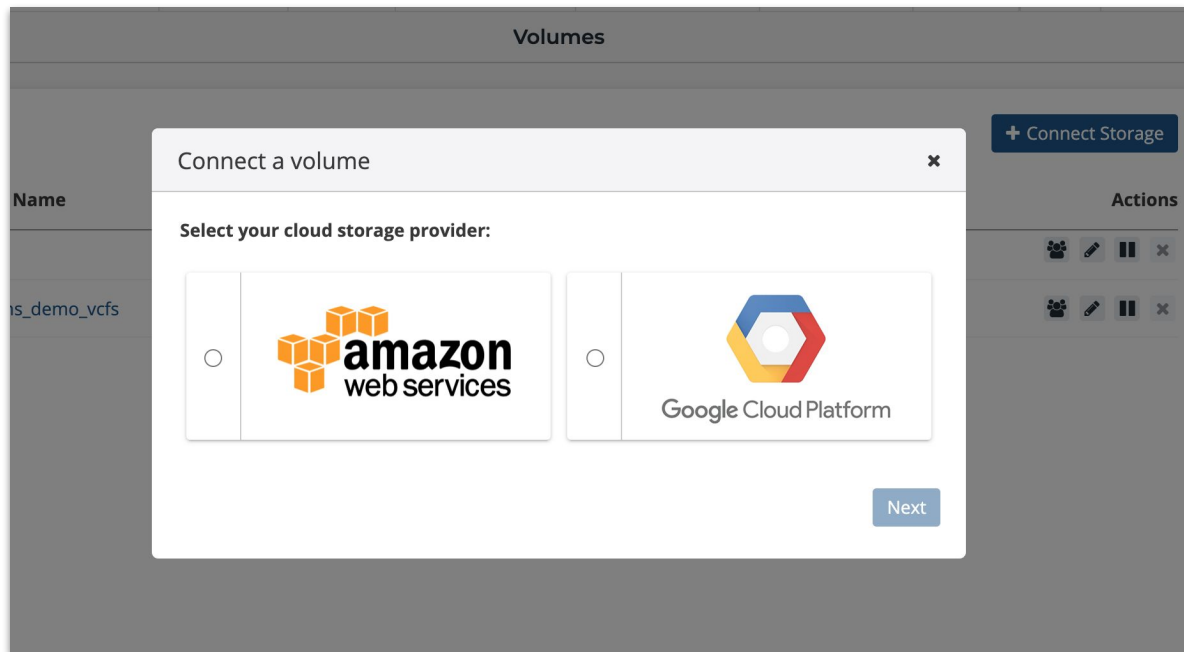
Paste the link of the file(s) you want to import

<ftp://john:mypass123@superseq.com/results/NA18507>

or [Browse file](#) on your computer containing the links

Import

# Connect private cloud storage directly to platform



Users retain full control over cloud storage access, management, and integrations.

# Organize and manage files within projects

Nested folder structure for organizing files

The screenshot displays the BioData CATALYST web application interface. The top navigation bar includes the NIH logo, 'BioData CATALYST Powered by Seven Bridges', and various dropdown menus for 'Projects', 'Data', 'Public Gallery', 'Public projects', 'Developer', 'Staff', and a user profile 'alisonleaf'. Below this, a secondary navigation bar shows 'Dashboard', 'Files' (selected), 'Apps', and 'Tasks'. A red 'CONTROLLED' badge is visible next to the project name 'Inflammation biomarkers'. The main content area is titled 'Files' and features a search bar, filter buttons for 'Type', 'Sample ID', 'Task ID', and 'Tags', and a 'Clear filters' button. A red arrow points to the 'New folder' button in the top right corner of the file management section. Below the filters is a table with columns for 'Name', 'Experimental strategy', 'Type', 'Size', and 'Gender'. The table lists four folders: 'Hispanic\_Community\_Health\_Study', 'Womens\_Health\_Initiative', 'Multi-Ethnic\_Study\_of\_Atherosclerosis', and 'CARDIA'. At the bottom, there is a 'Refresh' button and a status bar indicating 'Showing 1-3 of 3'.

<input type="checkbox"/>	Name	Experimental strategy	Type	Size	Gender
<input type="checkbox"/>	Hispanic_Community_Health_Study	-	-	-	-
<input type="checkbox"/>	Womens_Health_Initiative	-	-	-	-
<input type="checkbox"/>	Multi-Ethnic_Study_of_Atherosclerosis	-	-	-	-
<input type="checkbox"/>	CARDIA	-	-	-	-

# Questions?

**Next up:** Tools, Workflows, and Interactive Analysis



# Tools, Workflows, and Interactive Analysis

Tony Patelunas, Program Manager at Seven Bridges

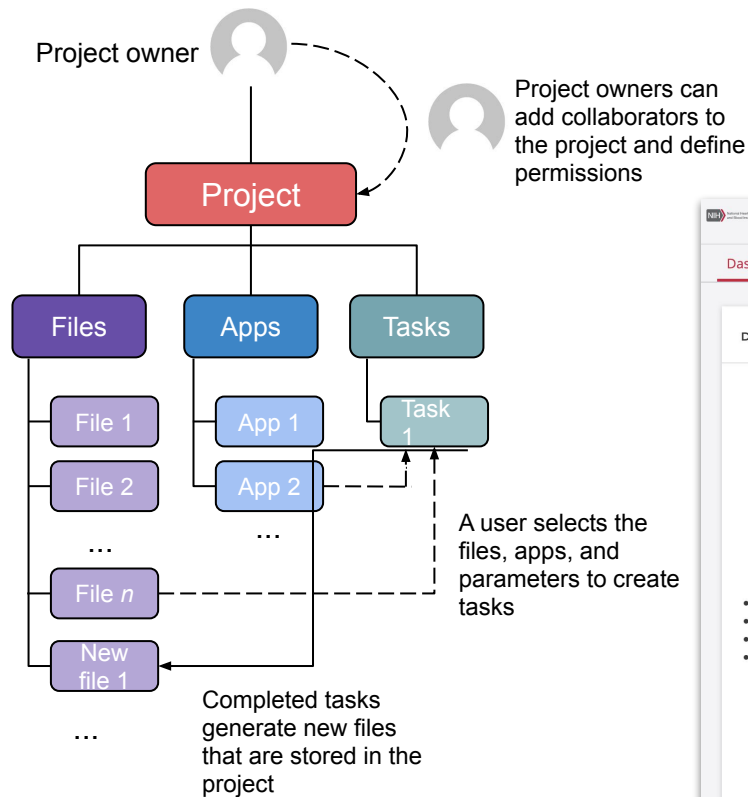


National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions

The screenshot shows the BioData CATALYST interface for a project named "Alison\_test\_GWAS". The top navigation bar includes "Projects", "Data", "Public Gallery", "Public projects", "Automations", "Developer", "Staff", and a user profile "alisonleaf". The main content area is divided into two sections: "DESCRIPTION" and "MEMBERS".

**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 any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the platform are logged on the task page. This notepad is just for your own notes.

You can also use markdown here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the [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**

Search

**Tasks** Data Cruncher

**COMPLETED** GENESIS Null Model run - 01-17-20 17:44:24  
Submitted by alisonleaf · Jan. 17, 2020 12:51

**COMPLETED** GENESIS VCF to GDS run - 01-17-20 17:39:50  
Submitted by alisonleaf · Jan. 17, 2020 12:43

# Interactive analysis



Studio®



**Fast prototyping** and implementation of custom tertiary analysis tools using interactive Java, Python and R in the JupyterLab environment as well as RStudio.

All project files available within JupyterLab, RStudio, and SAS. Over 50 instances to select from.

A screenshot of a 'Create new analysis' dialog box. The dialog has a title bar with a close button. Below the title bar are two tabs: 'Basic information' (active) and 'Compute requirements'. Under 'Basic information', there is a text input field for 'Analysis name' containing 'My first analysis'. Below that is a section for 'Environment' with three options: 'JupyterLab' (Web-based UI for Project Jupyter), 'RStudio' (IDE for R), and 'SAS Studio BETA' (Analytics and data management platform). The 'SAS Studio BETA' option is highlighted with a blue border. At the bottom, there is a section for 'Environment setup' with a dropdown menu currently showing 'SAS Data Science'. At the bottom right are 'Previous' and 'Next' buttons.

# User friendly workflow editor enables reproducibility by default

Common Workflow Language enables **portability, reproducibility, and scalability**

Use or combine 600+ optimized tools and workflows to construct your analysis

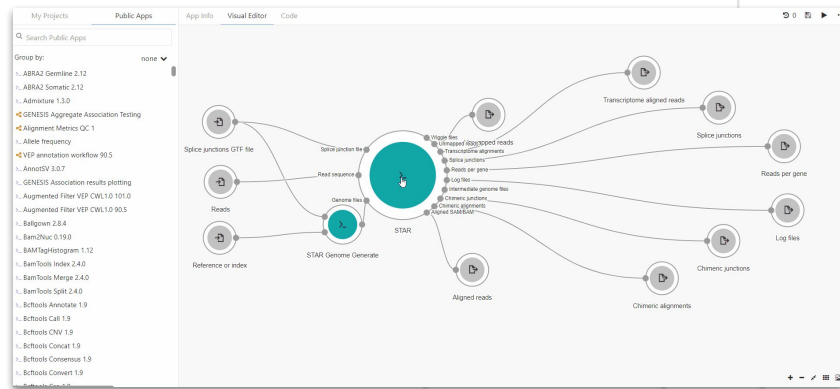
Seamlessly import workflows from external public repos (e.g. Dockstore)

Create your own tools with our CWL Tool Editor

Expose or lock parameters appropriately



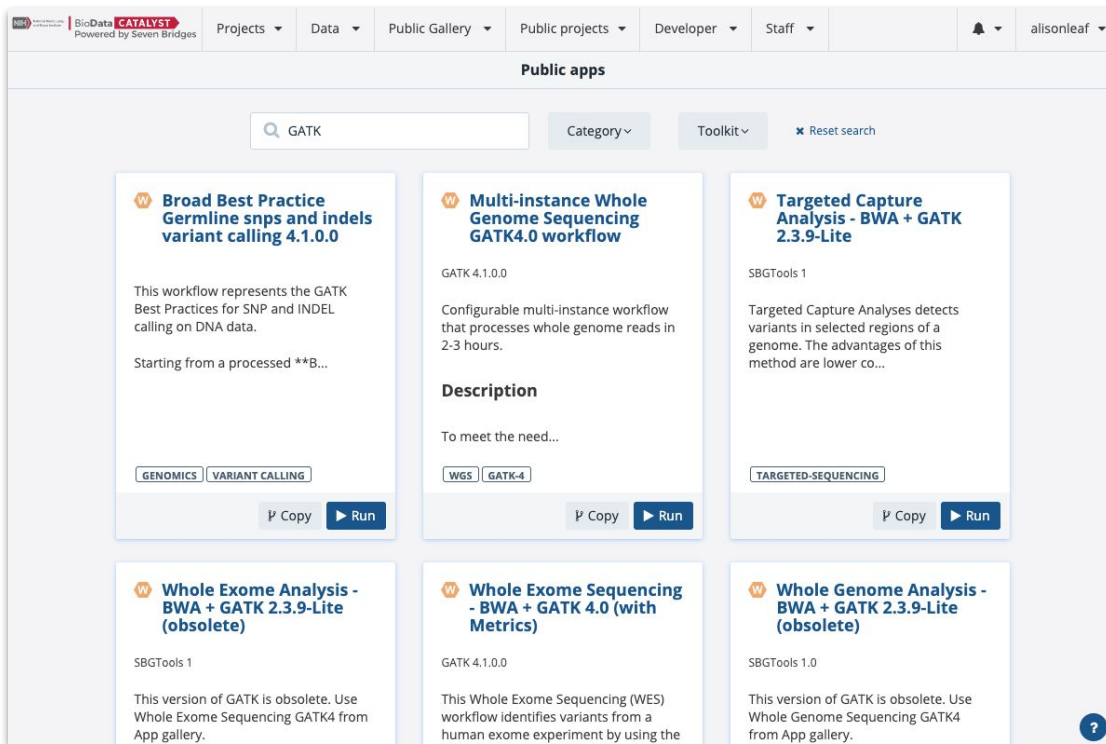
COMMON  
WORKFLOW  
LANGUAGE



# Find the tools you need in the Public Apps Gallery

A curated collection of **600+** bioinformatics tools & workflows:

- Optimized for speed & cost in the cloud
- Fully parameterized & customizable
- Accessible via the user interface & API
- Tool descriptions and helpful hints



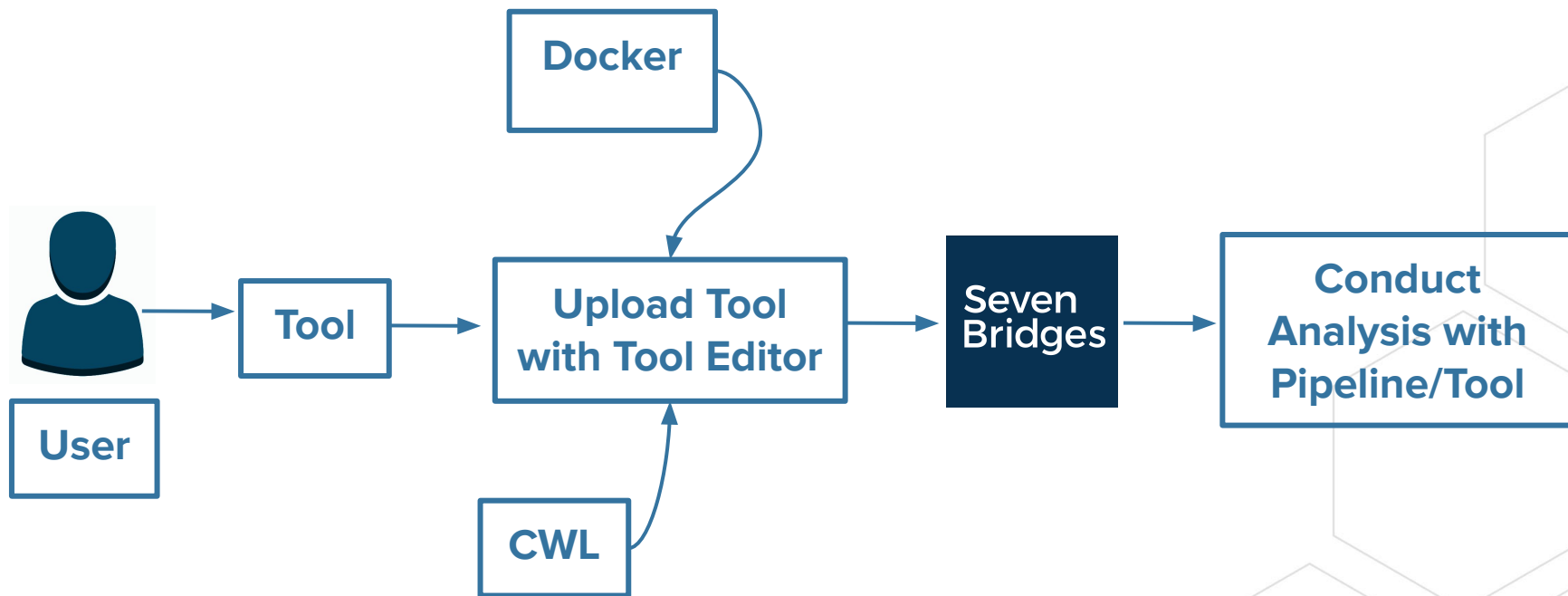
# Run association pipelines out of the box

- GENESIS
- Plink
- EPACTS
- STAAR (coming soon)

The screenshot displays the BioData CATALYST Public apps interface. At the top, there is a navigation bar with the NIH logo, the text 'BioData CATALYST Powered by Seven Bridges', and several dropdown menus: Projects, Data, Public Gallery, Public projects, Developer, and Staff. A user profile 'alisonleaf' is visible on the right. Below the navigation bar, the page is titled 'Public apps'. A search bar contains the text 'GENESIS'. To the right of the search bar are dropdown menus for 'Category' and 'Toolkit', and a 'Reset search' link. The main content area features six workflow cards arranged in a 2x3 grid. Each card has a 'W' icon, a title, a description, a list of supported data formats, and 'Copy' and 'Run' buttons.

Workflow Name	Description	Supported Data Formats	Buttons
GENESIS Aggregate Association Testing	Aggregate Association Testing workflow runs aggregate association tests, using Burden, SKAT [1], fastSKAT [2], SMM...	GWAS, CWL1.0	Copy, Run
GENESIS Null Model	Null Model workflow fits the regression or mixed effects model under the null hypothesis of no genotype effects. i...	GWAS, CWL1.0, GENOMICS	Copy, Run
GENESIS Single Variant Association Testing	Single Variant workflow runs single-variant association tests. It consists of several steps. Define Segments divid...	GWAS, CWL1.0, GENOMICS	Copy, Run
GENESIS Sliding Window Association Testing	Sliding Window Association Testing workflow runs sliding-window association tests. It can use Burden, SKAT [1], fa...		
GENESIS VCF to GDS	VCF to GDS workflow converts VCF or BCF files into Genomic Data Structure (GDS) format. GDS files are required by ...		
Fusion Transcript Detection - ChimeraScan	Fusion Transcript Detection - ChimeraScan 1.0 Fusion Transcript Detection - ChimeraScan detects and identifies fusion transcripts from paired-end RNA-Seq data using		

# Bringing custom tools to the platform



# Scale to 100's and 1000's of tasks in parallel using batching

Only one input per task can be selected for batching.

- Turn on the batching option on the draft task page, and select batch criteria: by File, or File metadata (e.g. Sample ID, Library ID).
- For each batch criteria match, a task will be created.

**BATCH 260 Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 03-22-19 13:2...** [Get support](#) [Discard](#) [Run](#)

Last update by shan.yeuz\_demo on Mar. 22, 2019 13:25  
App: Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) - Revision: 4

**Task Inputs** **Execution Settings**

**Inputs**

Batching ☒ On [Change selection](#)

Batch by: File

This will create one task for each selected item.

- 0.cram (1 item) x
- 1.cram (1 item) x
- 10.cram (1 item) x
- 100.cram (1 item) x
- 101.cram (1 item) x
- 102.cram (1 item) x
- 103.cram (1 item) x
- 104.cram (1 item) x
- 105.cram (1 item) x
- 106.cram (1 item) x
- 107.cram (1 item) x
- 108.cram (1 item) x
- 109.cram (1 item) x
- 11.cram (1 item) x

**App Settings**

[Edit parameters](#) [Show editable](#)

**GATK HaplotypeCaller (RGATK\_HaplotypeCaller)**

Memory Per Job

**GATK BaseRecalibrator (RGATK\_BaseRecalibrator)**

Intervals String

**SAMtools Index (IFSAMtools\_Index)**

Number of threads

**Picard MarkDuplicates (IPicard\_MarkDuplicates)**

Memory per job

**BWA MEM Bundle 0.7.17**

(BWA\_MEM\_Bundle\_0\_7\_17)

**Outputs**

BAM

Indexed CRAM

Realigned CRAM md5sum

VCF

VCF md5sum

gVCF

gVCF md5sum

metrics

multiqc\_report

**BATCH 260 Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04** [Get support](#) [Edit and rerun](#)

Executed on Nov. 29, 2018 03:26 by nevennameu Batch by: File  
Spot Instances: On ☐ Memorization: Off ☐ Price: \$2392.30 ☐

App: Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) - Revision: 2

Search task names  Status: All

Task Name	Submitted by	Submitted on	App	Duration	Status	Actions
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 1.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	17 hours, 29 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 10.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 57 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 11.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 50 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 6.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	17 hours, 24 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 18.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	17 hours, 10 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 17.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	15 hours, 58 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 8.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 24 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 7.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 39 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 19.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 35 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 23.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 58 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 16.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 27 minutes	COMPLETED	<a href="#">C</a>
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 22.cram	nevennameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 57 minutes	COMPLETED	<a href="#">C</a>



# Detailed documentation and tutorials

## Comprehensive tips for reliable and efficient analysis set-up

BIODATA CATALYST POWERED BY SEVEN BRIDGES

### Objective

### Helpful terms to know

### User Accounts & Billing Groups

#### Further reading

### Tips for Running Tools/Workflows

#### Start with the descriptions

#### Test the workflow

#### Specify computational resources

#### Learn about Instance Profiles

#### Scale up with Batch Analysis

#### Parallelize with Scatter

#### Configuring default computational resources

### Further analysis and interpretation of your Results

#### Getting started

#### JupyterLab environment

#### Accessing the files

#### Saving the created files

## OBJECTIVE

We have prepared this guide to help you with your first set of projects on BioData Catalyst powered by Seven Bridges. Each section has specific examples and instructions to demonstrate how to accomplish each step. We also highlight potential stumbling blocks so you can avoid them as you get set up. If you need more information on a particular subject, our [Knowledge Center](#) has additional information on all of the platform features. Additionally, our [support team](#) is available 24/7 to help!

## HELPFUL TERMS TO KNOW

**Tool** refers to a stand-alone bioinformatics tool or its Common Workflow Language (CWL) wrapper that is created or already available on the platform.

**Workflow / Pipeline** (interchangeably used) – denotes a number of tools connected together in order to perform multiple analysis steps in one run.

**App** stands for a CWL wrapper of a tool or a workflow that is created or already available on the platform.

**Task** – represents an execution of a particular tool or workflow on the platform. Depending on what is being executed (tool or workflow), a single task can consist of only one tool execution (tool case) or multiple executions (one or more per each tool in the workflow).

**Job** – this refers to the “execution” part from the “Task” definition (see above). It represents a single run of a single tool found within

## Troubleshooting Failed Tasks

BIODATA CATALYST POWERED BY SEVEN BRIDGES

### Helpful terms to know

### Getting started

### Examples: Quick & Unambiguous

[Task 1: Docker image not found](#)

[Task 2: Insufficient disk space](#)

[Task 3: Scatter over a non-list input](#)

[Task 4: Automatic allocation of the required instance is not possible](#)

[Task 5: JavaScript evaluation error due to lack of metadata](#)

[Task 6: Invalid JavaScript indexing](#)

[Task 7: Insufficient memory for Java process](#)

### Examples: File compatibility challenges

[Task 8: STAR reports incompatible chromosome names](#)

[Task 9: RSEM reports incompatible chromosome names](#)

[Task 10: Incompatible alignment coordinates](#)

Examples: When error messages are not enough

[Task 11: Invalid command line](#)

Tasks and examples described in this guide are available as a public project on the Platform.

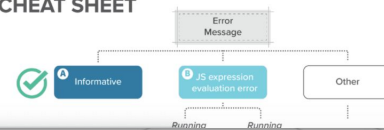
Often the first step to a user becoming comfortable using BioData Catalyst powered by Seven Bridges is their gaining confidence in resolving issues they encounter on their own. This confidence usually comes with experience – the experience with bioinformatics tools and Linux environment in general, but also the experience with the platform features.

However, one of the reasons for developing the platform in the first place is to enable an additional level of abstraction between the users and low-level command line work in the terminal. Even though there are a number of platform features that help with tracking down the issues, the less-experienced users can still face challenges with troubleshooting because the whole process might assume familiarity digging through the tool and system messages.

Fortunately, there is a set of steps that most often brings us to the solution. Based on internal knowledge and experience, the Seven Bridges team has come up with the **Troubleshooting Cheat Sheet** (Figure 1) which should help you navigate through the process of resolving the failed tasks.

## Troubleshooting CHEAT SHEET

SevenBridges



Visit the Knowledge Center

# Getting Help - Contacting Support from the platform

24/7 Help Desk can help you with failed analyses, login issues, or any other platform issue.

The screenshot displays the BioData CATALYST platform interface. At the top, there is a navigation bar with tabs for Projects, Data, Public Gallery, Public projects, Developer, and Staff. Below this is a sub-navigation bar with Dashboard, Files, Apps, and Tasks. The main content area is titled 'Genesis tutorial' and includes sections for DESCRIPTION and MEMBERS. A modal window titled 'Need help?' is open, providing links to documentation and project management options. A red circle highlights a 'Help and support' button in the bottom right corner of the interface, with a red arrow pointing to it.

**Need help?**  
Learn from the documentation below.

- Create a project
- Manage the project dashboard
- Add notes to your project
- Leave a project
- Delete a project

Not finding what you need? Visit our [Knowledge Center](#)

**Contact our support**

Describe your issue or share your ideas

Send

**Help and support**

# Questions?

**Next up:** Understanding, Estimating, and Managing Cloud Costs

# Understanding, Estimating, and Managing Cloud Costs

Tony Patelunas, Program Manager at Seven Bridges



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Agenda

- What are Cloud Costs?
- Estimating Cloud Costs
  - Categories of costs
  - Benchmarking
- Managing Cloud Costs
  - Billing groups
  - Task information
- Funding Cloud Costs
  - Apply for Pilot Credits
  - Grant writing

# What are cloud costs?

Three categories of costs:

- Storage
- Compute
- Egress

Web resource: [Cloud Costs and 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.

# Estimating Cloud Costs

# Estimating Cloud Costs

Researchers incur fees for:

- Data Storage
- Computing / Analysis
- Egress charges

## Estimate and Manage Your Cloud Costs



### Overview

In this tutorial, you will learn how cloud costs are incurred on BioData Catalyst Powered by Seven Bridges (the Platform), and the steps you should take to estimate your project cloud costs in advance of scaling up analyses.

Learning to estimate and manage your cloud costs will prepare you to effectively budget for your research projects. These estimates can be included in grant proposals, or be used to request cloud credits offered by the National Institutes of Health.

### Background

The Platform is a [multi-cloud](#) bioinformatics solution. This means that you can run compute jobs on regions of both Amazon Web Service (AWS) and Google Cloud Platform (GCP) (Figure 1). By running analyses on the cloud in the location where data is stored, it saves you time that would otherwise be spent copying large datasets. This multi-cloud functionality can also lead to cost savings, since data egress charges can be avoided. These concepts will be expanded upon throughout this tutorial.

New Platform users may be accustomed to working with an on-site HPC.

[View Cloud Cost Guide](#)



# Estimating Cloud Costs

## Data Storage

Charges are billed on all files in your workspace that belong to your project.

- **Includes**: All files you upload to BioData Catalyst and any results files generated by your workflows and analysis.
- **Does NOT include**: Controlled dataset files hosted by BioData Catalyst for general use.

Costs vary based on the amount of data you store, what type of disk or service you use for storing the data, and the service you select (AWS or GCP).

**Up-to-date information on storage rates:** Amazon S3 and Google Cloud

# Estimating Cloud Costs

## Computing / Analysis

Compute costs vary and depend on a range of factors:

- Platform and cloud infrastructure provider where an analysis is performed
- Your workspace & cloud instance settings
- Length of time to workflow completion

**Resources:** BioData Catalyst Powered by Terra and BioData Catalyst Powered by Seven Bridges

# Estimating Cloud Costs

## Egress Charges

Data uploaded or generated in your workspace is stored on a single cloud provider instance. If you move files you will be charged **Egress fees**. These fees will occur if you:

- Transfer files to another cloud provider, **OR**
- Download files to a local machine

Fees for data egress vary based on your service provider and what actions you take.

# Planning costs for GWAS pipelines

## GENESIS Benchmarking Guide

### Introduction

The objective of the GENESIS Benchmarking Guide is to instruct users on the drivers of cloud costs when running GENESIS workflows on the NHLBI BioData Catalyst Powered by Seven Bridges.

For all GENESIS workflows, the Seven Bridges team has performed comprehensive benchmarking analysis on Amazon Web Services scenarios:

- 2.5k samples (1000G data)
- 10k samples (TOPMed Freeze5)
- 36k samples (TOPMed Freeze5)
- 50k samples (TOPMed Freeze5)

The resulting execution times, costs found in the sections below. In the benchmarking results and some tips. Lastly, we included a Methods section for your reference.

View GENESIS Guide and Benchmarking

					AWS Instance					Google Instance					
Analysis	Samples	Variants	Relatedness matrix	Instance type	Parallel instances	Instance	CPU	RAM (GB)	Time	Cost	Instance	CPU	RAM (GB)	Time	Cost
Single test	2.5K		w/o	Spot	8	r4.8	1	2	1 h, 8 min	3\$	n1-standard-64	1	2	1h	7\$
Single test	2.5K		Dense	Spot	8	r4.8	1	2	1 h, 6 min	5\$	n1-standard-64	1	2	1h	7\$
Single test	10K		w/o	On dm	8	c5.18	1	2	50 min	10\$	n1-standard-4	1	2	1 h, 12 min	13\$
Single test	10K		Sparse	On dm	8	c5.18	1	2	58 min	11\$	n1-standard-4	1	2	1 h, 13 min	14\$
Single test	10K		Sparse	On dm	8	r4.8	1	2	1 h, 30 min	11\$	n1-standard-4	1	2	1 h, 13 min	14\$
Single test	10K		Dense	On dm	8	r5.4	1	8	3 h	24\$	n1-highmem-32	1	8	2 h, 20 min	30\$
Single test	36K		w/o	On dm	8	r5.4	1	5	3 h, 20 min	27\$	n1-standard-64	1	5	1 h, 30 min	35\$
Single test	36K		Sparse	On dm	8	r5.4	1	5	4 h	32\$	n1-highmem-16	1	5	4 h, 30 min	35\$
Single test	36K		Sparse	On dm	8	r5.12	1	5	1 h, 20 min	32\$	n1-standard-64	1	5	1 h, 30 min	35\$
Single test	36K		Dense	On dm	8	r5.12	1	50	1 d, 15 h	930\$	n1-highmem-96	1	50	1 d, 6 h	1,300\$
Single test	36K		Dense	On dm	8	r5.24	1	50	17 h	800\$					
Single test	50K		w/o	On dm	8	r5.12	1	8	2 h	44\$	n1-standard-96	1	8	2 h	73\$
Single test	50K		Sparse	On dm	8	r5.12	1	8	2 h	48\$	n1-standard-96	1	8	2 h	73\$
Single test	50K		Dense	On dm	8	r5.24	48	100	11 d	13,500\$	n1-highmem-96	16	100	6 d	6,600\$

# Managing Cloud Costs

# Tasks have detailed credit usage information

**COMPLETED** **GENESIS Single Variant Test w/ GDS Conversion and Null Model Fitting ...** [Get support](#) [View stats & logs](#) [Publish](#) [Edit and rerun](#)

Executed on Aug. 25, 2021 10:16 by dave

Spot Instances: **On** [?](#) | Memoization (WorkReuse): **Off** [?](#) | Price: **\$0.24** [?](#) | Duration: **11 minutes** [?](#)

App: GENESIS Single Variant Test w/ GDS Conversion and Null Model Fitting

**Instances:** \$0.21  
**Attached disks:** \$0.03  
**Data transfer:** \$0.00

**Inputs** [?](#)

**Phenotype file** [?](#) [📎](#)  
sample\_phenotype\_pcs.RData

**Relatedness matrix file** [?](#) [📎](#)  
kinship.RData

**Variants Files** [?](#) [📎](#)  
1KG\_phase3\_subset\_chr1.vcf.gz  
1KG\_phase3\_subset\_chr10.vcf.gz  
1KG\_phase3\_subset\_chr11.vcf.gz  
1KG\_phase3\_subset\_chr12.vcf.gz  
1KG\_phase3\_subset\_chr13.vcf.gz  
...and 17 more items

**App Settings** [?](#)

**GENESIS Null Model** (#null\_model)

- Covariates** [?](#)
  - sex
  - age
  - PC1
  - PC2
  - PC3
  - PC4
  - gaussian
  - height
  - demo\_height
  - FALSE
- Family [?](#)
- Outcome [?](#)
- Output prefix (Output prefix) [?](#)
- Two stage model [?](#)

**Outputs** [?](#)

**Association test plots** [?](#) [📎](#)

- demo\_height\_manh.png
- demo\_height\_qq.png

**Association test results** [?](#) [📎](#)

- demo\_height\_chr1.RData
- demo\_height\_chr2.RData
- demo\_height\_chr3.RData
- demo\_height\_chr4.RData
- demo\_height\_chr5.RData
- ...and 17 more items

**Null Model HTML Report** [?](#) [📎](#)

- demo\_height\_report.html

# Track costs on platform payments page

See cumulative costs for **Analysis** (Tasks and Data Cruncher) and **Storage**

**Billing Group settings: BDC Fellow - Einat Granot-Hershkov**

**Info**

**Organization**

Creator	alisonleaf
Primary contact	Seven Bridges
Address	One Broadway, 14th Fl. , Massachusetts, United States
Payment Method	CC - Last 4 Digits: **** * Expiration Date: <a href="#">Change Payment Method</a>

**Current usage** **\$ 309.42**

Analysis usage		Storage usage	
Analysis charges	<b>\$ 303.64</b>	Storage charges	<b>\$ 5.78</b>
Tasks	<b>\$ 285.24</b>	Active	<b>\$ 5.45</b>
Data Cruncher analyses	<b>\$ 18.41</b>	Downloaded	<b>\$ 0.33</b>
		Storage deduction	<b>\$ 0.00</b>
Additional charges	<b>\$ 0.00</b>	Credits	<b>\$ 0.00</b>

Instance limits  
Total number of instances that can be run in parallel

Current usage: 0 of 60 ?

**Alison Leaf**

- Account settings
- Payments**
- Email notifications

[Sign out](#)

# Funding Cloud Costs



# Try out the ecosystem 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](https://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

# Requesting grant funding for BioData Catalyst

- Understand your potential costs
  - Storage
  - Computation
- Use sample text
- Request Letter of Support from the BioData Catalyst Coordinating Center

## Writing BioData Catalyst into a Grant Proposal

Guidance on writing BioData Catalyst into a research proposal and the various costs you should budget for.

## Writing BioData Catalyst into your proposal's budget

NHLBI BioData Catalyst is a cloud-based ecosystem which seeks to empower researchers analyzing phenotypic and genotypic heart, lung, blood, and sleep data. Researchers on NHLBI BioData Catalyst have access to a number of controlled and open datasets, as well as the power to bring their own data to the ecosystem for analysis.

This document intends to serve as a resource for researchers writing NHLBI BioData Catalyst into grant proposals.

The BioData Catalyst ecosystem leverages two well-known cloud computing services, Google Cloud Platform (GCP) and Amazon Web Services (AWS), to perform computational analysis and store data. Users may scale their workloads up or down by toggling the virtual machine (VM) instance size and attached storage, as well as horizontally scale workloads by specifying a number of parallel instances. Increasing compute power, storage, and parallelization has an associated increase in cost, which is estimated for the researcher.

**View BioData Catalyst  
Grant Guide**

# Questions?

**Next up:** Interoperability with Gen3, Terra and Dockstore

# Interoperability with Gen3, Terra and Dockstore

Dr. Alisa Manning



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# BioData Catalyst is an ecosystem of platforms

User flows through the ecosystem are specialized to each user community.

## Explore Available Data

### BioData Catalyst Powered by Gen3

Gen3 is a software platform that allows partner organizations and grant approved researchers to search and access harmonized datasets. Users can search over project and study-specific genomic and phenotypic data and export selected cohorts to analytical workspaces in a scalable, reproducible, and secure manner.

[Launch](#)

| [Documentation](#) 

| [Learn](#)

### BioData Catalyst Powered by PIC-SURE

Explore available data through BioData Catalyst Powered by PIC-SURE with interactive search and visualizations for feasibility assessment. Use query results to create a cohort, with the ability to choose specific variables of interest to export into an analysis environment.

[Launch](#)

| [Documentation](#) 

| [Learn](#)

# BioData Catalyst is an ecosystem of platforms

User flows through the ecosystem are specialized to each user community.

## Analyze Data in Cloud-based Shared Workspaces

### BioData Catalyst Powered by Seven Bridges

Utilize collaborative workspaces for analyzing genomics data at scale. Access hosted datasets along with Common Workflow Language (CWL) and GENESIS R package pipelines for analysis. This platform also enables users to bring their own data for analysis and work in RStudio and Jupyterlab Notebooks for interactive analysis.

[Launch](#)

| [Documentation](#)

| [Learn](#)

### BioData Catalyst Powered by Terra

Share and compute across large genomic and genomic-related datasets. Terra offers a stand-alone computational workspace model that provides a secure collaborative place to organize data, run and monitor Workflow Description Language (WDL) analysis pipelines, and perform interactive analysis using applications such as RStudio, Jupyter Notebooks, and the Hail GWAS tool.

[Launch](#)

| [Documentation](#)

| [Learn](#)

# BioData Catalyst is an ecosystem of platforms

User flows through the ecosystem are specialized to each user community.

Use Community Tools on Controlled-access Datasets   Imputation Server

## Dockstore

Search from a catalog of high-quality Docker-based workflows that export to Terra or Seven Bridges. Explore organization pages to find collections of workflows from labs, institutions, and consortiums or create a page to share your work with the wider bioinformatics community.

[Launch](#)

[Documentation](#) 

[Learn](#)

## Access the Imputation Server

### Imputation Server developed by the University of Michigan

Upload your own phased or unphased GWAS genotypes to the server and receive phased and imputed genomes in return. The server offers imputation from various reference panels including the TOPMed reference panel.

[Launch](#)

[Documentation](#) 



# Introduction to *BioData Catalyst Powered by Gen3*

Gen3 is a data platform for building data commons and data ecosystems.

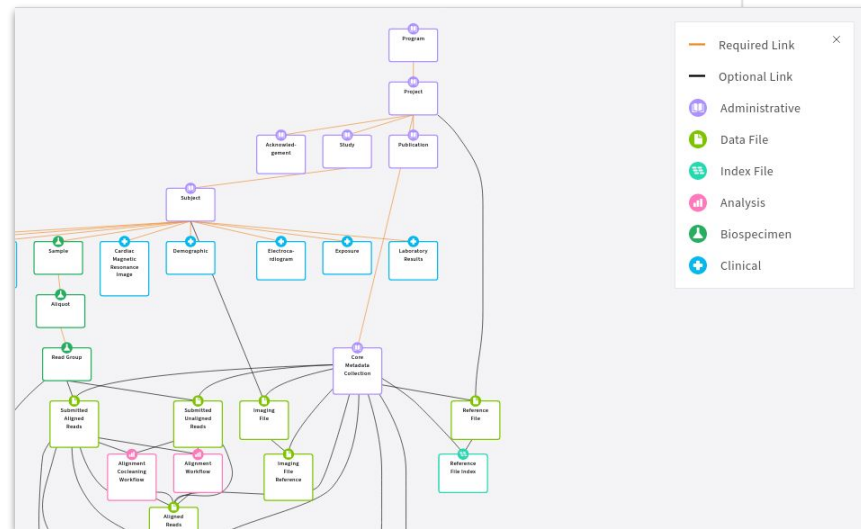
- creates pointers to data files and links them to metadata (**file information**) .

## Indexing data files

- Globally Unique IDs (GUIDs)
- Creates a pointer for the data file

## Graph Model

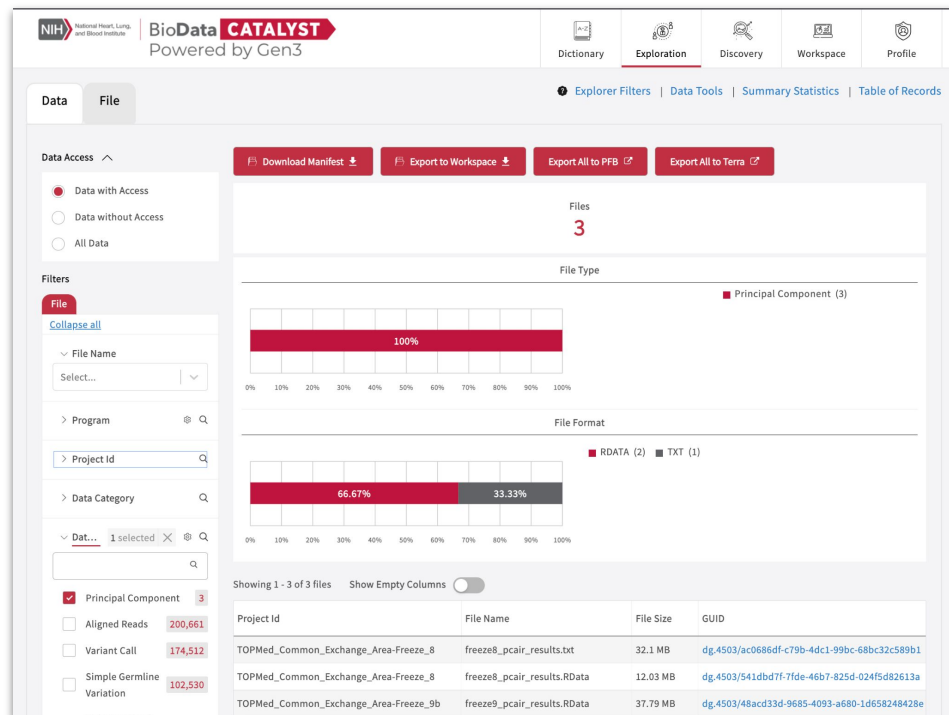
- The ability to relate metadata (**file information**) via nodes and edges
- Allows for linkage between data files and clinical information



# Introduction to *BioData Catalyst Powered by Gen3*

## Exploration

- Displays metadata (*file information*) found within the graph model
- Search and filter functionality
- Interoperability feature:  
Export the selected files to BioData Catalyst Powered by Terra



# About *BioData Catalyst Powered by Terra*

Terra is a scalable platform for biomedical research

- **Access Data:** Browse closed and open access datasets
- **Collaborate:** Organize your data and tools in a workspace. Work with your project team in one place
- **Workflows:** Utilize batch analysis workflows from others (Dockstore, Galaxy) or write your own
- **Interactive Analysis:** Interact with your data in your workspace with Jupyter Notebooks, Rstudio, the command line, or bring your own software via Docker containers

# Terra differentiators

Workflow Language	Workflow Description Language (WDL)
Cloud Provider	Google Cloud Platform, Azure ( <i>coming</i> )
Applications	<ul style="list-style-type: none"><li>• Preloaded applications and options to bring-your-own through a user-friendly interface.</li><li>• Galaxy, IGV, Seqr</li></ul>
Interactive Analysis Features	<ul style="list-style-type: none"><li>• Highly customizable machines with persistent disks set up to save your work</li><li>• Bioconductor, Hail, GATK and other popular bioinformatics tools preloaded.</li><li>• "Best practices" workspaces from the tool developers</li></ul>



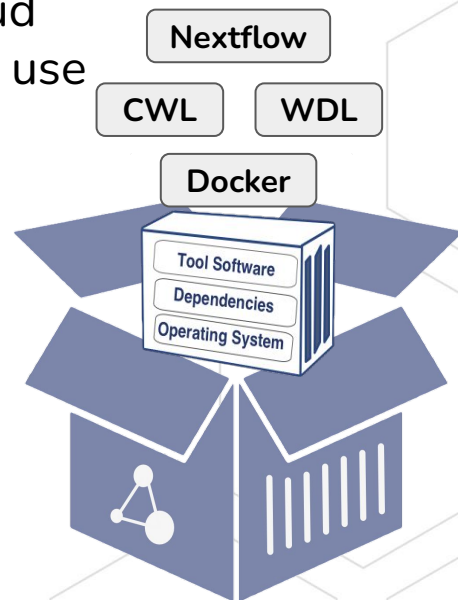
# Introduction to Dockstore

“an app store for bioinformatics”

Users can launch workflows from Dockstore directly into cloud workspaces like Seven Bridges or Terra or download them to use locally.

## Advantages

- Increases reproducibility of computational analysis using combination of containers and workflow languages
- Increases the transparency of analysis methods
- Allows others to verify results and apply existing methods into their own research



[dockstore.org](https://dockstore.org)

# Publish your workflows on Dockstore!

- Sharing your workflows on Dockstore makes them more **accessible** and your research methods **transparent** and **reproducible**.
- Dockstore integrates with GitHub and automatically updates your Dockstore entry every time an update is made to the GitHub repository.
- Get started by following the BioData Catalyst [Bring Your Own Tool documentation](#).

**Dockstore** Search Organizations About Docs Forum Login Register

Organizations / Broad Institute of MIT and Harvard / Viral Genomics

**BROAD INSTITUTE** Collection: **Viral Genomics**  
Viral Genomic Workflows, see [viral-pipelines.readthedocs.io](#) for details.

**Workflows & Tools**

- [github.com/broadinstitute/viral-pipelines/assemble\\_refased](#)  
Last updated Aug 18, 2021 assembly WDL View
- [github.com/broadinstitute/viral-pipelines/fetch\\_sra\\_to\\_bam](#)  
Last updated Aug 18, 2021 ncbi WDL View
- [github.com/broadinstitute/viral-pipelines/genbank](#)  
Last updated Aug 18, 2021 ncbi WDL View
- [github.com/broadinstitute/viral-pipelines/fetch\\_annotations](#)  
Last updated Aug 18, 2021 ncbi WDL View

**About the Collection**

**Viral NGS Workflows**

The workflows in this collection provide the ability for users to perform viral genomic data analysis. These workflows enable users to perform assembly, QC, kraken metagenomics and aggregate statistics. Additionally, we've provided workflows for users to go from raw reads (uBAM), through to producing a phylogenetic tree.

These workflows allow users to work with either their own and/or public data, such as from NCBI SRA and GenBank. This collection contains a workflow that allows users to pull data from SRA (via SRA accession #), and a workflow to prepare their data files for bulk upload to GenBank.

Detailed documentation is available at [ReadTheDocs](#).

**Overview of analytical workflows available**

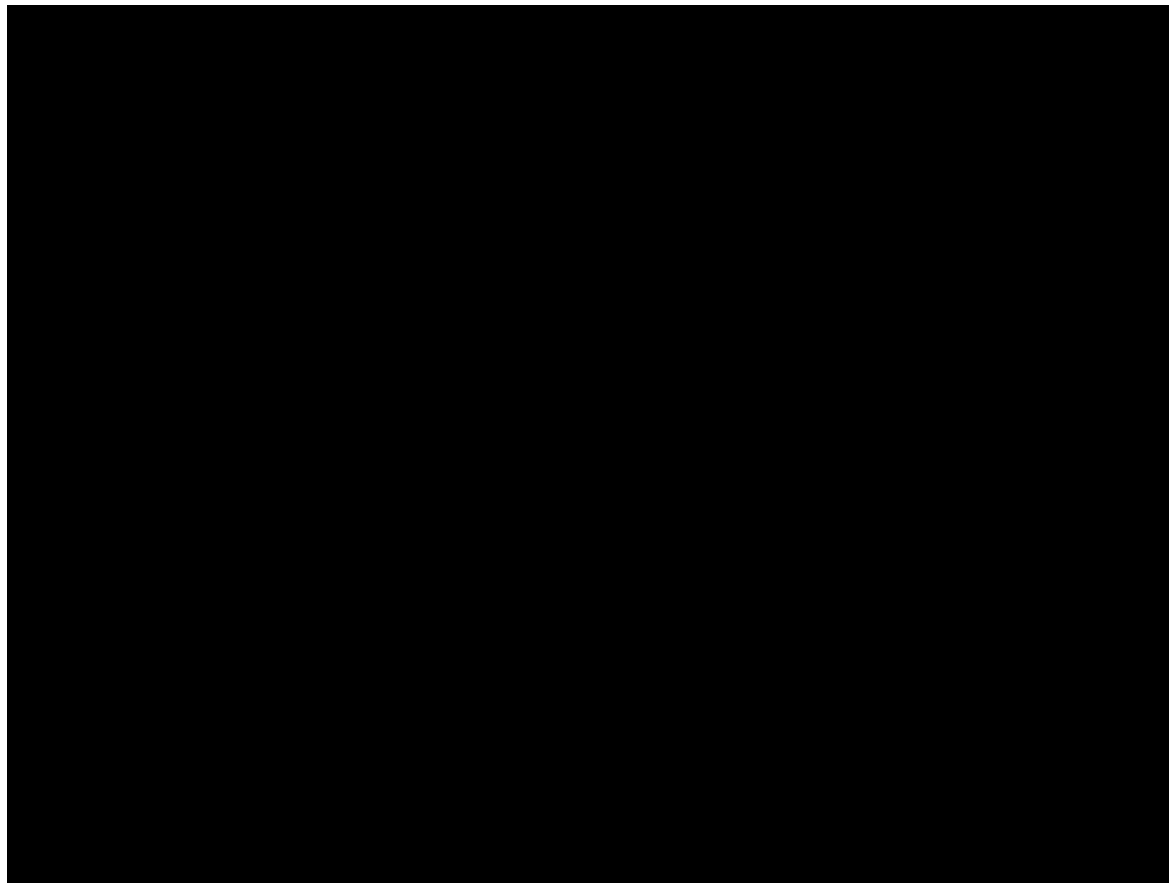
```
graph TD
    FASTQ --> ImportFASTQ[Import FASTQ Data]
    SRAID[SRA ID] --> ImportSRA[Import SRA Data]
    ImportFASTQ --> uBAM
    ImportSRA --> uBAM
    uBAM --> KrakenClassifier[Kraken Classifier]
    uBAM --> ViralAssembly[Viral Assembly]
    KrakenClassifier --> FASTA
    ViralAssembly --> FASTA
    FASTA --> BuildAugurTree[Build Augur Tree]
    BuildAugurTree --> Nextstrain
    uBAM -.-> MapRef[Map to reference sequence]
    MapRef -.-> Scaffold[Sc scaffolding]
    Scaffold -.-> RemoveGap[Removing to remove gap sequences & low quality reads]
    RemoveGap -.-> Refine[Refine assembly]
```

**Tutorials**

The following Terra workspaces outline in detail the steps to set up and execute the listed workflows and they additionally contain example inputs and references.

[COVID-19 Broad Viral NGS](#)  
[COVID-19](#)

[github.com/broadinstitute/viral-pipelines/beast\\_gpu](#)



Import **xvcfmerge** workflow from Dockstore



# Additional Information

## Useful links

- [Gen3 website](#)
- [BioData Catalyst documentation: Discovering Data using Gen3](#)

## Accessing genomic data via the GA4GH DRS standard

- [Terra documentation: Data access with the GA4GH Data Repository Service \(DRS\)](#)

## Workspace tutorial on Gen3 data

- [Terra documentation: Working with Workspaces](#)
- [BioData Catalyst documentation: Genome Wide Association Study with 1000 Genomes Data Tutorial](#)

# Questions?

**Next up:** Researcher Presentation and Q&A with Ravi Mathur

Researcher Presentation and Q&A:

# Use of TOPMed WGS as Public Controls on BioData Catalyst

Ravi Mathur, Statistician, RTI International



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Overview

- Background
- Results
  - Identification of Public Controls
  - Testing for Technology Bias
  - Assessing Type 1 Error
  - Validating True Positives
- Method: GAWMerge Protocol
- Conclusions

# Background

- Matching population/public controls to case only cohorts is a standard epidemiologic approach
  - Control confounding factors and cut cost
  - 217 Studies in dbGaP are case-only datasets from over 136K Samples
- Successful applications in GWAS<sup>1,2</sup> requires
  - Consistency in ancestry
  - Substantial overlap in the genotyped variants between datasets
- NHLBI TOPMed cohorts with whole-genome sequencing (WGS) data could address both issues

# NHLBI Trans-Omics for Precision Medicine (TOPMed)

- Sponsored by NHLBI and part of the Precision Medicine Initiative
- Provides integration of WGS and other omics (e.g., metabolic profiles, epigenomics, protein, and RNA expression patterns) for a diversity of heart, lung, blood, and sleep (HLBS) disorders.
- Data is available via dbGaP and hosted on BioData Catalyst

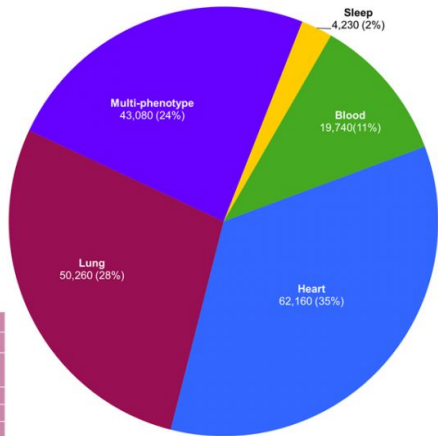


# TOPMed Diversity

- Over 180K Participants
- ~60% of the Cohort is of non-European Ancestry

## Phenotype Focus

Phases 1-7 (~180K Participants)



### Blood:

Hemophilia  
Sickle Cell Disease  
Platelets  
Lipids  
Blood Cancers

### Heart:

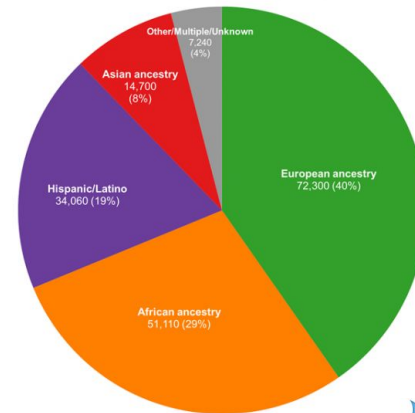
Hypertension  
Myocardial Infarction  
Coronary Artery Disease  
Stroke  
Small Vessel Disease  
Venous Thromboembolism  
Congenital Heart Disease  
Atrial Fibrillation  
Coronary Artery Calcification  
Adiposity  
Congestive Heart Failure

### Lung:

Asthma  
Chronic Obstructive Pulmonary Disease  
Idiopathic Pulmonary Fibrosis  
Sarcoidosis  
Interstitial Lung Disease

## Ancestry & Ethnicity

Phases 1-7 (~180K Participants)



# Questions to Answer

1. How to select WGS cohorts for the available array data?
2. Are Array genotyping data comparable with WGS data?
3. Will the population control method produce false-positive signals?
4. Can we reproduce GWAS hits by combining array genotyping data and WGS data?



# Array Data Used

- **COPDGene**

- One of the largest studies to investigate the genetic factors of Chronic Obstructive Pulmonary Disease (COPD)
- 9,994 total samples
- African American and European American ancestries
- Available via dbGAP
  - Imported into BDC via BYOD

- **COGEND**

- Genetic study of nicotine dependence initiated in 2001
- Nicotine dependent cases and non-dependent smoking controls were identified and recruited from Detroit and St. Louis.
- Over 2,900 donated blood samples for genetic studies using the HumanOmni2.5 array with about 2.5M SNPs
- AA and EA ancestries

# Step 1: Identifying WGS Data

- TOPMed Harmonized Variables
- Extensive effort by the TOPMed DCC to harmonize 63 phenotypes across 17 TOPMed Studies
  - Includes variables for atherosclerosis, sleep, inflammation, lipids, VTE, blood cell count, blood pressure, demographic, and common covariates
- Data is hosted on BioData Catalyst

➤ [Am J Epidemiol. 2021 Oct 1;190\(10\):1977-1992. doi: 10.1093/aje/kwab115.](#)

## A System for Phenotype Harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program

Adrienne M Stilp, Leslie S Emery, Jai G Broome, Erin J Buth, Alyna T Khan, Cecelia A Laurie, Fei Fei Wang, Quenna Wong, Dongquan Chen, Catherine M D'Augustine, Nancy L Heard-Costa, Chancellor R Hohensee, William Craig Johnson, Lucia D Juarez, Jingmin Liu, Karen M Mutalik, Laura M Raffield, Kerri L Wiggins, Paul S de Vries, Tanika N Kelly, Charles Kooperberg, Pradeep Natarajan, Gina M Peloso, Patricia A Peyser, Alex P Reiner, Donna K Arnett, Stella Aslibekyan, Kathleen C Barnes, Lawrence F Bielak, Joshua C Bis, Brian E Cade, Ming-Huei Chen, Adolfo Correa, L Adrienne Cupples, Mariza de Andrade, Patrick T Ellinor, Myriam Fornage, Nora Franceschini, Weiniu Gan, Santhi K Ganesh, Jan Graffelman, Megan L Grove, Xiuqing Guo, Nicola L Hawley, Wan-Ling Hsu, Rebecca D Jackson, Cashell E Jaquish, Andrew D Johnson, Sharon L R Kardia, Shannon Kelly, Jiwon Lee, Rasika A Mathias, Stephen T McGarvey, Braxton D Mitchell, May E Montasser, Alanna C Morrison, Kari E North, Seyed Mehdi Nouraie, Elizabeth C Oelsner, Nathan Pankratz, Stephen S Rich, Jerome I Rotter, Jennifer A Smith, Kent D Taylor, Ramachandran S Vasan, Daniel E Weeks, Scott T Weiss, Carla G Wilson, Lisa R Yanek, Bruce M Psaty, Susan R Heckbert, Cathy C Laurie

PMID: 33861317 PMCID: [PMC8485147](#) DOI: [10.1093/aje/kwab115](#)

# Step 1: Identify Public Control Cohorts via Gen3

Criteria: COPDGene Cohort, EA and AA race, current or former cigarette smoker

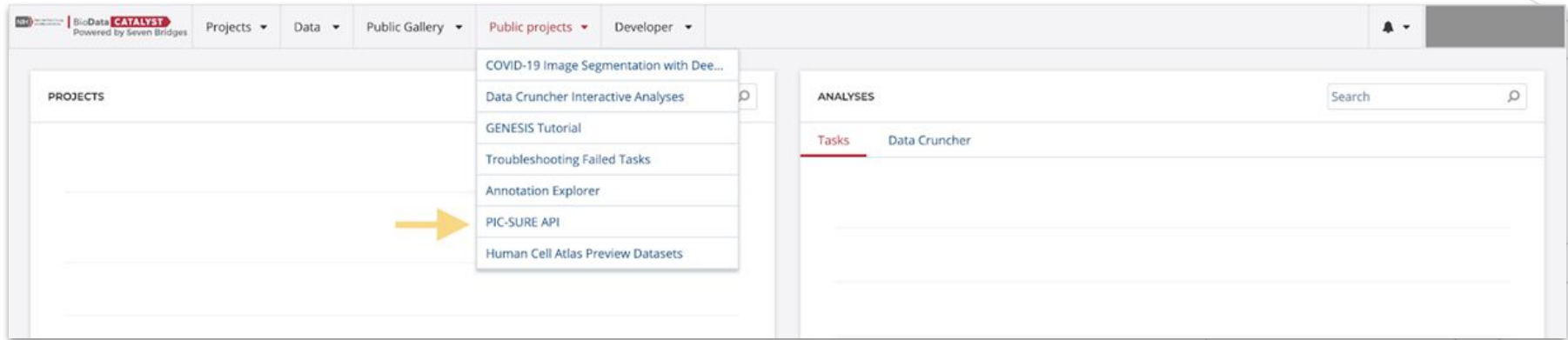
The screenshot displays the BioData CATALYST interface, which is powered by Gen3. The top navigation bar includes links for Dictionary, Exploration, Discovery, Workspace, and Profile. The main interface is divided into several sections:

- Data Access:** Radio buttons for "Data with Access" (selected), "Data without Access", and "All Data".
- Filters:** Tabs for "Project", "Subject", and "Harmonized Variables". Under "Project", "parent" is selected with a count of 9,994. Under "Project Id", "parent-COPDGene\_HMB\_" is selected with a count of 9,994. Other filters include "parent-ARIC\_HMB-IRB\_" (8,780), "parent-FHS\_HMB-IRB-MDS\_" (7,162), "parent-MESA\_HMB\_" (3,667), and "parent-CHS\_HMB-MDS\_" (2,969).
- Export Options:** Buttons for "Export All to Terra", "Export All to Seven Bridges", "Export to PFB", and "Export to Workspace".
- Summary Statistics:** A central area showing "Projects: 1" and "Subjects: 9,994". It includes two donut charts: "Annotated Sex" (male: 5,341 (53.4%), female: 4,653 (46.6%)) and "Race" (white: 67.37%, black or african american: 32.63%).
- Table of Records:** A table showing the first 20 of 9,994 subjects. The table has columns for Project Id, Data Format, Race, Annotated Sex, Ethnicity, and BP Diastolic.

Project Id	Data Format	Race	Annotated Sex	Ethnicity	BP Diastolic
parent-COPDGene_HMB_	CRAM, VCF	white	male	not hispanic or latino	68
parent-COPDGene_HMB_	CRAM, VCF	black or african american	male	not hispanic or latino	77
parent-COPDGene_HMB_	CRAM, VCF	white	female	not hispanic or latino	80

# Step 1: Identify Public Control Cohorts via PIC-SURE

- Queried TOPMed Harmonized variables for race, smoking status variables for the COPDGene cohort
- Accessed via Jupyter Notebook as Python code using the Data Cruncher within BioData Catalyst Powered by Seven Bridges



# Questions to Answer



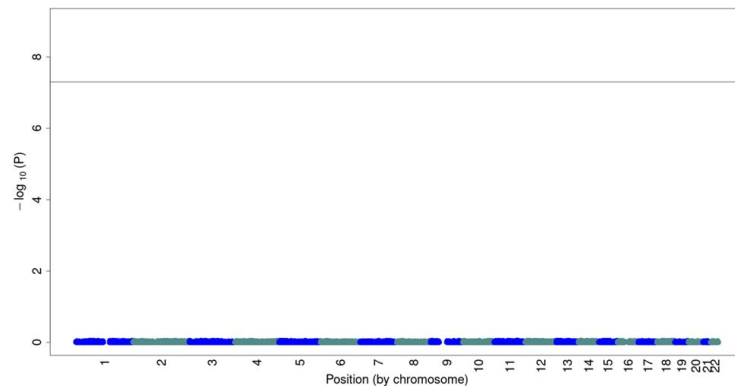
1. How to select WGS cohorts for the available array data?
2. Are Array genotyping data comparable with WGS data?
3. Will the population control method produce false-positive signals?
4. Can we reproduce GWAS hits by combining array genotyping data and WGS data?

# Cases with Array Genotyping Data vs Controls with WGS data

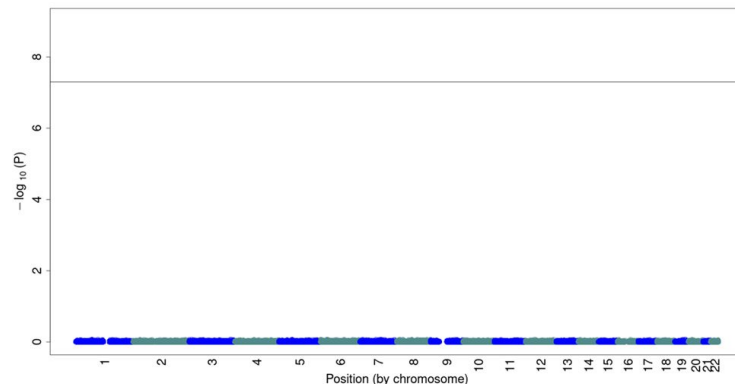
COPDGene Cohort

Same Set of Samples with Array and WGS data

A European ancestry



B African ancestry



# Questions to Answer



1. How to select WGS cohorts for the available array data?



2. Are Array genotyping data comparable with WGS data?

3. Will the population control method produce false-positive signals?

4. Can we reproduce GWAS hits by combining array genotyping data and WGS data?

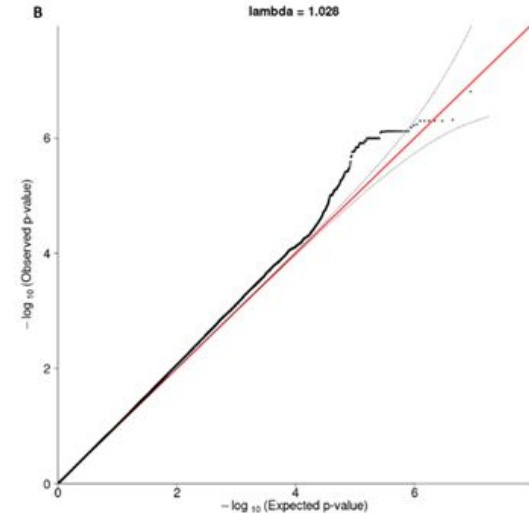
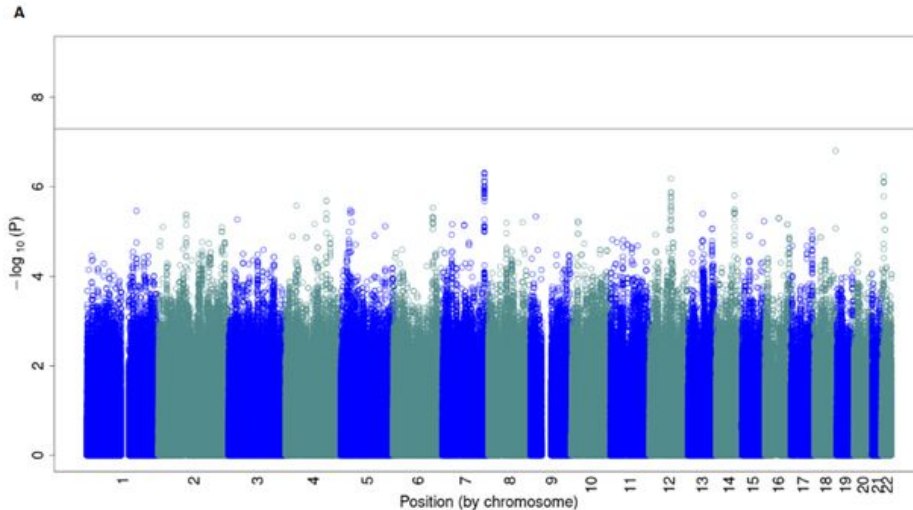
# Controlling False Positives: All Smokers

Control: All COPDGene samples with WGS genotyping

Cases: All COGEND samples with Array genotyping




Analysis: Meta-Analysis of GWAS (EV 1-15 and sex as covariates) within EA and AA Ancestries

Expectation: No Signal

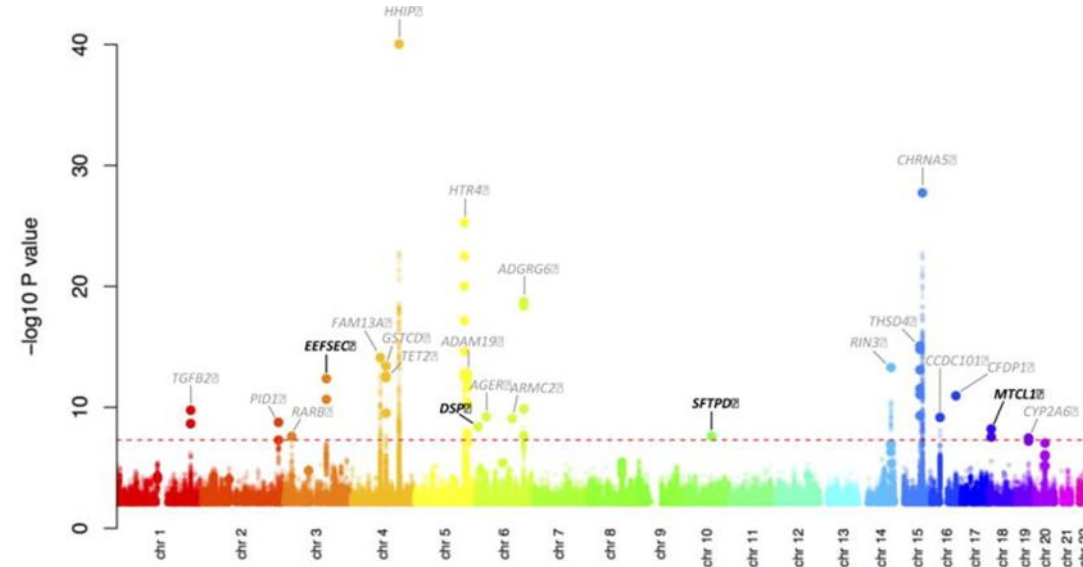




# Questions to Answer

-  1. How to select WGS cohorts for the available array data?
-  2. Are Array genotyping data comparable with WGS data?
-  3. Will the population control method produce false-positive signals?
4. Can we reproduce GWAS hits by combining array genotyping data and WGS data?

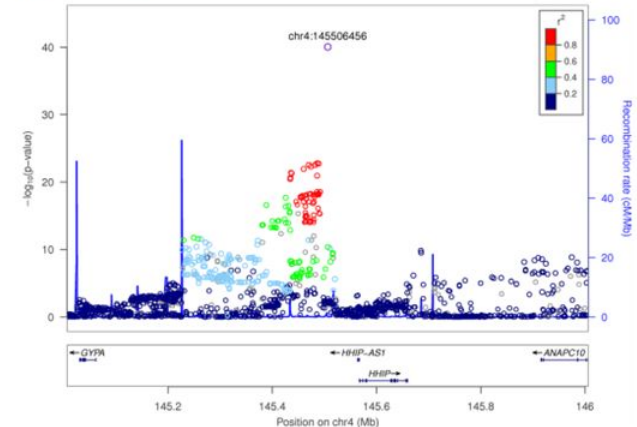
# Replicating Known COPD Signal



Hobbs, Brian D., et al. "Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis." *Nature genetics* 49.3 (2017): 426.

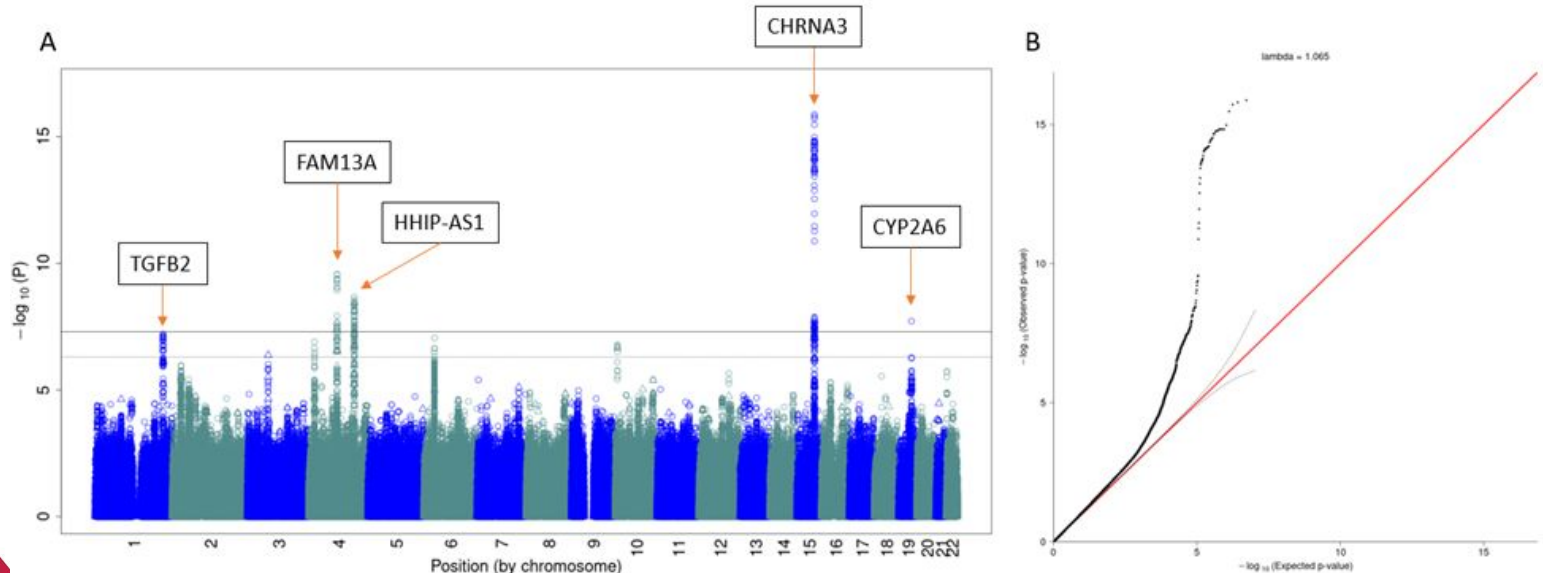
- Covariates: sex, pack-years of smoking, ever-smoking status, current-smoking status, and ancestry-based principal components
- Hit: HHIP on chromosome 4

Supplementary Figure 2a: LocusZoom for rs13141641 (HHIP locus)







# Replicating Known COPD Signal

- Data: COGEND Array Data vs COPDGene COPD cases WGS data
- Analysis: Meta-Analysis of GWAS (EV 1-15 and sex as covariates) within EA and AA Ancestries

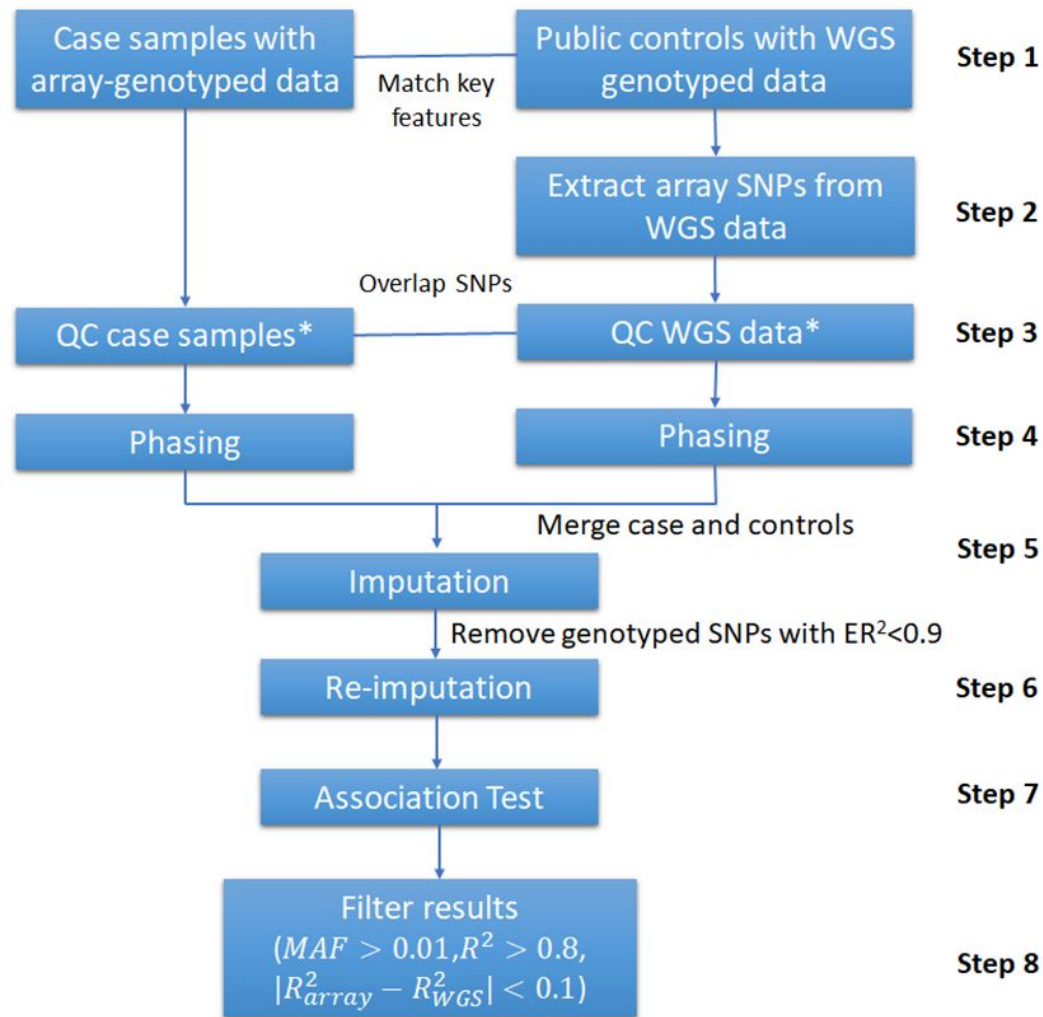


# Questions to Answer

-  1. How to select WGS cohorts for the available array data?
-  2. Are Array genotyping data comparable with WGS data?
-  3. Will the population control method produce false-positive signals?
-  4. Can we reproduce GWAS hits by combining array genotyping data and WGS data?

# GAWMerge: Genotyping Array and WGS Merging

- Protocol for integrating array and WGS genotyping data
- Implemented as a Common Workflow Language (CWL) Workflow on BioData Catalyst Powered by Seven Bridges



# Conclusions

- Utilized valuable *BioData Catalyst Powered by Gen3* and *BioData Catalyst Powered by PIC-SURE* to Identify Cohorts for integrating with available SNP array genotyping data
- Utilized *BioData Catalyst Powered by Seven Bridges* to implement the GAWMerge Protocol for integrating array and WGS genotyped data for GWAS
- BioData Catalyst is easy to use for conducting research, driving science, and new discoveries
  - Data from dbGaP and other sources are already decrypted and ready to analyze
  - GUI interface for creating and editing workflows is easy to use

# Acknowledgements



Manuscript: <https://doi.org/10.1101/2021.10.19.464854>

# Q&A with Ravi



# Announcements



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# 30 MIN BREAK

We will reconvene at 3:15 pm ET

**UP NEXT: GENESIS Workflows**



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# GENESIS Workflows

Alisa Manning, Terra  
Tony Patelunas, Seven Bridges



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# The TOPMed analysis Pipeline & GENESIS R/Bioconductor package

Components of the TOPMed analysis pipeline, originally written for the TOPMed Data Coordinating Center at UW have been translated to workflows in BioData Catalyst.

Documentation and more information:

[https://github.com/UW-GAC/analysis\\_pipeline](https://github.com/UW-GAC/analysis_pipeline)

## Analysis Steps:

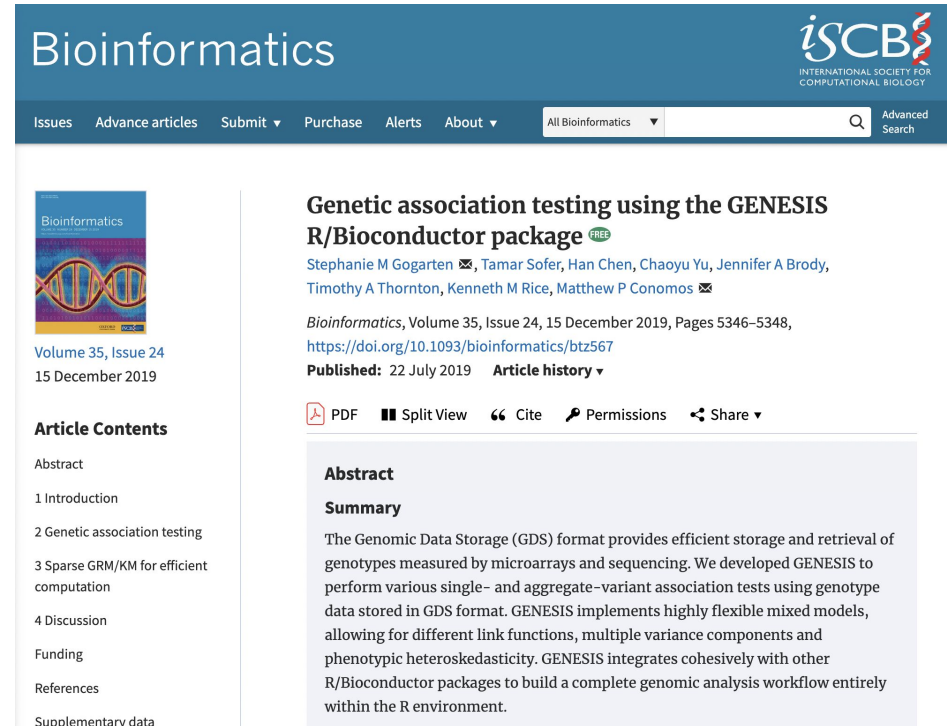
- Conversion to GDS
- Relatedness and Population structure
- Genetic Relationship Matrix
- Association testing
  - Null model
  - Single Variant
  - Rare variant

# The TOPMed analysis Pipeline & GENESIS R/Bioconductor package

## Genetic Association in two steps

**Null Model:** Creates an RData object with the results of fitting the regression model under the “Null Hypothesis” – i.e. no genetic association.

**Single variant / Rare Variant:** Uses the genetic data in ‘gds’ format to scan the files and perform genetic association tests



The screenshot shows the front page of a Bioinformatics journal article. The header includes the journal title 'Bioinformatics' and the ISCB logo. Navigation links for Issues, Advance articles, Submit, Purchase, Alerts, and About are present. A search bar is located on the right. The article title 'Genetic association testing using the GENESIS R/Bioconductor package' is prominently displayed, along with the authors' names and affiliations. The publication details, including the volume, issue, date, and page numbers, are listed. A 'Published' date and an 'Article history' link are also visible. The 'Abstract' section is partially visible, showing the summary of the article. The 'Article Contents' section lists the table of contents, including the introduction, genetic association testing, sparse GRM/KM for efficient computation, discussion, funding, references, and supplementary data.

Bioinformatics

INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Issues Advance articles Submit Purchase Alerts About All Bioinformatics Advanced Search

**Genetic association testing using the GENESIS R/Bioconductor package** FREE

Stephanie M Gogarten ✉, Tamar Sofer, Han Chen, Chaoyu Yu, Jennifer A Brody, Timothy A Thornton, Kenneth M Rice, Matthew P Conomos ✉

Bioinformatics, Volume 35, Issue 24, 15 December 2019, Pages 5346–5348, <https://doi.org/10.1093/bioinformatics/btz567>

**Published:** 22 July 2019 **Article history** ▼

PDF Split View Cite Permissions Share ▼

**Abstract**

**Summary**

The Genomic Data Storage (GDS) format provides efficient storage and retrieval of genotypes measured by microarrays and sequencing. We developed GENESIS to perform various single- and aggregate-variant association tests using genotype data stored in GDS format. GENESIS implements highly flexible mixed models, allowing for different link functions, multiple variance components and phenotypic heteroskedasticity. GENESIS integrates cohesively with other R/Bioconductor packages to build a complete genomic analysis workflow entirely within the R environment.

**Article Contents**

Abstract

1 Introduction

2 Genetic association testing

3 Sparse GRM/KM for efficient computation

4 Discussion

Funding

References

Supplementary data

<https://academic.oup.com/bioinformatics/article/35/24/5346/5536872>

# GENESIS Workflows on BioData Catalyst

Tutorial project on BioData Catalyst Powered by SevenBridges:

<https://platform.sb.biodatacatalyst.nhlbi.nih.gov/u/biodatacatalyst/genesis-tutorial/>

This project is designed to introduce the user to the GENESIS R package and related R packages (SeqArray, SeqVarTools, and SNPRelate) used to perform mixed model association testing in sequence data.

It consists of an interactive analysis with examples that will help the user understand the code that is used in GENESIS public apps, prepare data for input to those apps, and interact with the results. Also, there are several task examples for performing the analysis that are equivalent with the code in the interactive analysis.

The code in this project was developed as a series of exercises for the Summer Institute in Statistical Genetics, and is also available on github: [https://uw-gac.github.io/SISG\\_2021](https://uw-gac.github.io/SISG_2021).

# Learning objectives

## Part 1: Getting Started

- Link hosted files
- Create a project
- Launch Data Cruncher

## Part 2: Interactive Analysis

- Work in a Seven Bridges interactive environment to:
  - Convert VCF files
  - Explore data
  - Harmonize phenotypes

## Part 3: Tools/Workflows

- Use CWL apps to:
  - Fit a Null Model
  - Run a single variant association test
  - Monitor task progress

# Questions?



# Thank for joining us

Join the Community

Interact with the forum



Subscribe to our [YouTube channel](#)

Register for May Community Hours