

# Starting Cloud-Based Research on BDC

BDC for PCGC Fellows

**Tuesday, September 26th**



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

®

# Statement of Conduct

The BDC 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.

Web Resource: [Statement of Conduct](#)

# Agenda

Topic	Time
<b>Welcome!</b> Introductions, Housekeeping, and Icebreaker	10 min
<a href="#"><u>BDC's Role in Data Science</u></a> - BDC Mission and Vision, Intro to the ecosystem	20 min
<a href="#"><u>Using BDC in a Research Project</u></a>	
Data Discovery and Exploration	30 min
Analysis on BDC	30 min



# Welcome!



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

®

# Introductions

What inspired you to work  
in the sciences?



# BDC's Role in Data Science



National Heart, Lung,  
and Blood Institute

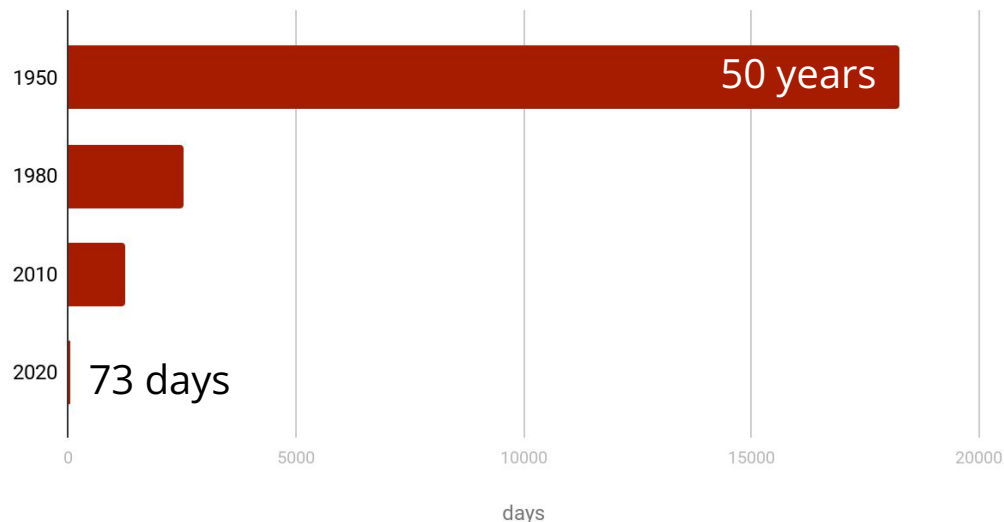
BioData

**CATALYST**

®

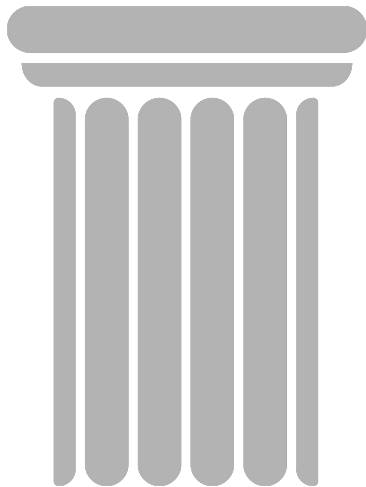
# The rate of data generation is accelerating rapidly

Doubling Time of Health Knowledge

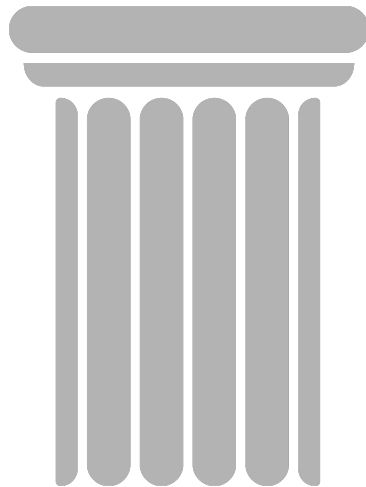


- More biomedical data will be generated this year than all previous years **combined**
- Diverse data modalities including EHR data, Survey, Sequencing, Transcriptomics, Metabolomics, Proteomics, Imaging, Sensor, E-Phys, Flow Cytometry, and so on

# Mission



# Vision



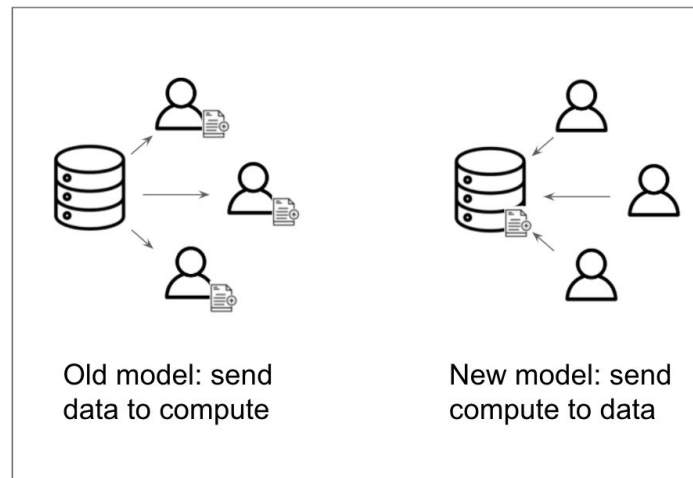
The **mission** is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.

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.



# Using the Cloud to store and analyze growing health data

- 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.



WHO?

WHAT?

WHERE?

SCIENCE!

WHY?



Genomics

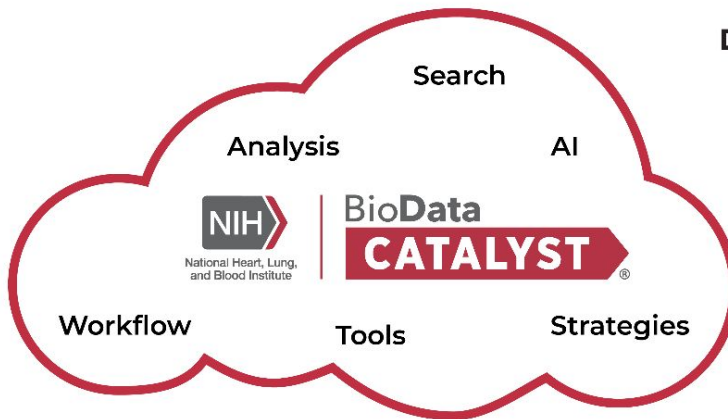


Clinical



Imaging

DATA  
HARMONIZATION



- UNDERSTAND
- OPEN SCIENCE
- CROSS-LINK

- COLLABORATE
- SCALE
- SHARE
- INTEROPERATE

HOW?

Diagnostic  
Tools

Therapeutic  
Options



DISCOVERY

Prevention  
Strategies



PATIENTS!

# What BDC offers



## Managing the Computing Environment

Elastic Computing

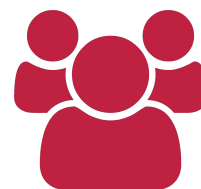


## Easier Access to many High Value Datasets



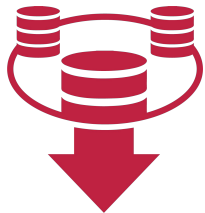
## Tooling

Data Discovery  
Statistical Analysis  
Tools (R, SAS)  
Other Specialized  
Workflows



## Community and Peer Interactions

# The Computing Environment



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



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



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



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

# Platforms and Services

## Explore Data

- PIC-SURE
- Gen3

## Analyze Data

- Seven Bridges
- Terra

→ View BDC Services

**What Do You Want to Do Today?**

**Explore Available Data**

*BDC-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](#)

*BDC-PIC-SURE*

Explore available data through *BDC-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](#)

**Analyze Data in Cloud-based Shared Workspaces**

*BDC-Seven Bridges*

Utilize collaborative workspaces for analyzing genomics data at scale. Access hosted datasets along with

*BDC-Terra*

Share and compute across large genomic and genomic-related datasets. Terra offers a stand-alone computational

# Community engagement and support

*Though the primary goal of BDC is to build a data science platform, at its core, this is a people-centric endeavor. BDC 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

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

# Community Hours

Monthly sessions on a variety of topics  
Materials made available for registrants

## Topics of interest:

- Exploring and Accessing Data
- Tour of the Analysis Workspaces
- Cloud Costs
- Community Showcases
- Interactive Analysis
- Reproducible Research Methods
- Reproducible Science

...and more!

**Budgeting for Scale**

Rep: Tony Patelunas, Seven Bridges

These are two examples of budgeting for scale, as you think about writing costs into a grant proposal. You need to do some calculations ahead of time: how many samples, what size are these samples, how much is that actually going to cost. The following are two examples - one for computation, one for storage - using the RNA-seq analysis for these examples.

**Slide 33: Budgeting for Scale: Computation**

**Platform prevents you from running further analyses when you reach billing group cap**

If you try to execute an analysis after you run out of funding in billing group, the platform will not allow the analysis to start and you will see an error message "insufficient funds in billing group."

Important for researchers to determine how to pay for additional cloud costs prior to reaching billing group cap (\$500) so that research is not delayed.

estimate cost. So if a user or RNA-seq, they want marking which is 4.5 GB about 4GB, it will cost cents per run times a

# Learning Resources

Many of the questions new users have may already be answered on either the BDC Gitbook or one of the Platform websites.

Our Gitbook documentation includes:

- Instructions on approvals and accounts needed to access BDC and how to check 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](#)



# Questions?

# Using BDC in a Research Project



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

®

# Curricula of:

## Using BDC in a Research Project

### Learning goals:

- **Search and select data relevant to your research question**
- Create and use a project in a cloud-computing analysis workspace
- Discover some available tools and workflows

# Platforms and Services

## Explore Data

- PIC-SURE
- Gen3

## Analyze Data

- Seven Bridges
- Terra

→ View BDC Services

**What Do You Want to Do Today?**

**Explore Available Data**

*BDC-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](#)

*BDC-PIC-SURE*

Explore available data through *BDC-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](#)

**Analyze Data in Cloud-based Shared Workspaces**

*BDC-Seven Bridges*

Utilize collaborative workspaces for analyzing genomics data at scale. Access hosted datasets along with

*BDC-Terra*

Share and compute across large genomic and genomic-related datasets. Terra offers a stand-alone computational

# Data on BDC

# Introduction to dbGaP

BDC ingests various datasets from the **Database of Genotypes and Phenotypes**, or **dbGaP** (<https://www.ncbi.nlm.nih.gov/gap/>)

What is dbGaP?

- Public repository for individual phenotype, exposure, genotype, and sequence data
- Main purpose is to archive and distribute the results of studies investigating the association between genotype and phenotype
- Researchers submit a Data Access Request (DAR) and are able to download the study files when authorized for research

# How is data organized in dbGaP?

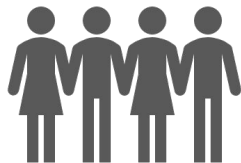
- Data is organized into **studies**
  - Each study has a specific **accession number** or unique identifier (e.g., phs000007)
- Studies have multiple **subjects**, or study participants
- Data organized by **consent groups**, based on consents given by subjects (research purposes their data can be used for)
- Studies consist of **phenotypic** and/or **genotypic** data
  - Phenotypic data is generally referred to as **variables**
  - Genotypic data is generally referred to as **samples**

# Data Available in BDC

3.42  
Petabytes of data



280,000+  
Participants



490,000+  
Data files



150,000+  
Whole genomes





# Data Available in BDC

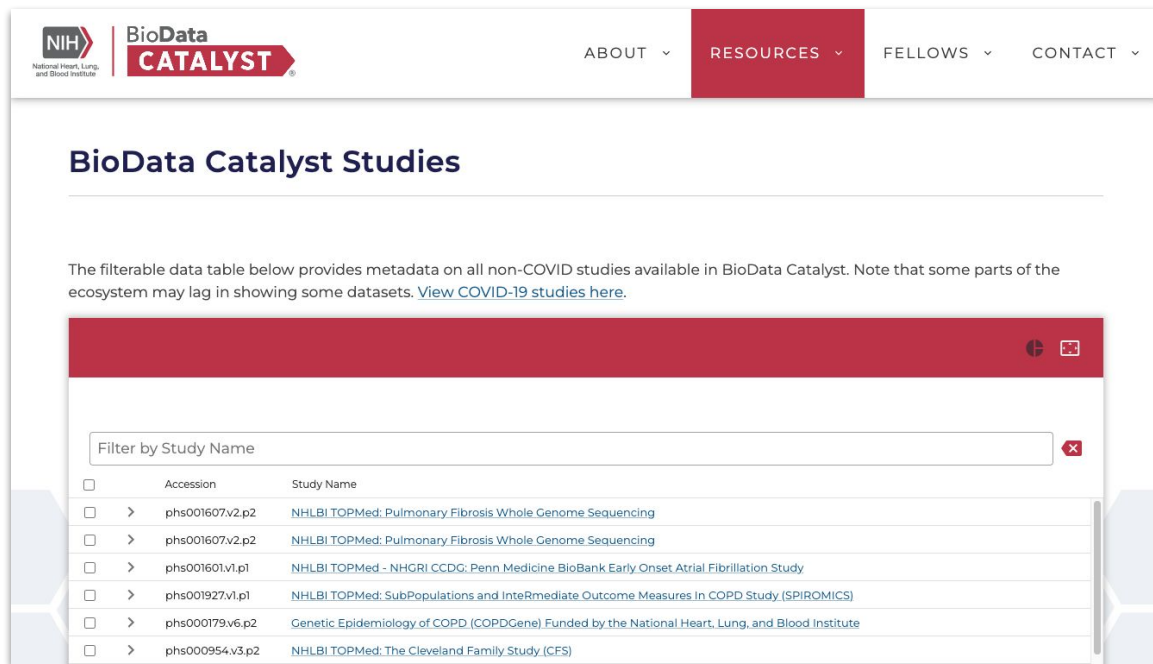
BDC is always ingesting  
new data

Check BDC website for a  
full list of studies available  
on the ecosystem

Resources → Data

Click “Explore Studies”

EXPLORE STUDIES 🔍



The screenshot shows the BioData Catalyst website. The header includes the NIH logo, the BioData CATALYST logo, and navigation links: ABOUT, RESOURCES (highlighted), FELLOWS, and CONTACT. The main heading is "BioData Catalyst Studies". Below this, a paragraph states: "The filterable data table below provides metadata on all non-COVID studies available in BioData Catalyst. Note that some parts of the ecosystem may lag in showing some datasets. [View COVID-19 studies here.](#)"

Below the text is a table with a search bar labeled "Filter by Study Name". The table has two columns: "Accession" and "Study Name".

Accession	Study Name
> phs001607.v2.p2	NHLBI TOPMed: Pulmonary Fibrosis Whole Genome Sequencing
> phs001607.v2.p2	NHLBI TOPMed: Pulmonary Fibrosis Whole Genome Sequencing
> phs001601.v1.p1	NHLBI TOPMed - NHGRI CCDC: Penn Medicine BioBank Early Onset Atrial Fibrillation Study
> phs001927.v1.p1	NHLBI TOPMed: SubPopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS)
> phs000179.v6.p2	Genetic Epidemiology of COPD (COPDGene) Funded by the National Heart, Lung, and Blood Institute
> phs000954.v3.p2	NHLBI TOPMed: The Cleveland Family Study (CFS)

<https://biodatacatalyst.nhlbi.nih.gov/resources/data/studies>

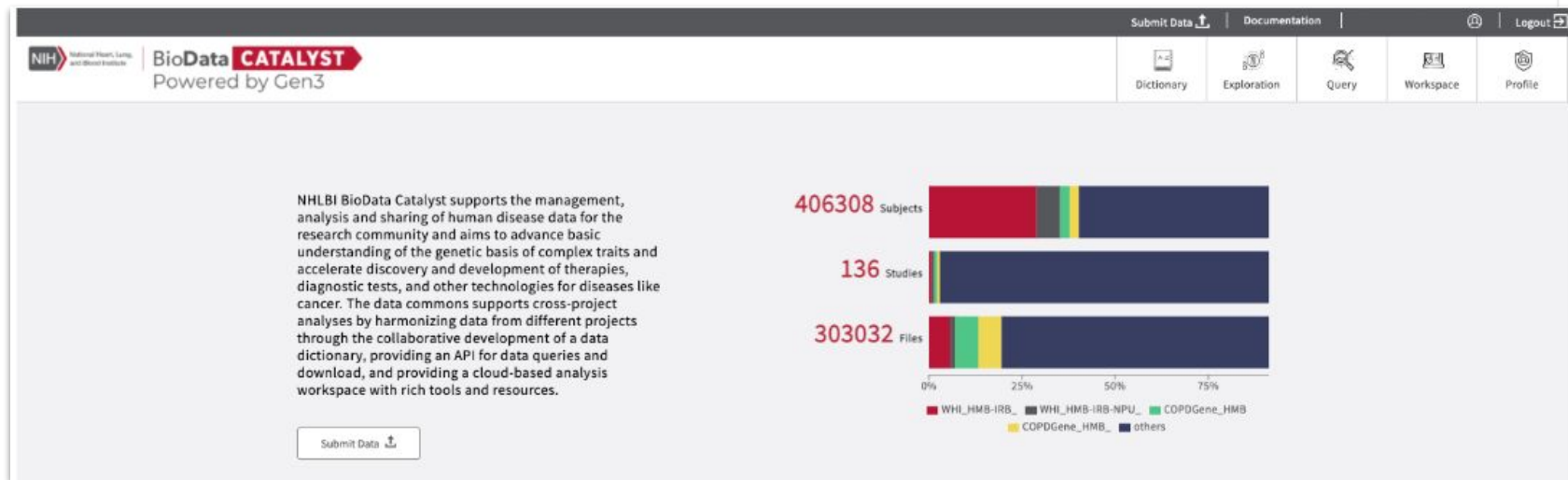
# Bring-Your-Own Data

- 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](#)


# Gen3 - Key Features

1. Source of Truth - File Object Persistence & Dataset Metadata
2. Interoperability - Standards-based integration points with other systems
3. Data Access - eRA Commons / dbGaP Authorization Inheritance
4. Data Ingestion - Robust data ingestion pipeline





# Gen3 - Discovery Page


A tool for discovery of released datasets (fully open, no required approval to discovery available data, [dbGaP FHIR](#)).


 National Heart, Lung, and Blood Institute


**BioData CATALYST**  
Powered by Gen3

 Dictionary

 Exploration

 **Discovery**

 Workspace

 Profile

Summary Statistics | Tags | Table of Records | Pagination

73  
STUDIES

296,115  
TOTAL SUBJECTS

🔍 heart

🔄 Reset Selection

▼ Study Filters

0 selected

STUDY NAME	FULL NAME	NUMBER OF SUBJECTS	DBGAP ACCESSION NUMBER	RELEASED	DATA AVAILABILITY	🗑️	▼
FHS_HMB-IRB-MDS_	Framingham Cohort	13,070	phs000007.v31.p12.c1	Yes	🔒		

See Grouping of Framingham Phenotype Datasets Startup of Framingham [Heart](#) Study. Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study (FHS) -- under the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute, NHLBI) -- embarked on a novel and ambitious project in health research. At the time, little was known about the general causes of heart disease and stroke, but the death rates for CVD had been increasing steadily since th...

Parent DCC Harmonized Clinical Phenotype dbGaP

# Gen3 - Exploration Page

A dynamic summary statistics display and cohort builder for export:

- Search facets leveraging harmonized variables.

Standardized Cohort Handoff support to move cohort to analysis workspaces (e.g. Broad's Terra System, Velsera's Seven Bridges system).

The screenshot displays the BioData CATALYST Exploration page. The top navigation bar includes the NIH logo, the text 'BioData CATALYST Powered by Gen3', and four tabs: 'Dictionary', 'Exploration' (which is active), 'Workspace', and 'Profile'. Below the navigation bar, there are two sub-tabs: 'Data' and 'File'. The 'Data' sub-tab is active, showing a 'Data Access' section with three radio buttons: 'Data with Access' (selected), 'Data without Access', and 'All Data'. To the right of the 'Data Access' section are four red buttons: 'Export All to Terra', 'Export All to Seven Bridges', 'Export to PFB', and 'Export to Workspace'. Below these buttons, a summary table shows 'Projects' as 4 and 'Subjects' as 10,835.

Projects	Subjects
4	10,835

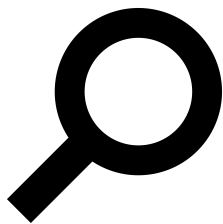
# Gen3 - Service and API Resources

- Useful Gen3 URL endpoints
  - <https://gen3.biodatacatalyst.nhlbi.nih.gov/submission>
  - <https://gen3.biodatacatalyst.nhlbi.nih.gov/query>
  - <https://gen3.biodatacatalyst.nhlbi.nih.gov/index/index/>
  - <https://gen3.biodatacatalyst.nhlbi.nih.gov/mds/metadata/<guid>>
  - <https://gen3.biodatacatalyst.nhlbi.nih.gov/DD>
- Gen3 SDK - <https://github.com/uc-cdis/gen3sdk-python>
- Gen3 Client - <https://gen3.org/resources/user/gen3-client/>

# Live Demo: Gen3 Discovery and Exploration Tools

# Empowering researchers to access data

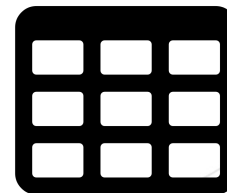
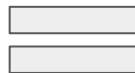
*BDC Powered by PIC-SURE* facilitates approachable research for all skill levels.



Search at the variable  
value and genomic  
variant level



Apply filters to create  
a cohort



Dataframe ready for  
research without  
opening any files or  
mapping to data  
dictionaries



# Live Demo: PIC-SURE Open Access

# 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

# Submitting a Data Access Request (DAR)

BDC uses dbGaP infrastructure for managing access to controlled-access data

Requirements:

1. An NIH eRA Commons account (or other valid NIH login). To learn more about this, visit [Understanding eRA Commons](#).
2. User must have Principal Investigator status. Those who are not PIs can ask their PI to add them as a data downloader.

# Questions?

# Analysis on BDC

# Curricula of:

## Using BDC in a Research Project

### Learning goals:

- Search and select data relevant to your research question
- **Create and use a project in a cloud-computing analysis workspace**
- **Discover some available tools and workflows**

# Analysis Platforms



National Heart, Lung,  
and Blood Institute

**BioData** **CATALYST**<sup>®</sup>

Powered by Seven Bridges



National Heart, Lung,  
and Blood Institute

**BioData** **CATALYST**<sup>®</sup>

Powered by Terra

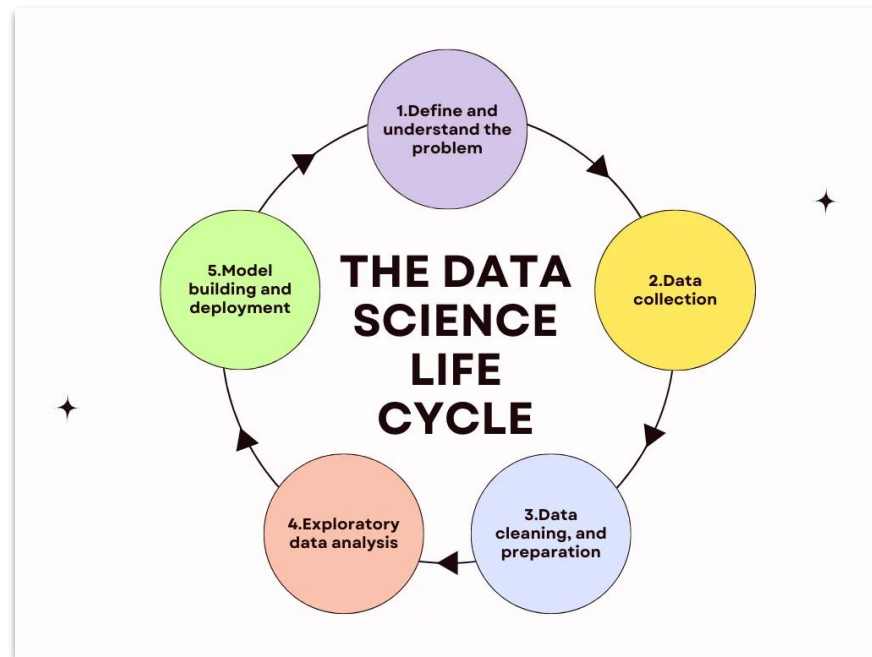
# Workflows

*Workflows (aka pipelines) are a series of steps performed by an external compute engine that are often used for automated, bulk analysis (such as aligning genomic reads)*



# Data Science Life Cycle

- Data science has an interactive nature
- BDC is here to support you on every step
- Examples
  - Data collection - PIC SURE / Gen3
  - Exploratory data analysis - Data Studio
  - Model building and deployment - CWL, WDL apps

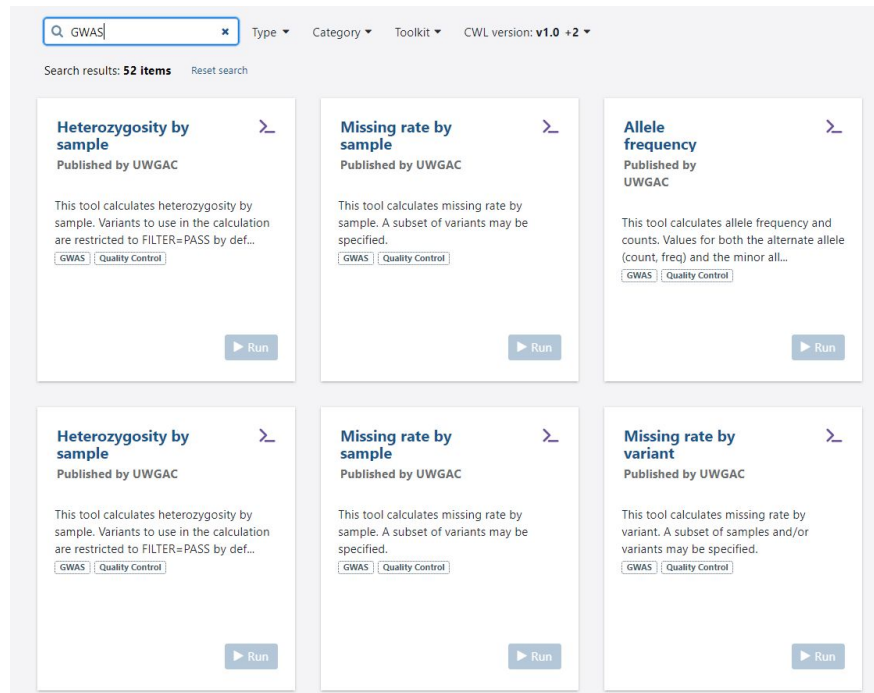


Credit: Madison Hunter | The data science project lifecycle.

# BDC-Seven Bridges

A curated collection of **800+** bioinformatics tools & workflows:

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



Open to the public @ [platform.sb.biodatacatalyst.nhlbi.nih.gov/public/apps](https://platform.sb.biodatacatalyst.nhlbi.nih.gov/public/apps)

# BDC-Terra

- Can write your own in WDL
- Can access 1,500+ public workflows in our methods repository

METHOD

amp-pd-workflows/rna-collect-rna-seq-metrics

SNAPSHOT  
1 ▼

Publicly Readable

Export to Workspace...

Summary

WDL

Configurations

Clone...

## Synopsis

This workflow runs Picard's CollectRnaSeqMetrics on a BAM file

## Snapshot Comment

## Method Owner

admin@amp-pd.org

## Created

October 29, 2019 at 2:02 PM

## Documentation

rna-collect-rna-seq-metrics.wdl

This workflow runs Picard's CollectRnaSeqMetrics on a BAM file

([https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/picard\\_analysis\\_CollectRnaSeqMetrics.php](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/picard_analysis_CollectRnaSeqMetrics.php)).

### Inputs:

- Per-sample:
- Sample name - Name of the sample associated with the BAM - used in names of outputs.
- BAM file - Path to a BAM file.
- Ref flat file - Gene annotations in refFlat form.
- Ribosomal intervals file - Location of rRNA sequences in genome, in interval\_list format.

METHOD

amp-pd-workflows/rna-collect-rna-seq-metrics

SNAPSHOT  
1 ▼

Publicly Readable

Export to Workspace...

Summary

WDL

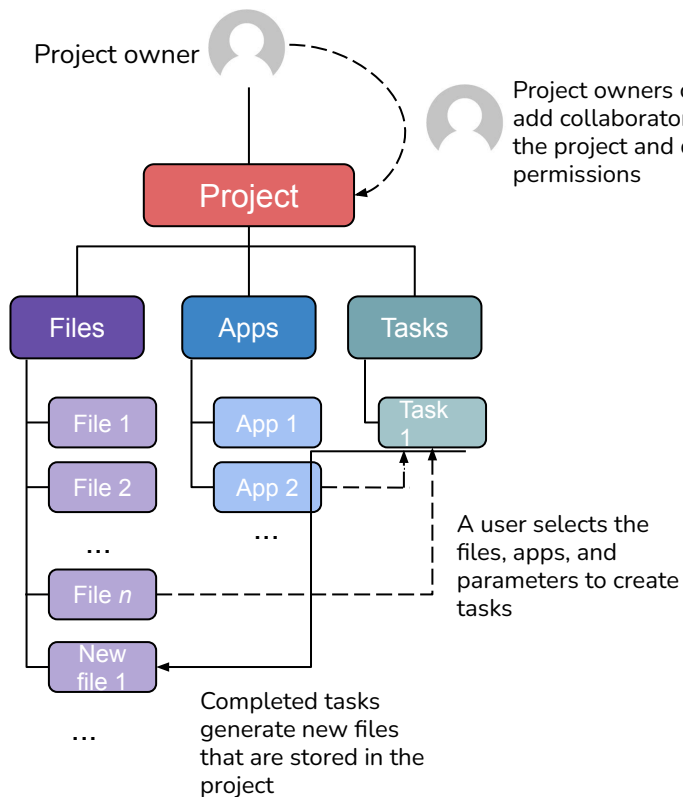
Configurations

```
1 ## rna-collect-rna-seq-metrics.wdl
2 ##
3 ## This workflow runs Picard's CollectRnaSeqMetrics on a BAM file
4 ## (https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/picard_analysis_CollectRnaSeqMetrics.php).
5 ##
6 ## Inputs:
7 ## - Per-sample:
8 ##   - Sample name - Name of the sample associated with the BAM - used in names of outputs.
9 ##   - BAM file - Path to a BAM file.
10 ##   - Ref flat file - Gene annotations in refFlat form.
11 ##   - Ribosomal intervals file - Location of rRNA sequences in genome, in interval_list format.
12 ##
13 ## - VM configuration
14 ##   - Docker image url
15 ##   - VM disk size
16 ##   - VM memory
17 ##   - Num CPU cores
18 ##   - Runtime zones, ex: "us-central1-a us-central1-b"
19 ##   - Number of times to try the workflow with a preemptible VM before
20 ##     falling back to a full-price VM.
21 ##
22 ## Outputs:
23 ## - Per-sample:
24 ##   - <sample-id>.RNA_Metrics
25 ##
26 workflow RNACollectRnaSeqMetrics {
27   String sample_name
28   File bam_file
29
30   ...
```

# Workspaces and Projects

*Workspaces (BioData Catalyst powered by Terra) and Projects (BioData Catalyst powered by Seven Bridges) are dedicated space where you and your collaborators can access and organize the same data and tools and run analyses together.*

# Seven Bridges - Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions

NIH BioData CATALYST Powered by Seven Bridges

Projects Data Public Gallery Public projects Automations Developer Staff alisonleaf

Dashboard Files Apps Tasks

Alison\_test\_GWAS Interactive Analysis Settings Notes

DESCRIPTION

**Welcome to your new project!**

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

**Within your project, you can:**

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

Please record the details of your project here, such as its aims, experimental context, and 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

# User friendly workflow editor enables reproducibility by default

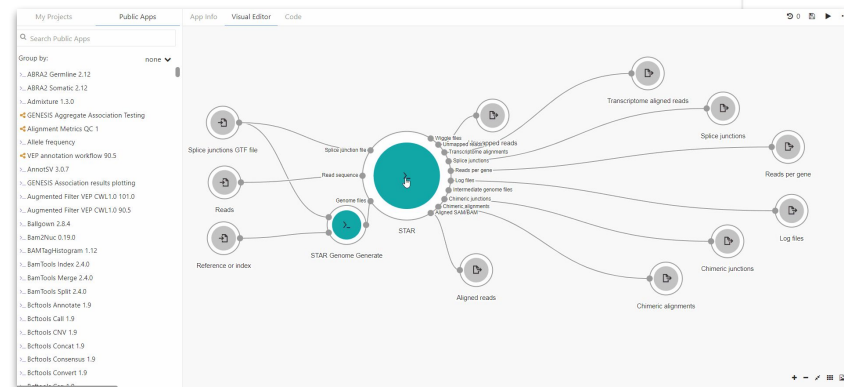
Common Workflow Language enables **portability**, **reproducibility**, and **scalability**

Use or combine 800+ optimized tools and workflows to construct your analysis

Seamlessly import workflows from external public repos

Create your own tools with our CWL Tool Editor

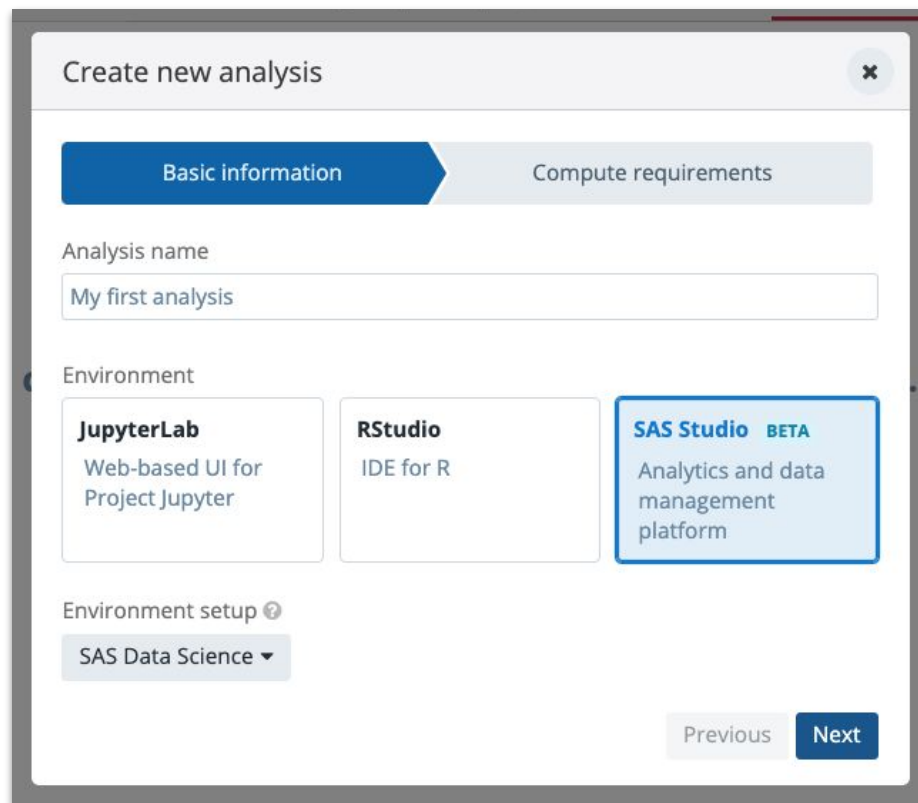
Expose or lock parameters appropriately



# Interactive analysis

**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.

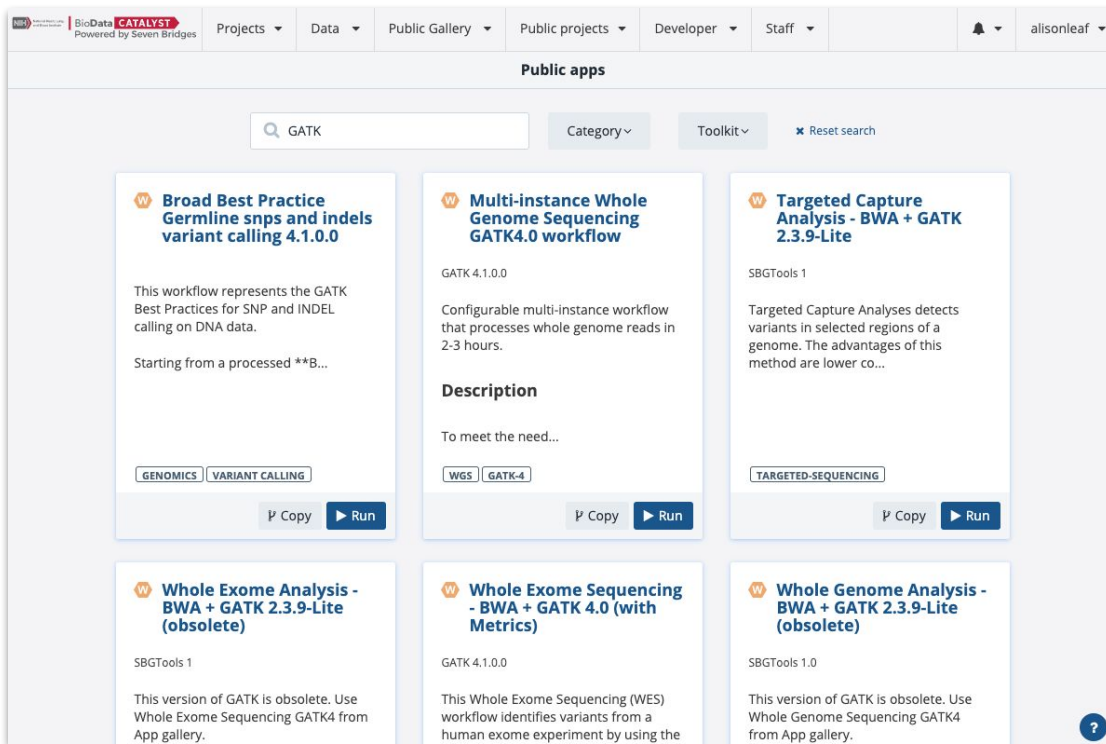


The screenshot shows a 'Create new analysis' dialog box with a close button (X) in the top right corner. It features two tabs: 'Basic information' (active) and 'Compute requirements'. Under 'Basic information', there is a text input field for 'Analysis name' containing 'My first analysis'. Below this is the 'Environment' section, which displays three selectable options: 'JupyterLab' (described as 'Web-based UI for Project Jupyter'), 'RStudio' (described as 'IDE for R'), and 'SAS Studio BETA' (described as 'Analytics and data management platform'). The 'SAS Studio BETA' option is highlighted with a blue border. At the bottom of the dialog is the 'Environment setup' section, which includes a dropdown menu currently set to 'SAS Data Science'. Navigation buttons 'Previous' and 'Next' are located at the bottom right.

# Find the tools you need in the Public Apps Gallery

A curated collection of **800+** bioinformatics tools & workflows:

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





# 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 (ISAMtools\_Index)**

Number of threads

**Picard MarkDuplicates (RPicard\_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 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 ☐ Memoization: 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

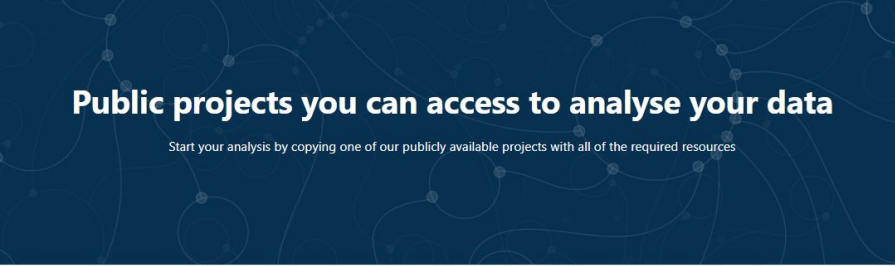
# Getting Help - Contacting Support from the platform

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 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 in the top right corner. Below the navigation bar, there is a sub-navigation bar with 'Dashboard' (highlighted), 'Files', 'Apps', and 'Tasks'. The main content area is titled 'Genesis tutorial' and includes links for 'Interactive Analysis', 'Settings', and 'Notes'. A 'DESCRIPTION' tab is active, showing a 'Welcome to' message and a list of actions: 'Start exploring', 'Install your tool', 'Upload your own datasets', and 'Collaborate securely'. A 'MEMBERS' tab is also visible, showing a list of users: 'milan.domazet' (OWNER), 'smgogarten', and 'mconomos03'. A 'Need help?' modal is open in the center, providing links to 'Create a project', 'Manage the project dashboard', 'Add notes to your project', 'Leave a project', 'Delete a project', 'Add a collaborator to a project', 'Set permissions', 'Interactive analysis', and 'Modify project settings'. Below these links, it says 'Not finding what you need? Visit our Knowledge Center' and 'Contact our support' with a text input field 'Describe your issue or share your ideas' and a 'Send' button. In the bottom right corner, a 'Help and support' button is circled in red, with a red arrow pointing to it from the left.

# Public Projects provide examples to follow

- Worked out example projects to get you started
- Contain files, apps, completed analysis tasks, Data Studio examples
- Can copy the entire project to your private dashboard



**Public projects you can access to analyse your data**

Start your analysis by copying one of our publicly available projects with all of the required resources

<b>Introduction to SAS</b> # Introduction to SAS __ ## About this project This project was built as a collaboration between the SAS team and Seven Bridges. The goal is for a user to learn the basics of the SAS programming language through annotated code and example analysis including importing data, cleaning data, running regressions, and working with genetic sequencing data. There are three parts to the project. <a href="#">Copy project</a>	<b>Automated Chest Imaging Platform (CIP) CT Phenotyping and Machine Learning Discovery in COPD</b> # Chest Imaging Platform (CIP) CT Phenotyping and Machine Learning Discovery in COPD ----- ####This project is developed and maintained by [Applied Chest Imaging Laboratory] ( <a href="https://acil.med.harvard.edu/">https://acil.med.harvard.edu/</a> ) at Harvard Medical School and hosted by Seven <a href="#">Copy project</a>	<b>TOPMed Freeze8 Variant Calling</b> # TOPMed Freeze8 Variant Calling Pipeline --- ## About the Analysis **TOPMed Variant Calling pipeline** detects and genotypes variants from a list of aligned sequences. It was developed by Hyun Min Kang and Jonathon LeFaive from the University of Michigan - [GitHub] ( <a href="https://github.com/statgen/topmed_variant_">https://github.com/statgen/topmed_variant_</a> ) <a href="#">Copy project</a>
<b>Data Interoperability</b> # Data Import with DRS ### Examples on How to Bring the Data From Other Cloud Environments As part of their [DRS API] ( <a href="https://github.com/ga4gh/data-repository-service-schemas">https://github.com/ga4gh/data-repository-service-schemas</a> ) effort, the GA4GH Cloud Work Stream group has <a href="#">Copy project</a>	<b>PIC-SURE API</b> This project contains JupyterLab and RStudio example notebooks for accessing PIC-SURE API. They can be located in the [Data Studio] ( <a href="https://platform.sb.biodatacatalyst.nhlbi.nih.gov/sure-api/analysis/cruncher">https://platform.sb.biodatacatalyst.nhlbi.nih.gov/sure-api/analysis/cruncher</a> ). You can <a href="#">Copy project</a>	<b>COVID-19 Image Segmentation with Deep Learning</b> This project provides deep learning image segmentation tools in a Jupyter notebook, along with a pretrained model for segmenting lung area from CT <a href="#">Copy project</a>

# Live Help - Seven Bridges Office Hours

## Questions? Need help?

We hold sessions twice a week:

**Tuesdays** at 10am ET

**Thursdays** at 2pm ET

Come chat with us about your research!

Does Thursdays at 2pm ET work  
for everyone for the next few  
weeks to meet?



Scan the QR code

Or browse to

<https://meet.google.com/kbs-ojnj-dcg>

# BDC-Terra

- Dashboard

- General overview of the workspace that includes documentation on the workspace itself, cloud information, owners, and tags
- Good documentation makes your analysis easy to share (with others, as well as with your future self) and reproduce.

The screenshot shows the BDC-Terra workspace dashboard. The header includes the NIH logo, BioData CATALYST logo (Powered by Terra), and the workspace name 'WORKSPACES'. The breadcrumb trail is 'Workspaces > biodata-catalyst/BioData Catalyst GWAS 1000 Genomes Tutorial > Dashboard'. The main navigation bar has tabs for DASHBOARD, DATA, ANALYSES, WORKFLOWS, and JOB HISTORY. The dashboard content is titled 'ABOUT THE WORKSPACE' and 'GWAS Tutorial in NHLBI's BioData Catalyst'. It provides a brief overview of the tutorial workspace and links to documentation. A 'Data Model' section describes the setup for the GWAS analysis. On the right, a sidebar contains 'WORKSPACE INFORMATION' (Last Updated: 7/20/2021, Creation Date: 3/13/2020, Access Level: Project Owner), 'CLOUD INFORMATION', 'OWNERS', 'TAGS', and 'NOTIFICATIONS'. A 'Rate: \$0.00 per hour' badge is visible in the top right corner.

NIH | BioData CATALYST Powered by Terra

WORKSPACES

Workspaces > biodata-catalyst/BioData Catalyst GWAS 1000 Genomes Tutorial > Dashboard

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

ABOUT THE WORKSPACE

### GWAS Tutorial in NHLBI's BioData Catalyst

This tutorial workspace offers example tools for conducting mixed-models GWAS from start to finish using the [NHLBI BioData Catalyst](#) ecosystem. We've created a set of documents [to get you started in the BioData Catalyst system](#). If you're ready to conduct an analysis, proceed with this dashboard:

#### Data Model

This template was set up to work with the NHLBI BioData Catalyst Gen3 data model. In this dashboard, you'll learn how to import open access data from the Gen3 platform into this Terra template and conduct an association test. If you have never used the Gen3 data model before, we suggest you start with the tutorial [Getting Started with Gen3 Data in Terra](#).

**WORKSPACE INFORMATION**

Last Updated	7/20/2021
Creation Date	3/13/2020
Access Level	Project Owner

**CLOUD INFORMATION**

**OWNERS**

**TAGS**

**NOTIFICATIONS**

Rate: \$0.00 per hour

# BDC-Terra

- Data

- Import your own data or access data that is stored in Terra
- Convenient spreadsheet formatted data tables help keep track of all project data, no matter where files are stored in the cloud.

NIH National Heart, Lung, and Blood Institute | BioData CATALYST Powered by Terra | BETA WORKSPACES | Workspaces > fc-product-demo/Terra-Notebooks-Quickstart > Data

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY | Workspace is locked and read only

+ IMPORT DATA | EDIT | OPEN WITH... | EXPORT | SETTINGS | 0 rows selected | ADVANCED SEARCH | Search | Rate: \$0.00 per hour

TABLES	subject_id	age	bmi_baseline	dbgap_accession_number
<input type="checkbox"/>	HG00096	75	25.3	synthetic_data_set_1
<input type="checkbox"/>	HG00097	63	26.9	synthetic_data_set_1
<input type="checkbox"/>	HG00099	48	23.9	synthetic_data_set_1
<input type="checkbox"/>	HG00100	46	24.3	synthetic_data_set_1
<input type="checkbox"/>	HG00101	37	24.9	synthetic_data_set_1
<input type="checkbox"/>	HG00102	37	25.1	synthetic_data_set_1
<input type="checkbox"/>	HG00103	41	25	synthetic_data_set_1
<input type="checkbox"/>	HG00105	61	28.5	synthetic_data_set_1
<input type="checkbox"/>	HG00106	88	27.5	synthetic_data_set_1

Search all tables

BigQuery\_table (2) | cohort (1) | **subject (2504)** | REFERENCE DATA | OTHER DATA

No references have been added. Add reference data

Workspace Data | Files

1 - 100 of 2504 | 1 2 3 4 5 | Items per page: 100



# BDC-Terra

- Analyses
  - Integrate and visualize your data in real time using Galaxy, Jupyter Notebooks, or RStudio
  - All three apps run on virtual machines or clusters of machines in a workspace Cloud Environment.

The screenshot displays the BioData CATALYST WORKSPACES interface. The top navigation bar includes the NIH logo, 'BioData CATALYST Powered by Terra', a 'BETA' badge, and 'WORKSPACES'. The breadcrumb trail shows 'Workspaces > biodata-catalyst/BioData Catalyst GWAS 1000 Genomes Tutorial > Analyses'. The main navigation tabs are DASHBOARD, DATA, ANALYSES (selected), WORKFLOWS, and JOB HISTORY. The 'Your Analyses' section features a '+ START' button and a search bar. Below is a table of analyses:

Application	Name	Last Modified
Jupyter	2-GWAS-preliminary-analysis.ipynb	Sep 2022
Jupyter	3-GWAS-genomic-data-preparation.ipynb	Sep 2022
Jupyter	1-Prepare-Gen3-data-for-exploration.ipynb	Sep 2022
Jupyter	terra_data_table_util.ipynb	Sep 2022

A sidebar on the right indicates a 'Rate: \$0.00 per hour'.



# BDC-Terra

- Workflows
  - Collect, configure (set up) and run workflows for bulk analyses

The screenshot displays the BioData CATALYST Workflows interface. At the top, the header includes the NIH logo, 'BioData CATALYST Powered by Terra', a 'BETA' badge, and the 'WORKSPACES' section. The breadcrumb trail shows 'Workspaces > biodata-catalyst/BioData Catalyst GWAS 1000 Genomes Tutorial > Workflows'. A navigation bar below the header contains links for 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS' (which is highlighted), and 'JOB HISTORY'. On the right side of the header, there is a notification bell icon with a '6' badge.

The main content area is titled 'WORKFLOWS' and features a search bar labeled 'SEARCH WORKFLOWS', a 'Sort By: Alphabetical' dropdown menu, and two view toggle buttons (grid and list). Below these, there are three workflow cards:

- Find a Workflow**: A card with a blue plus icon and a circular information icon.
- 1-vcfToGds**: A card showing 'V. master' and 'Source: Dockstore' with a circular information icon.
- 2-genesis\_GWAS**: A card showing 'V. v1\_4\_1' and 'Source: Dockstore' with a circular information icon.

On the far right, a vertical sidebar displays a 'Rate: \$0.00 per hour' and a lightning bolt icon.

# BDC-Terra

- Workflows
  - Can edit WDL script directly or use UI to assign inputs and outputs

← Back to list

## ① 2-genesis\_GWAS

Version: v1\_4\_1

Source: [github.com/AnalysisCommons/genesis\\_wdl/genesis\\_GWASv1\\_4\\_1](https://github.com/AnalysisCommons/genesis_wdl/genesis_GWASv1_4_1)

Synopsis:

No documentation provided

☐ Run workflow with inputs defined by file paths

☒ Run workflow(s) with inputs defined by data table

**Step 1**

Select root entity type:

**Step 2**

No data selected

☒ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks ☐ Retry with more memory ☐ Ignore empty outputs

SCRIPT \*\* INPUTS \*\* OUTPUTS \*\*

Hide optional inputs [Download json](#) | [Drag or click to upload json](#) | [Clear inputs](#)

Task name ↓	Variable	Type	Attribute
genesis_gwas_wf	these_genotype_files	Array[File]	<input type="text" value="this.gds"/> <a href="#">[-]</a>
genesis_gwas_wf	this_outcome_name	String	<input type="text" value="this.outcome_name"/> <a href="#">[-]</a>
genesis_gwas_wf	this_pheno_file	File	<input type="text" value="this.phenotypes"/> <a href="#">[-]</a>

## ② 2-genesis\_GWAS

Version: v1\_4\_1

Source: [github.com/AnalysisCommons/genesis\\_wdl/genesis\\_GWASv1\\_4\\_1](https://github.com/AnalysisCommons/genesis_wdl/genesis_GWASv1_4_1)

Synopsis:

No documentation provided

☐ Run workflow with inputs defined by file paths

☒ Run workflow(s) with inputs defined by data table

**Step 1**

Select root entity type:

**Step 2**

No data selected

☒ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks ☐ Retry with more memory ☐ Ignore empty outputs

SCRIPT \*\* INPUTS \*\* OUTPUTS \*\*

```
1 task genesis_nullmodel {
2   String outcome_name
3   String? outcome_type
4   String? covariates_string
5   File pheno_file
6   File genotype_file
7   String results_file
8   File? kinship_matrix
9   String? pheno_id
10  String? conditional
11  String? het_varsIn
12  String? transform
13  String? transform_rankNorm
14  String? transform_rescale
```

# ***BDC-Terra***

- Useful Workspaces
  - [Working with GnomAD data](#)
  - [Whole Genome Analysis Pipeline](#)
  - [Workflows Tutorial](#)
- Useful Workflows
  - [Processing for Variant Discovery](#)
  - [CRAM to BAM](#)
  - [Generate Sample Map](#)
  - [HaploType Caller](#)



# Live Demo: Seven Bridges



# Working Session

1. Create your own project/sandbox on BDC-Seven Bridges
  - a. Name project in this format - **PCGC\_First\_Last**
  - b. Add any of the BDC reps or other PCGC Fellows to the project
  - c. Copy files, apps, and JupyterLab from the main project
2. **Discuss:**
  - a. What software do you need to run on BDC for **your research**?
    - i. Example: GWAS will require different software than RNA-Seq differential expression
  - b. What is your timeline?

Feel free to ask any questions.

# Live Help - Seven Bridges Office Hours

## Questions? Need help?

We hold sessions twice a week:

**Tuesdays** at 10am ET

**Thursdays** at 2pm ET

Come chat with us about your research!

Does Thursdays at 2pm ET work  
for everyone for the next few  
weeks to meet?



Scan the QR code

Or browse to

<https://meet.google.com/kbs-ojnj-dcg>

# Closing

# Session Recap

- BDC is a cloud-based ecosystem of tools that enable research, including:
  - Data discovery and exploration
  - Research analysis and workflow development
- Projects and Workspaces are secure, private containers for your files, software and analysis tasks
- Workflows are scalable analysis tools that can be written yourself in either WDL or CWL or found in one of BDC's various methods repositories



# Next steps

- Practice: Log-in and try out what we discussed
- Engage on the forum
- Join Seven Bridges Office Hours
- Discussion: What would be most helpful for **YOU**?



# Questions?