Workshop Forum: https://bit.ly/BDC-NDSA-Forum

Starting Cloud-Based Research on NHLBI BioData Catalyst® (BDC)

BDC for NDSA Workshop

Day 1: July 13 | 11 AM ET

NIH
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

Workshop Forum: https://bit.ly/BDC-NDSA-Forum

Agenda: Day 1

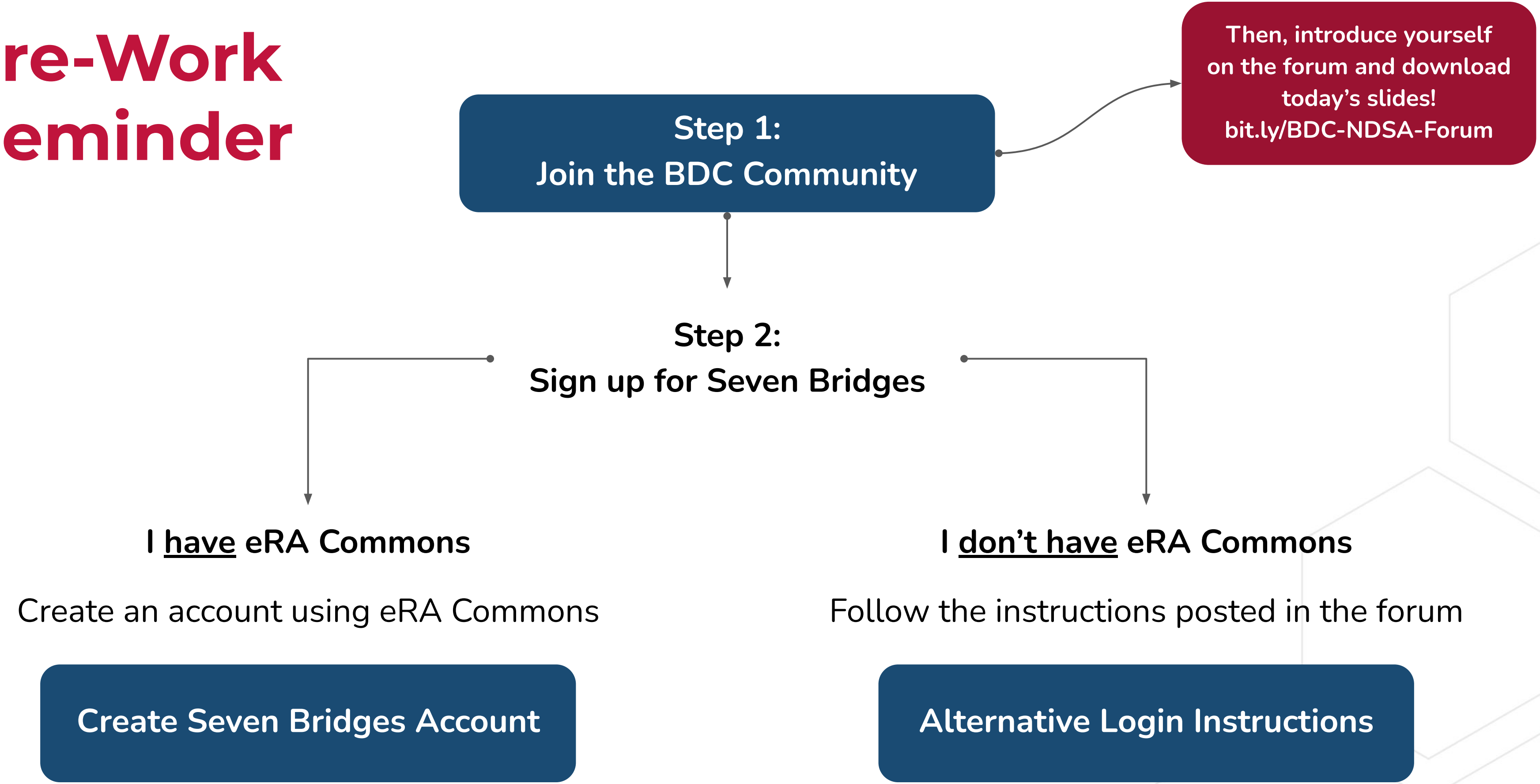
Topic	**Time**
Welcome! Introductions, Housekeeping, and Icebreaker	30 min
BDC's Role in Data Science + Discussion → Present Day Data Science Challenges → BDC's Approach → Future of Data Science and discussion	1 hour
30 MINUTE LUNCH BREAK @ 12:30 ET	
Using BDC in a Research Project → Data Discovery and Exploration → Analysis on BDC	> 2 hours
Closing & Recap + Working Session	40 min

Workshop Forum: https://bit.ly/BDC-NDSA-Forum

Agenda: Day 2

Topic	Time
Reproducible Research	1.5 hour
30 MINUTE LUNCH BREAK @ 12:40 ET	
Using BDC as a teaching platform → Use Cases and Guest Speaker with Q&A → Benefits and Challenges of Cloud Teaching with discussion	> 1.5 hour
Workflows and Cost Estimation → Sign up for $500 in Pilot Credits	> 1 hour
Closing & Recap	40 min

Pre-Work Reminder



Step 1:
Join the BDC Community

Then, introduce yourself on the forum and download today's slides!
bit.ly/BDC-NDSA-Forum

Step 2:
Sign up for Seven Bridges

I <u>have</u> eRA Commons

Create an account using eRA Commons

Create Seven Bridges Account

I <u>don't have</u> eRA Commons

Follow the instructions posted in the forum

Alternative Login Instructions

Introductions

Meet Your Instructors

Emily Hughes

BDC Powered by PIC-SURE

Harvard Medical School

Kat Thayer

BDC Powered by Terra

Broad Institute

David Roberson

BDC Powered by Seven Bridges

Seven Bridges/Velsera

Justin Dorsheimer

BDC Powered by Gen3

University of Chicago

TAs and Live Support

Cera Fisher

BDC Powered by Seven Bridges

Seven Bridges/Velsera

Aarthi Krishnan

BDC Powered by Seven Bridges

Seven Bridges/Velsera

Michael Corace

BDC Powered by PIC-SURE

Harvard Medical School

Amber Voght

User Engagement Specialist,
BDC Coordinating Center

Kaleena Narwani

User Engagement Specialist,
BDC Coordinating Center

Icebreaker Activity

What inspired you to work in the sciences?

→ Answer now on Jamboard

BDC’s Role in Data Science

Curricula of:

BDC's Role in Data Science

Learning goals

- Acknowledge the state of the data science landscape
- Assess BDC's impact on field of data science
- Consider the trajectory of where data science is going
- Discover ways to support your and your students' science data interests using BDC

Present Day
Data Science Challenges

Perspective

Industry Representation

N = 4,367

Citizenship

73.3% US Citizens

Top 3 areas of Employment

1. Genetic Counseling (45.7%)
2. Research (30.4%)
3. Academic (23.4%)

Employment Status

78% Permanent Positions

Age

45.1% 25-40 years old

Gender Identity and Sexual Orientation

74.7% Women

23.3% Men

0.5% Nonbinary/Transgender

6.9% LGBTQIA

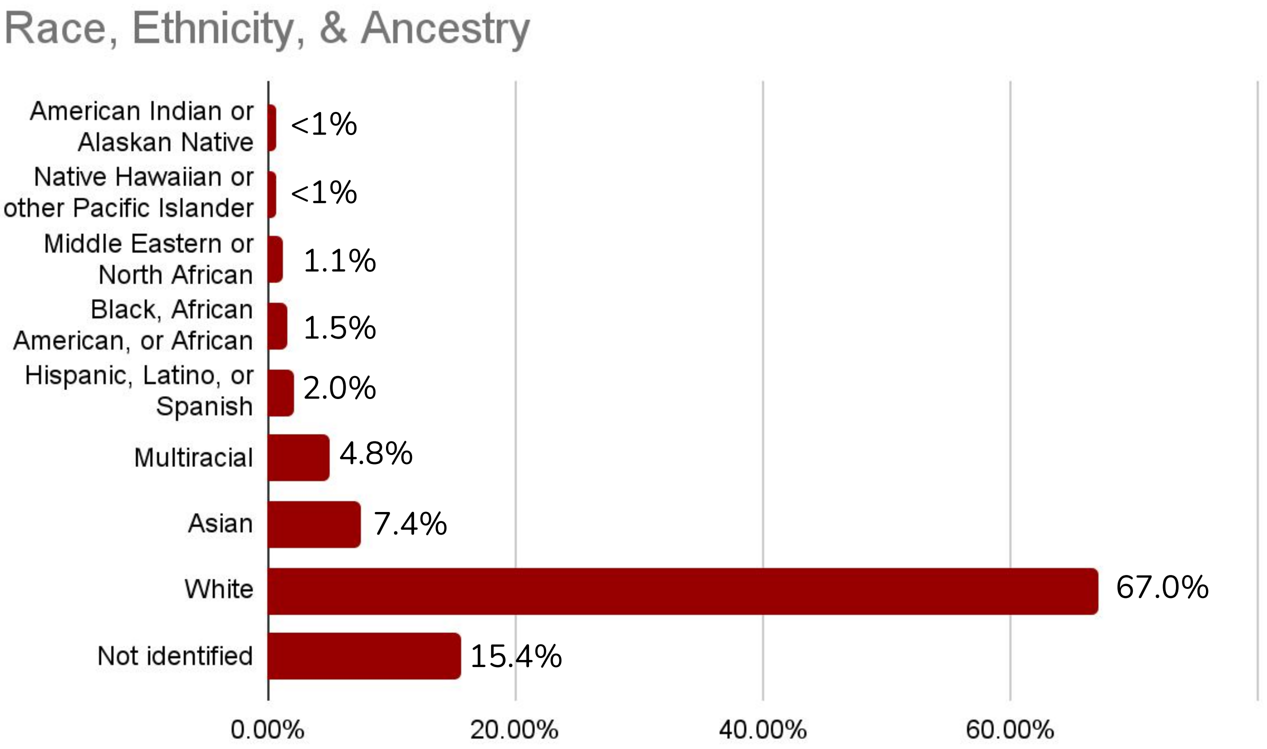
Disability Status

3.4% Reported having a disability

ASHG: Human Genetics & Genomics Workforce Survey

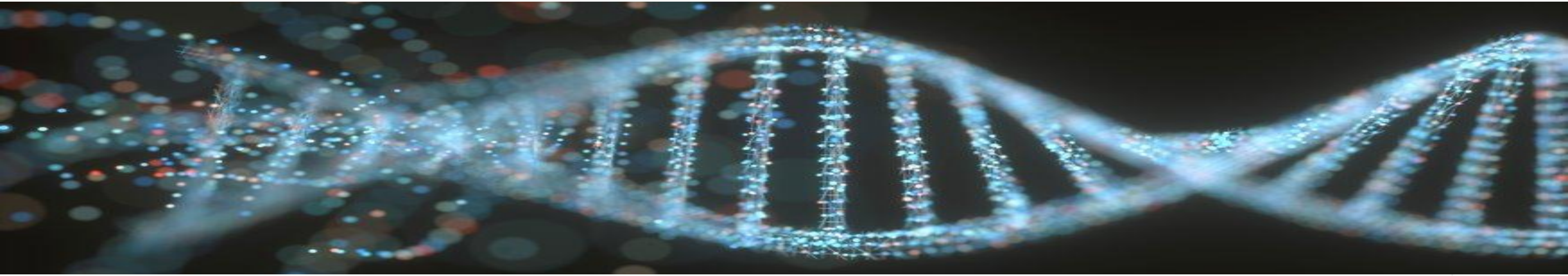
Industry Representation

N = 4,367



ASHG: Human Genetics & Genomics Workforce Survey

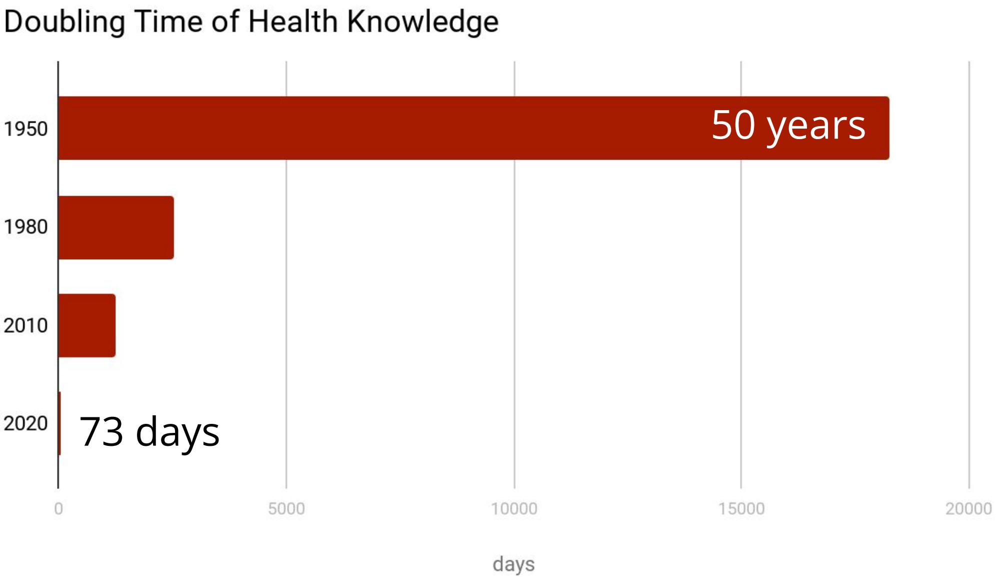
Data Representation



- 0.1% genomic differences come from variations among ~3 billion bases in our DNA
- Most permitted DNA samples used in research come from people with European Ancestry (78%)
- Active Initiatives: Polygenic Risk Methods in Diverse Populations Consortium, Electronic Medical Records and Genomics Network, and the Human Pangenome Reference Consortium

Genome.gov: Diversity in Genomics Research Fact Sheet

The rate of data generation is accelerating rapidly



- 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

Scalability

- Capturing data
  - Size of raw data
  - Funding considerations
  - Saving information
- Analyzing data
  - Computing time and power
  - Comparing data sets and harmonization
- Sharing results
  - Can be slow to impact other fields

Sources - Genome.gov:

Cost of Sequencing a Human Genome

Genomic Data Science Fact Sheet

Reproducibility

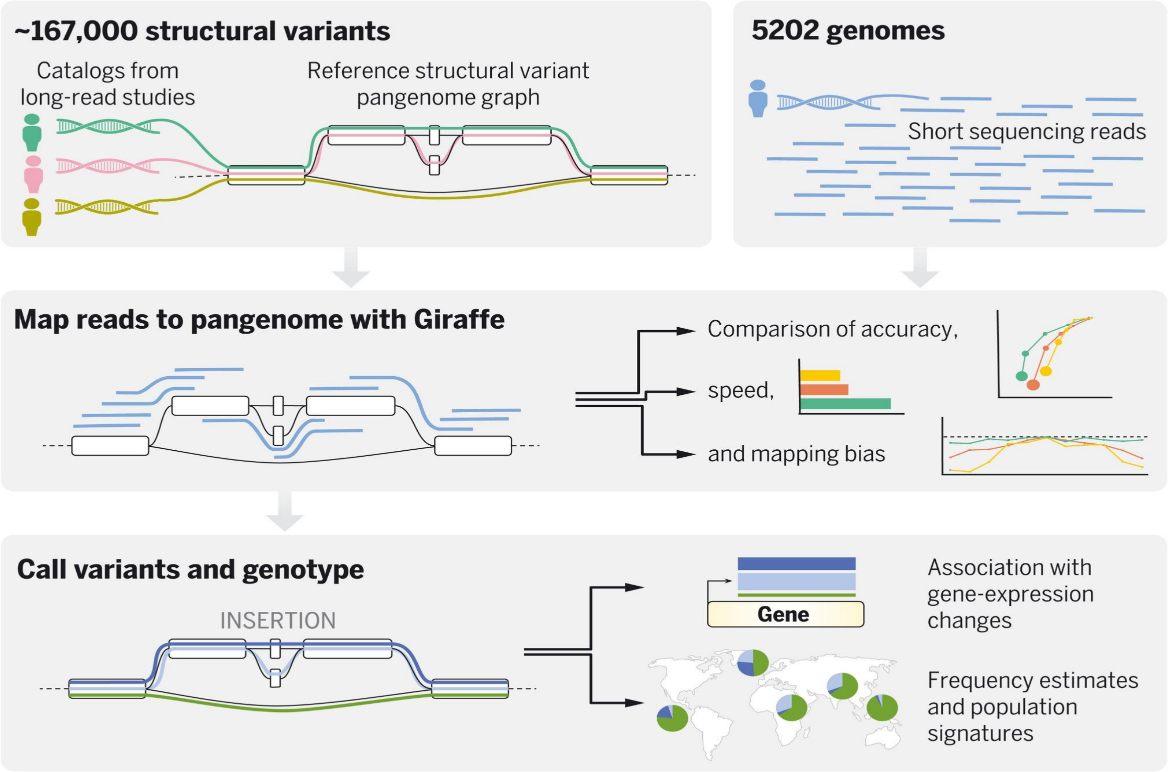
- Data harmonization and documentation
- Use diverse, yet comparable datasets
- Working in silos
  - Hyper local
  - Breaking through the noise
- Study errors
  - Type I Errors - false positives and generalizability

Find the Needle in the Haystack, Then Find it Again

Access

- IRB and data sharing
- Data privacy laws
  - California Privacy Law
- Special omics considerations
  - Patient De-identification
  - Secondary data scope allowance

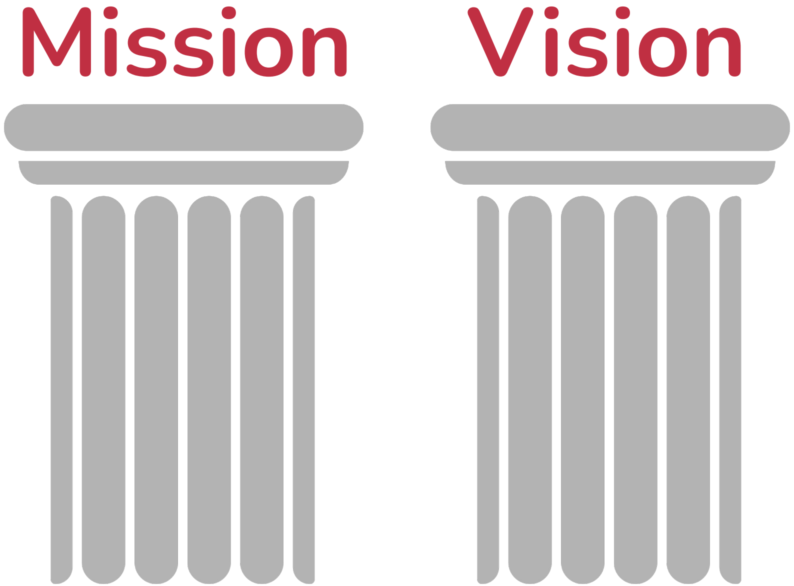
Case Study: Giraffe


~167,000 structural variants
Catalogs from long-read studies
Reference structural variant pangenome graph
5202 genomes
Short sequencing reads
Map reads to pangenome with Giraffe
Comparison of accuracy,
speed,
and mapping bias
Call variants and genotype
INSERTION
Gene
Association with gene-expression changes
Frequency estimates and population signatures

Brief Discussion

Of the challenges discussed, what resonated with you most?
What other challenges come to mind?

BDC’s Approach

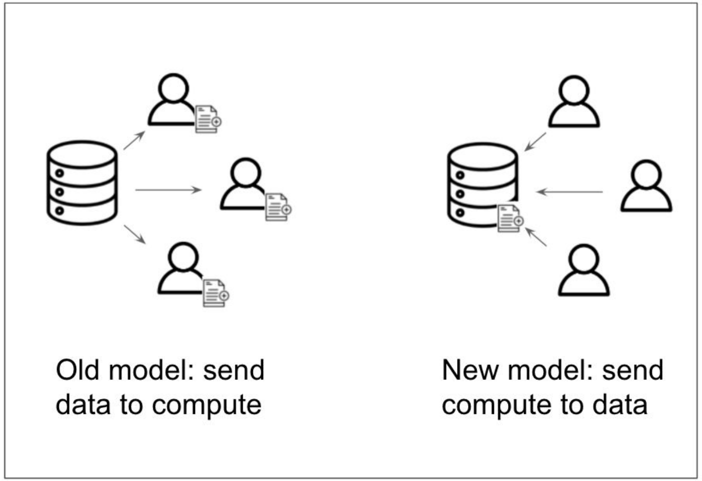


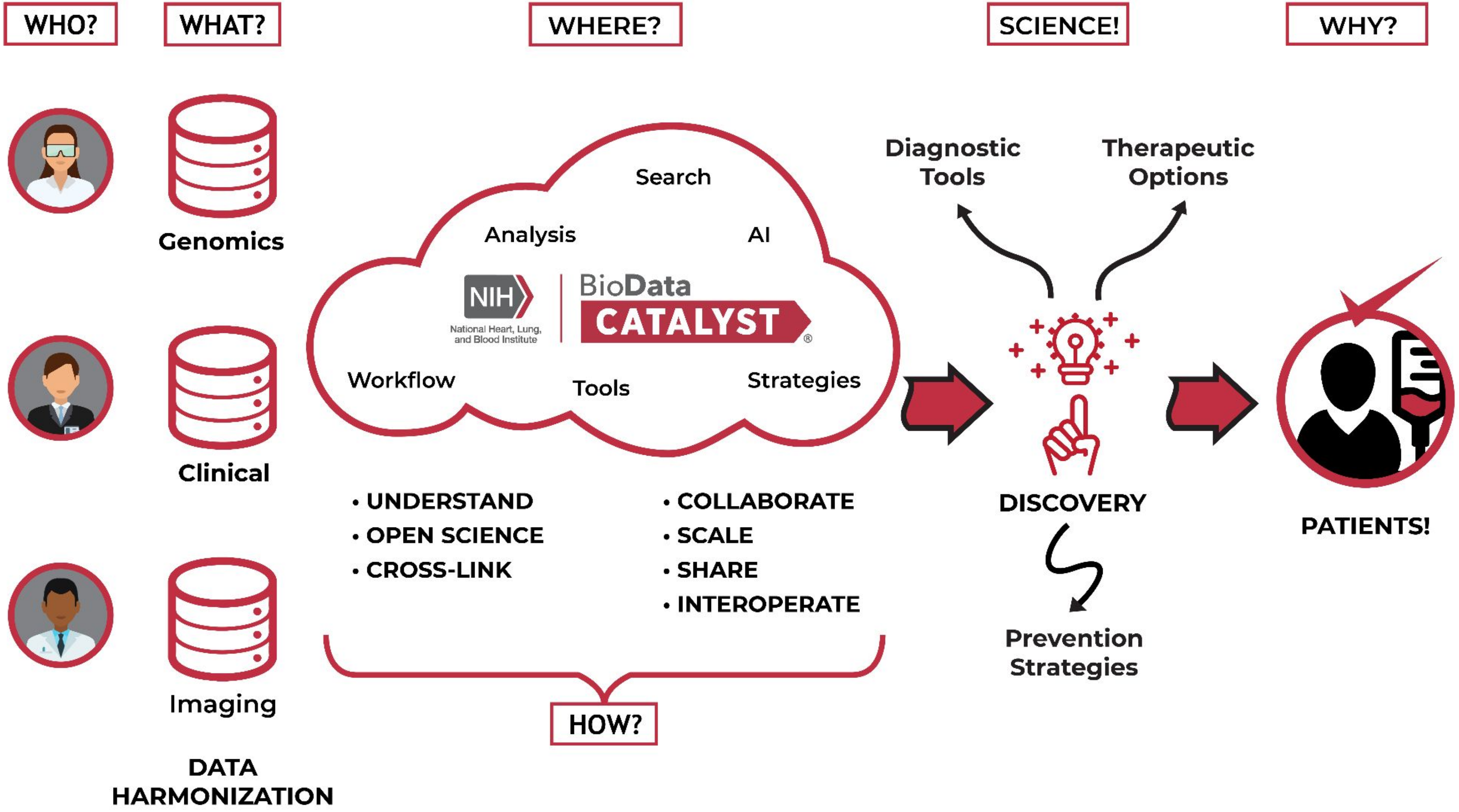
The ***mission*** is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.

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.


Old model: send data to compute
New model: send compute to data



WHO?

WHAT?

Genomics

Clinical

Imaging

DATA HARMONIZATION

WHERE?

Search

Analysis

AI

NIH National Heart, Lung, and Blood Institute | BioData CATALYST®

Workflow

Tools

Strategies

- UNDERSTAND
- OPEN SCIENCE
- CROSS-LINK

- COLLABORATE
- SCALE
- SHARE
- INTEROPERATE

HOW?

SCIENCE!

Diagnostic Tools

Therapeutic Options

DISCOVERY

Prevention Strategies

WHY?

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

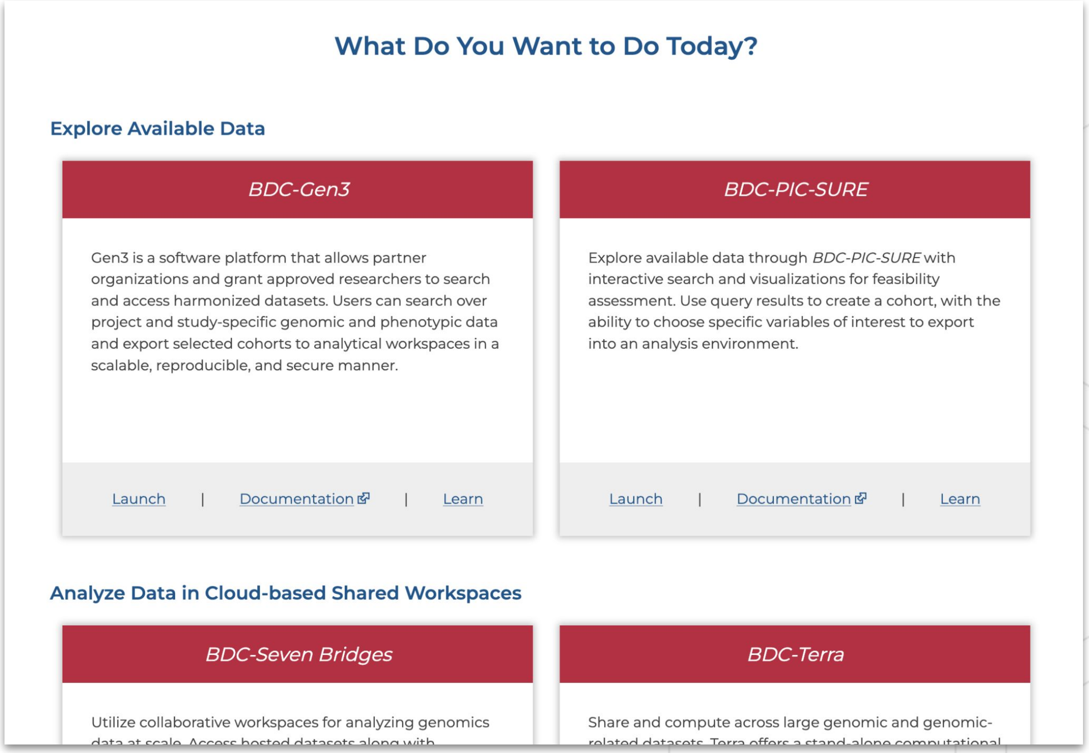
Explore Data

- 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

NIH National Heart, Lung, and Blood Institute | BioData CATALYST®

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

Questions?

The Future of Data Science

- How do you see the field of data science **evolving** in the **next five years**?
- What are **topics** we in the profession should be paying attention to right now? Why?
- What can be done to help better **prepare** future data scientists?

→ Answer now on Jamboard

Data Science Community Engagement

Fellowship programs

52 BDC Fellows over three cohorts

- Hands-on, modularized training:
  - Introduction to the platform
  - Creating Your Own Tool/Workflow
  - Performing GWAS
  - Estimating Cloud Costs
- Scientific projects were executed which otherwise might not have been possible
- Retrospective interviews for user-defined development

Four Bench2Bassinet Fellows

- Hands-on, modularized training:
  - Introduction to the platform
  - Creating Your Own Tool/Workflow
  - Interoperability
  - Estimating Cloud Costs

Read about the BDC Fellows →

Workshops & Courses

- **UW Summer Institute for Statistical Genetics**
  - >300 students over four years
- **American Thoracic Society**
  - Asynchronous “flipped” approach for 50 attendees
- **CHARGE Consortium Annual Meeting, 2022**
  - Hands-on workshop for 75+ attendees, with interactive demo
- **Howard University Data Science Collaboration**
- **AIM-AHEAD PRIME**
- **PRIDE Programs**
  - Held six sessions with 50+ students in 2022

image credit: https://www.biostat.washington.edu/suminst/sisg

Questions?

Break

While you wait...

If you **have an eRA Commons account** and would like to follow along with the upcoming demonstration, visit https://picsure.biodatacatalyst.nhlbi.nih.gov/ and log into the platform.

BREAK

Using BDC in a Research Project

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

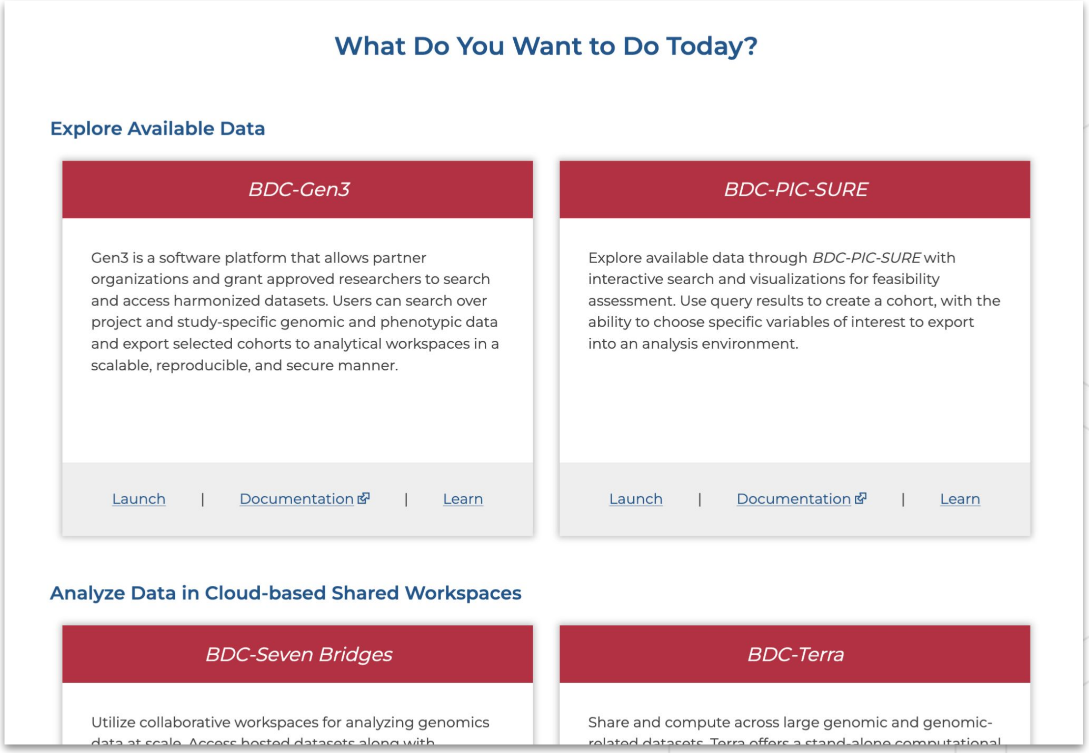
Explore Data

- 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

Finding 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

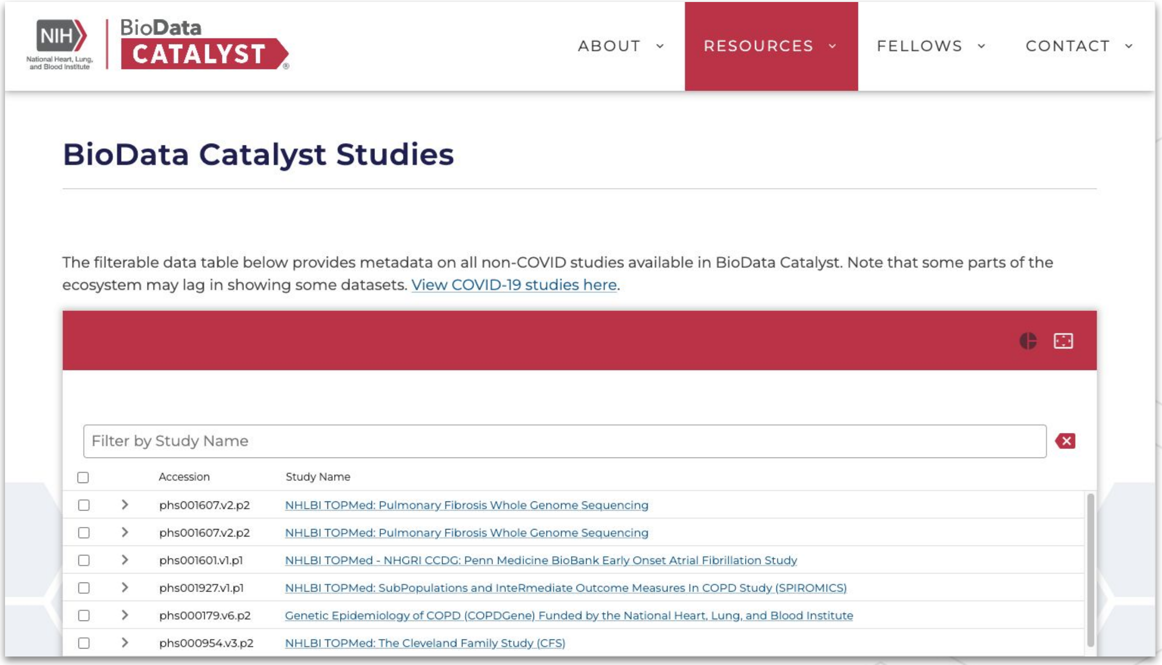
BDC is always ingesting new data

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

Resources → Data

Click "Explore Studies"

EXPLORE STUDIES



BioData Catalyst Studies

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.

Filter by 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 CCDG: 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

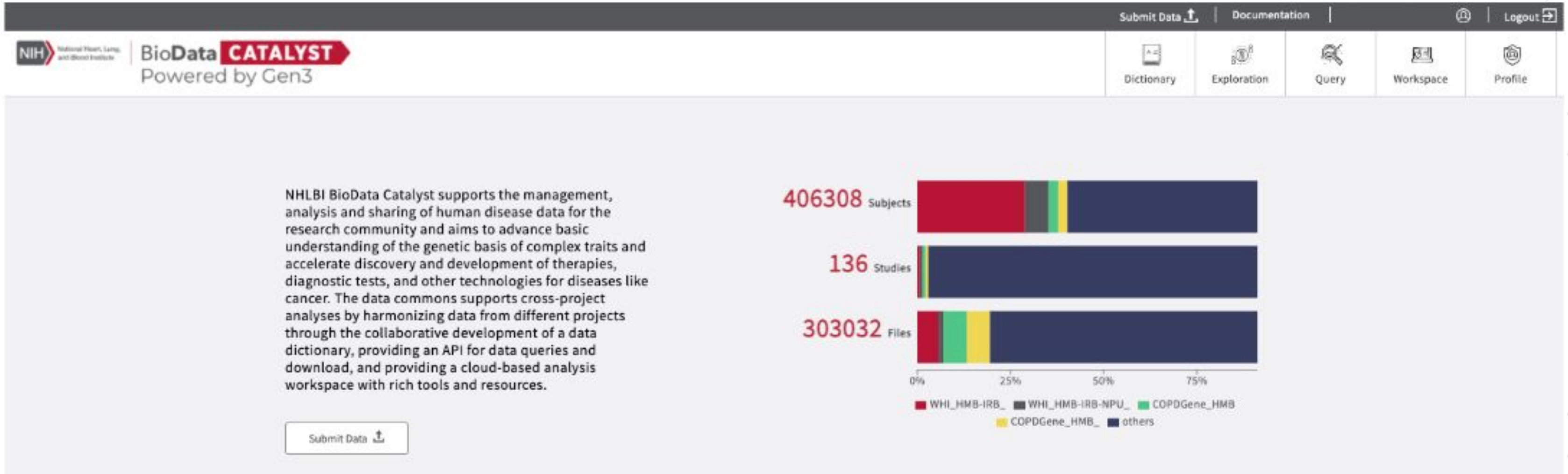
Bring-Your-Own Data

- 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

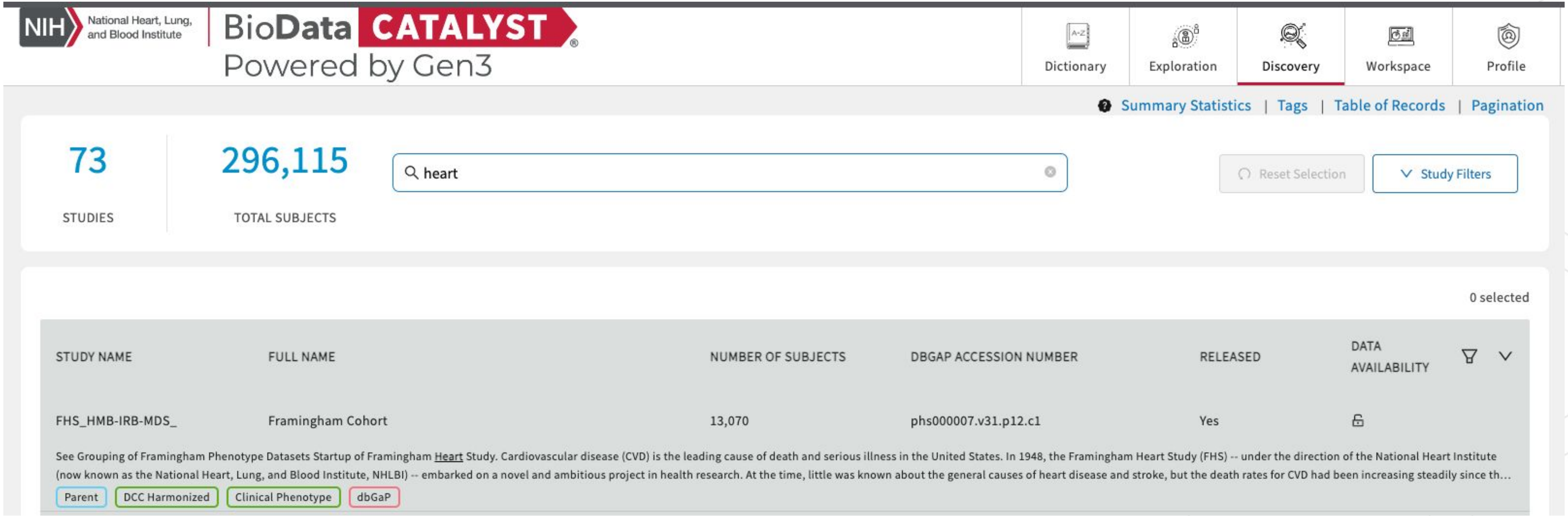
1. Source of Truth - File Object Persistence & Dataset Metadata
2. Interoperability - Standards-based integration points with other systems
3. Data Access - eRA Commons / dbGaP Authorization Inheritance
4. Data Ingestion - Robust data ingestion pipeline



NHLBI BioData Catalyst supports the management, analysis and sharing of human disease data for the research community and aims to advance basic understanding of the genetic basis of complex traits and accelerate discovery and development of therapies, diagnostic tests, and other technologies for diseases like cancer. The data commons supports cross-project analyses by harmonizing data from different projects through the collaborative development of a data dictionary, providing an API for data queries and download, and providing a cloud-based analysis workspace with rich tools and resources.

Gen3 - Discovery Page

A tool for discovery of released datasets (fully open, no required approval to discovery available data).



73 STUDIES

296,115 TOTAL SUBJECTS

heart

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...

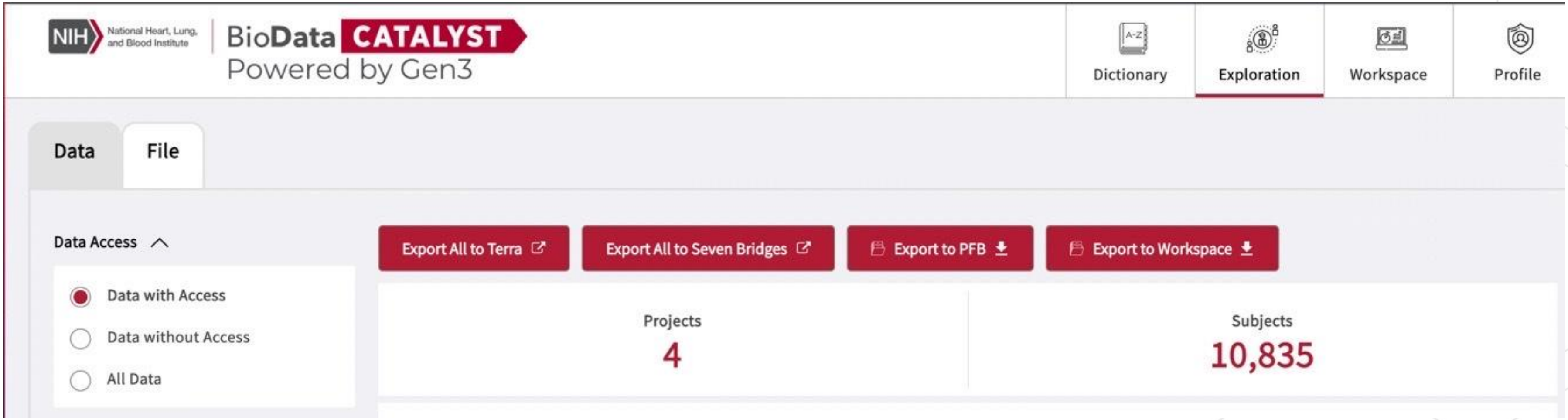
Parent | DCC Harmonized | Clinical Phenotype | dbGaP

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).



○ Data with Access
○ Data without Access
○ All Data

Projects
4

Subjects
10,835

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

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.

Submitting a Data Access Request (DAR)

Components of a DAR:

- Research Use Statement (2200 characters)
- Non-technical Summary (1100 characters)
- BioData Catalyst-specific Cloud Use Statement [Template language available]

For more information, step-by-step instructions, and template language, visit the "Submitting a dbGaP Data Access Request" page of the BDC documentation.

Live Demo:
PIC-SURE
Authorized Access

Breakout room

In your breakout rooms:

1. **Try:** exploring data in PIC-SURE. If you are not authorized to access data, explore available data in Open Access.
2. **Discuss:** How could PIC-SURE be useful in your research process?

Feel free to ask TAs any questions.

Breakout rooms - converge

Would someone like to discuss their PIC-SURE use case?

How could PIC-SURE be helpful in your research process?

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 environment
- Discover some available tools and workflows

Analysis Platforms

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

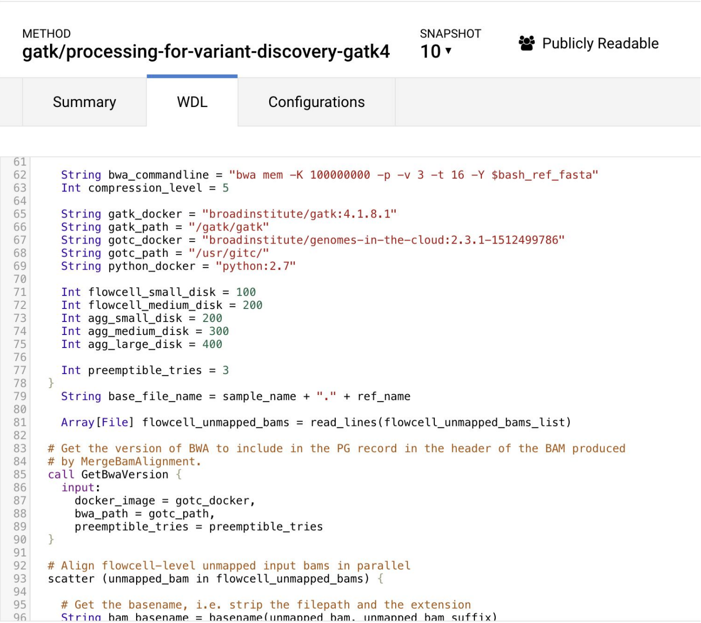
NIH National Heart, Lung, and Blood Institute | BioData CATALYST®
Powered by Terra

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)

BDC-Terra

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

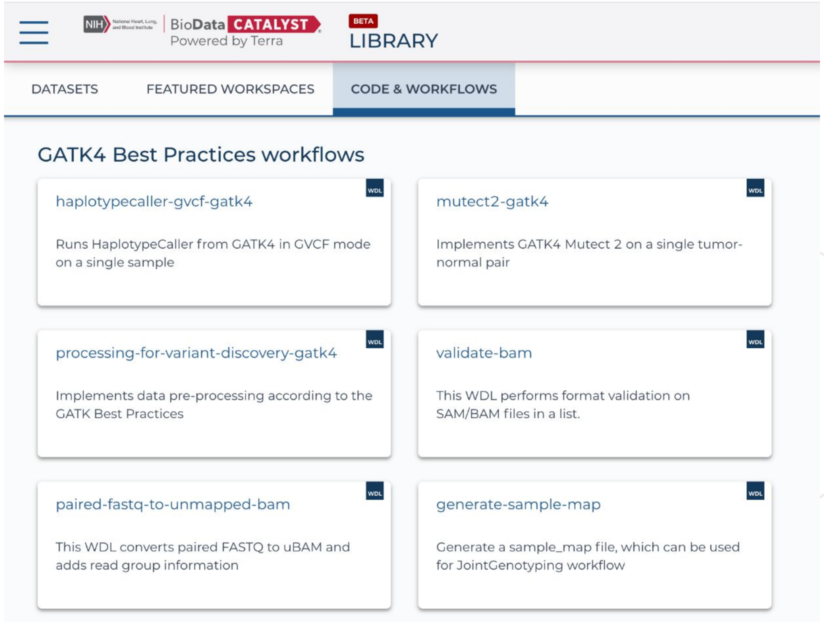


METHOD
gatk/processing-for-variant-discovery-gatk4

SNAPSHOT
10 ▾

Publicly Readable

```
    String bwa_commandline = "bwa mem -K 100000000 -p -v 3 -t 16 -Y $bash_ref_fasta"
    Int compression_level = 5

    String gatk_docker = "broadinstitute/gatk:4.1.8.1"
    String gatk_path = "/gatk/gatk"
    String gotc_docker = "broadinstitute/genomes-in-the-cloud:2.3.1-1512499786"
    String gotc_path = "/usr/gitc/"
    String python_docker = "python:2.7"

    Int flowcell_small_disk = 100
    Int flowcell_medium_disk = 200
    Int agg_small_disk = 200
    Int agg_medium_disk = 300
    Int agg_large_disk = 400

    Int preemptible_tries = 3
  }
    String base_file_name = sample_name + "." + ref_name

    Array[File] flowcell_unmapped_bams = read_lines(flowcell_unmapped_bams_list)

  # Get the version of BWA to include in the PG record in the header of the BAM produced
  # by MergeBamAlignment.
  call GetBwaVersion {
    input:
      docker_image = gotc_docker,
      bwa_path = gotc_path,
      preemptible_tries = preemptible_tries
  }

  # Align flowcell-level unmapped input bams in parallel
  scatter (unmapped_bam in flowcell_unmapped_bams) {

    # Get the basename, i.e. strip the filepath and the extension
    String bam_basename = basename(unmapped_bam, unmapped_bam_suffix)
```



GATK4 Best Practices workflows

haplotypecaller-gvcf-gatk4 (WDL)
Runs HaplotypeCaller from GATK4 in GVCF mode on a single sample

mutect2-gatk4 (WDL)
Implements GATK4 Mutect 2 on a single tumor-normal pair

processing-for-variant-discovery-gatk4 (WDL)
Implements data pre-processing according to the GATK Best Practices

validate-bam (WDL)
This WDL performs format validation on SAM/BAM files in a list.

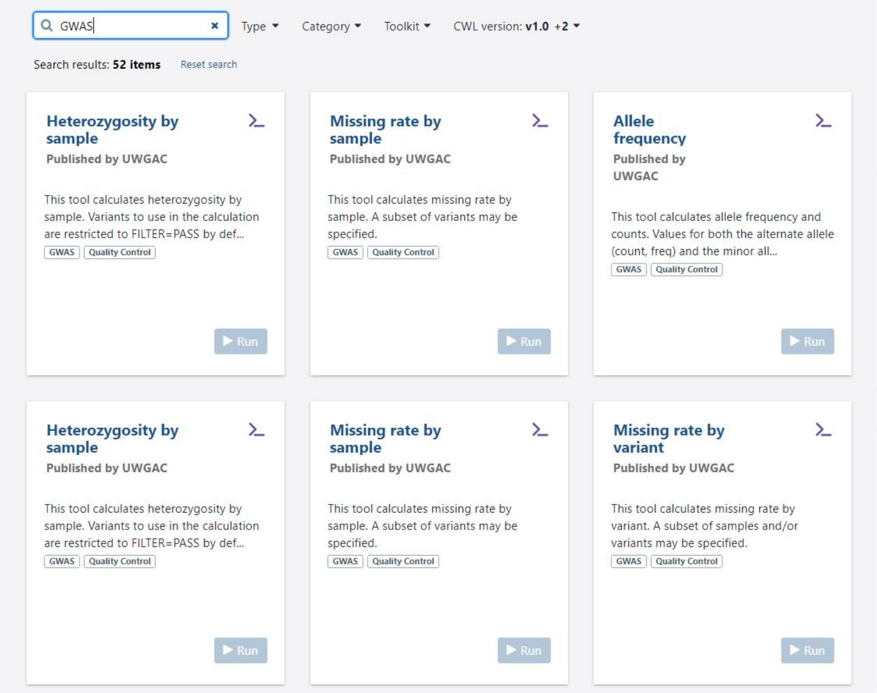
paired-fastq-to-unmapped-bam (WDL)
This WDL converts paired FASTQ to uBAM and adds read group information

generate-sample-map (WDL)
Generate a sample_map file, which can be used for JointGenotyping workflow

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



Search results: **52 items** Reset search

Heterozygosity by sample
Published by UWGAC
This tool calculates heterozygosity by sample. Variants to use in the calculation are restricted to FILTER=PASS by def...
GWAS | Quality Control

Missing rate by sample
Published by UWGAC
This tool calculates missing rate by sample. A subset of variants may be specified.
GWAS | Quality Control

Allele frequency
Published by UWGAC
This tool calculates allele frequency and counts. Values for both the alternate allele (count, freq) and the minor all...
GWAS | Quality Control

Heterozygosity by sample
Published by UWGAC
This tool calculates heterozygosity by sample. Variants to use in the calculation are restricted to FILTER=PASS by def...
GWAS | Quality Control

Missing rate by sample
Published by UWGAC
This tool calculates missing rate by sample. A subset of variants may be specified.
GWAS | Quality Control

Missing rate by variant
Published by UWGAC
This tool calculates missing rate by variant. A subset of samples and/or variants may be specified.
GWAS | Quality Control

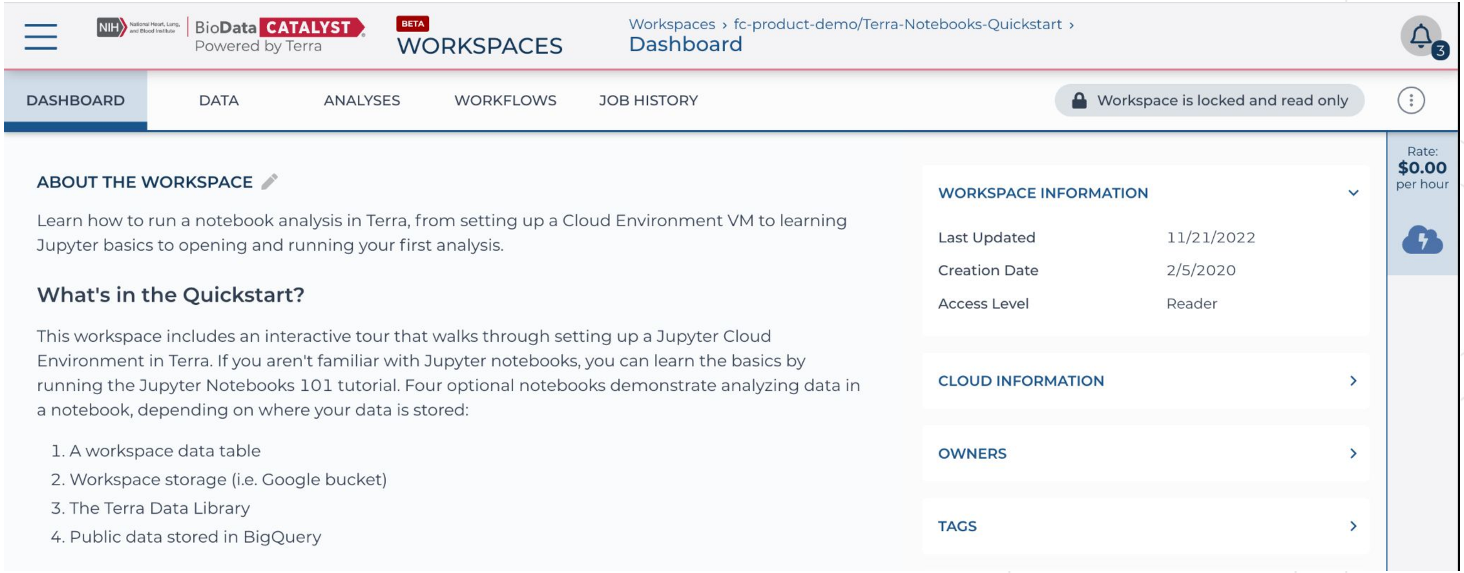
Open to the public @ platform.sb.biodatacatalyst.nhlbi.nih.gov/public/apps

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.

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.



NIH National Heart, Lung, and Blood Institute
BioData CATALYST
Powered by Terra
BETA
WORKSPACES
Workspaces › fc-product-demo/Terra-Notebooks-Quickstart ›
Dashboard
3
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
Workspace is locked and read only
Rate:
$0.00
per hour

ABOUT THE WORKSPACE

Learn how to run a notebook analysis in Terra, from setting up a Cloud Environment VM to learning Jupyter basics to opening and running your first analysis.

What's in the Quickstart?

This workspace includes an interactive tour that walks through setting up a Jupyter Cloud Environment in Terra. If you aren't familiar with Jupyter notebooks, you can learn the basics by running the Jupyter Notebooks 101 tutorial. Four optional notebooks demonstrate analyzing data in a notebook, depending on where your data is stored:

1. A workspace data table
2. Workspace storage (i.e. Google bucket)
3. The Terra Data Library
4. Public data stored in BigQuery

WORKSPACE INFORMATION

Last Updated	11/21/2022
Creation Date	2/5/2020
Access Level	Reader

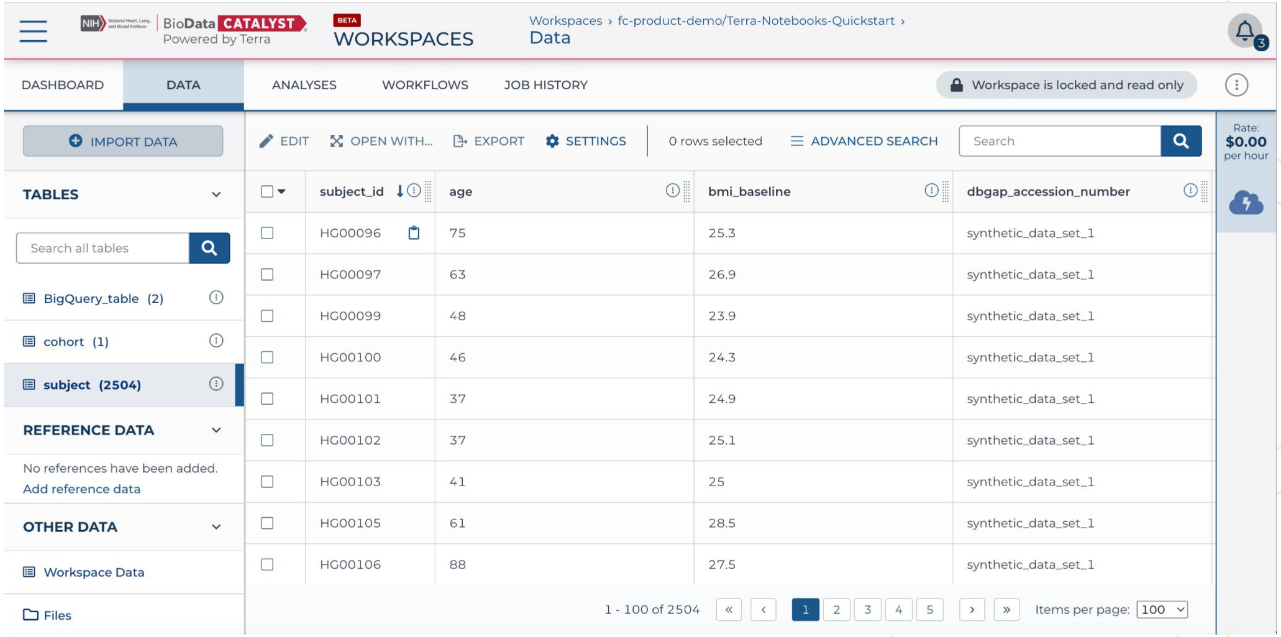
CLOUD INFORMATION ›

OWNERS ›

TAGS ›

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.



No references have been added.
Add reference data

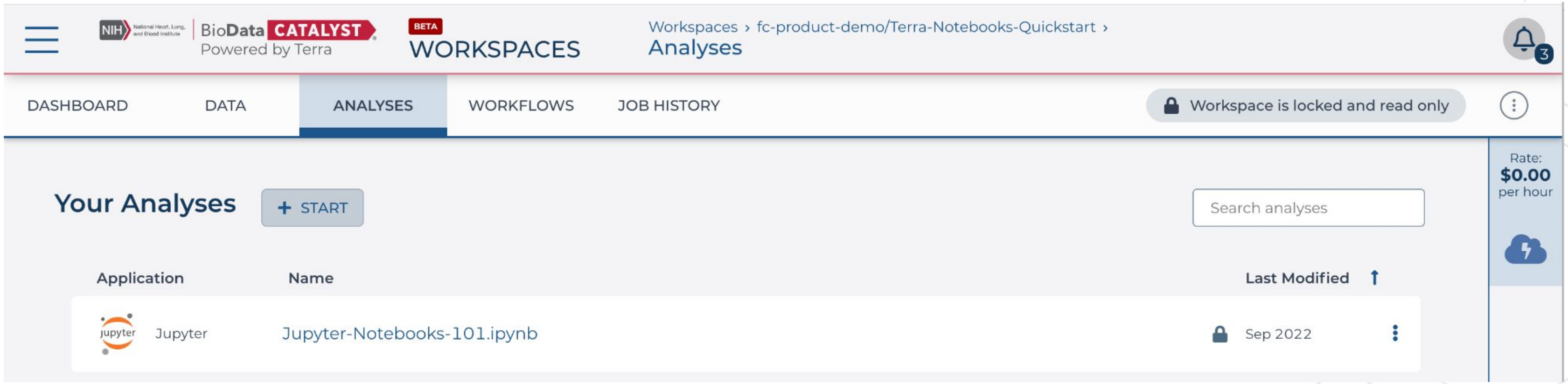
0 rows selected

subject_id	age	bmi_baseline	dbgap_accession_number
HG00096	75	25.3	synthetic_data_set_1
HG00097	63	26.9	synthetic_data_set_1
HG00099	48	23.9	synthetic_data_set_1
HG00100	46	24.3	synthetic_data_set_1
HG00101	37	24.9	synthetic_data_set_1
HG00102	37	25.1	synthetic_data_set_1
HG00103	41	25	synthetic_data_set_1
HG00105	61	28.5	synthetic_data_set_1
HG00106	88	27.5	synthetic_data_set_1

1 - 100 of 2504 « ‹ 1 2 3 4 5 › » Items per page: 100

BDC-Terra

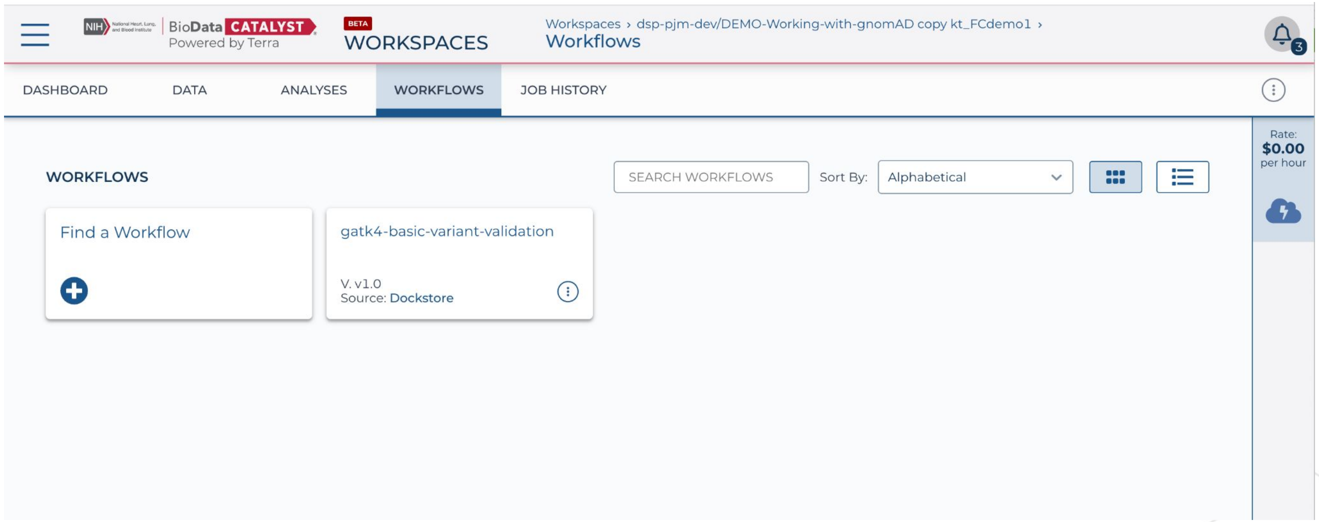
- Analyses
  - Interrogate and visualize your data in real time using Galaxy, Jupyter Notebooks, or RStudio
  - All three apps run on virtual machines or clusters of machines in a workspace Cloud Environment.



Application	Name	Last Modified
Jupyter	Jupyter-Notebooks-101.ipynb	Sep 2022

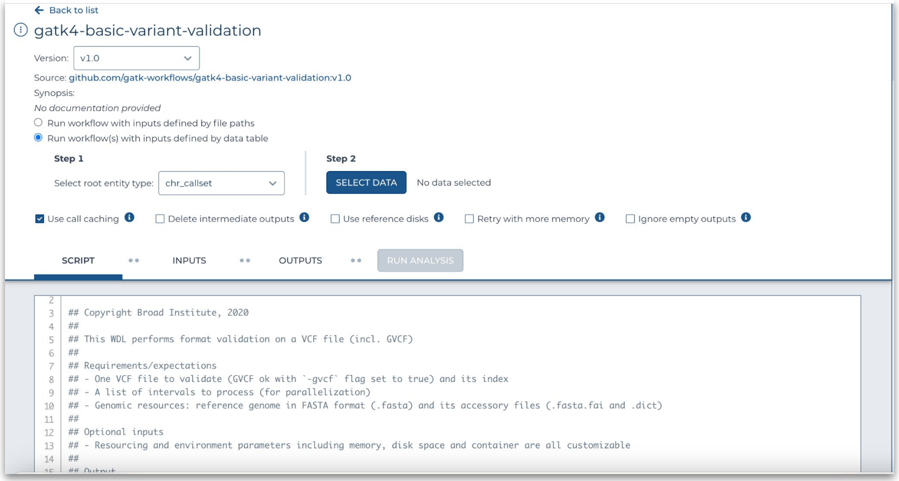
BDC-Terra

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



BDC-Terra

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



← Back to list

gatk4-basic-variant-validation

Version: v1.0

Source: github.com/gatk-workflows/gatk4-basic-variant-validation:v1.0

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

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

Step 1

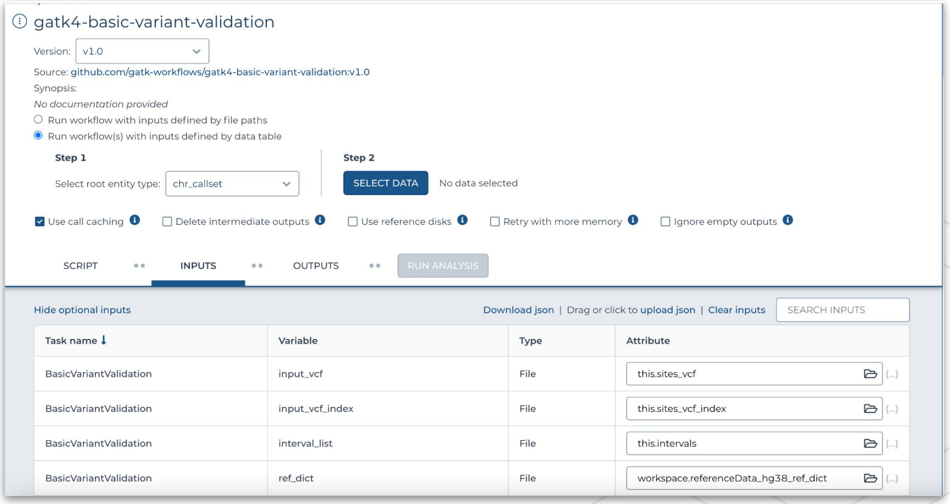
Select root entity type: chr_callset

Step 2

SELECT DATA No data selected

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

```

## Copyright Broad Institute, 2020
##
## This WDL performs format validation on a VCF file (incl. GVCF)
##
## Requirements/expectations
## - One VCF file to validate (GVCF ok with `-gvcf` flag set to true) and its index
## - A list of intervals to process (for parallelization)
## - Genomic resources: reference genome in FASTA format (.fasta) and its accessory files (.fasta.fai and .dict)
##
## Optional inputs
## - Resourcing and environment parameters including memory, disk space and container are all customizable
##
## Output
```



gatk4-basic-variant-validation

Version: v1.0

Source: github.com/gatk-workflows/gatk4-basic-variant-validation:v1.0

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: chr_callset

Step 2

SELECT DATA No data selected

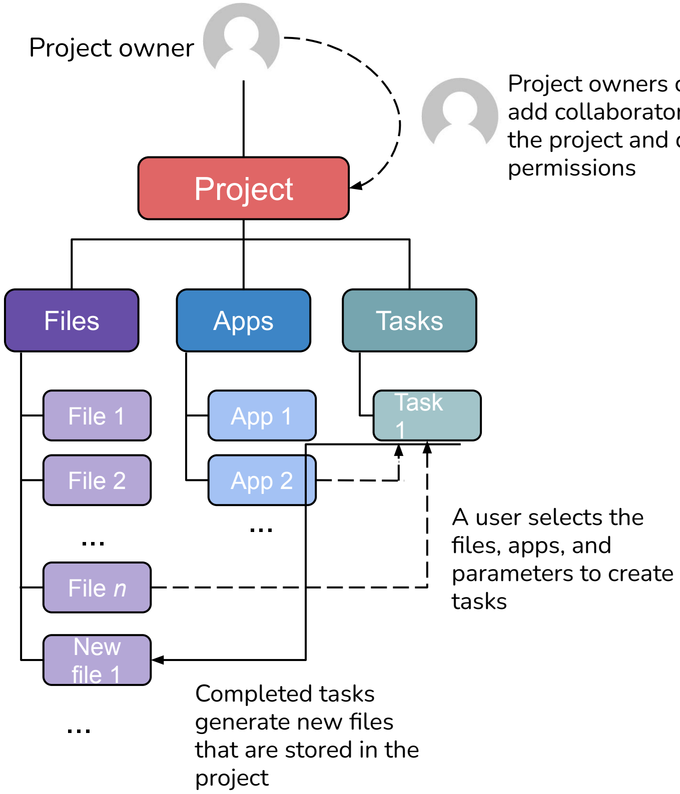
☑ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks ☐ Retry with more memory ☐ Ignore empty outputs

Task name ↓	Variable	Type	Attribute
BasicVariantValidation	input_vcf	File	this.sites_vcf
BasicVariantValidation	input_vcf_index	File	this.sites_vcf_index
BasicVariantValidation	interval_list	File	this.intervals
BasicVariantValidation	ref_dict	File	workspace.referenceData_hg38_ref_dict

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

Seven Bridges - Projects organize files, methods, and results



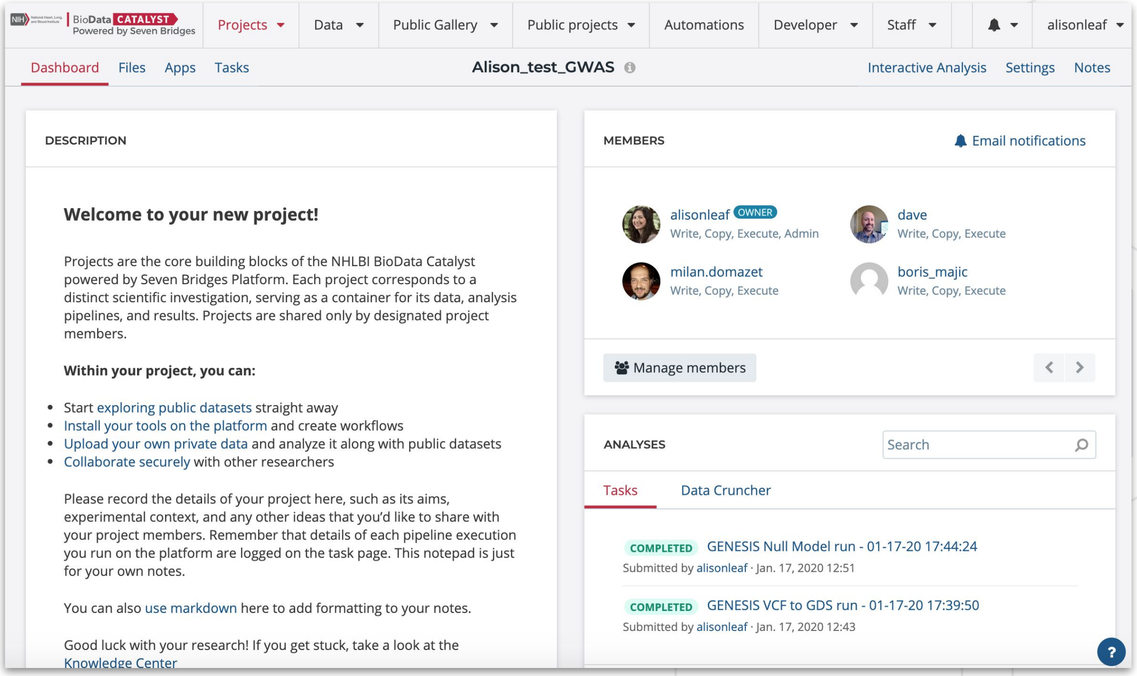
Project owners can add collaborators to the project and define permissions

A user selects the files, apps, and parameters to create tasks

Completed tasks generate new files that are stored in the project

Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions



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

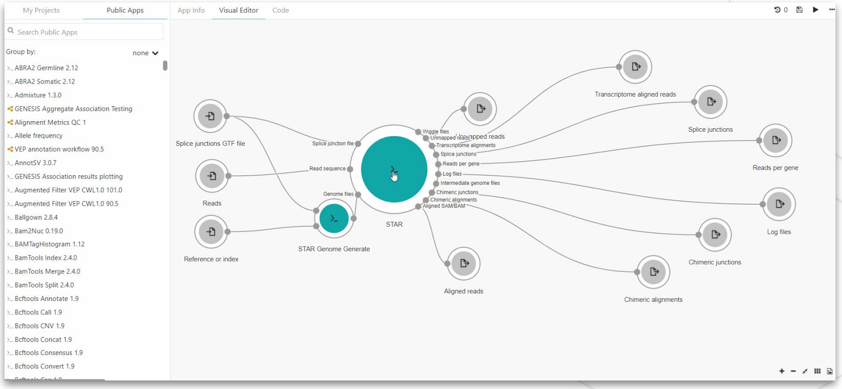
Common Workflow Language enables **portability**, **reproducibility**, and **scalability**

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

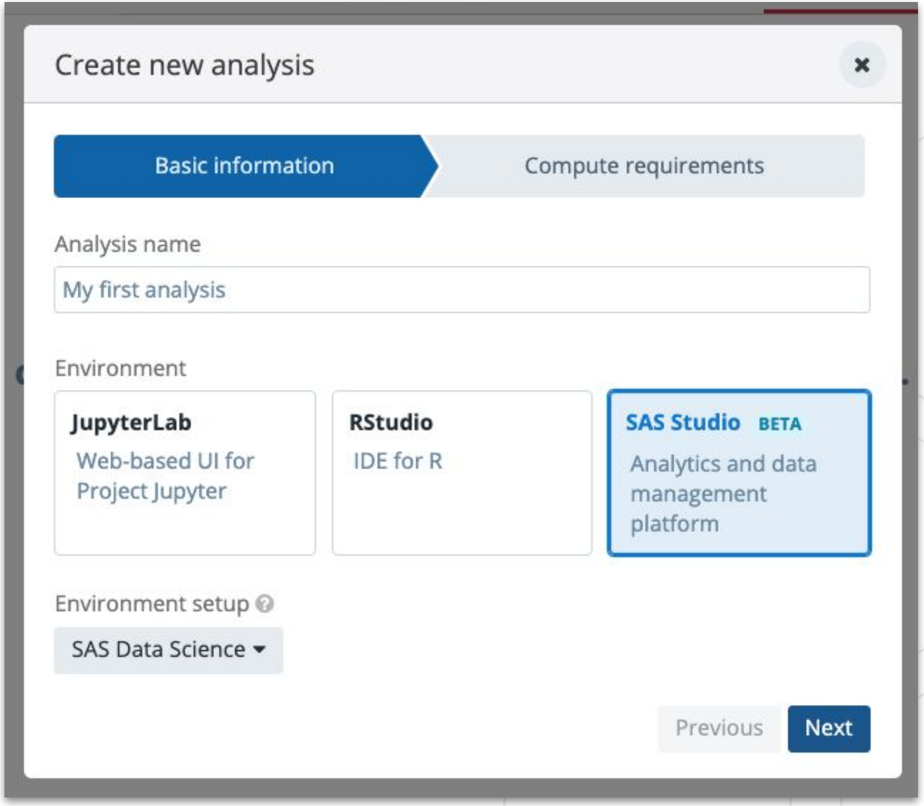
Expose or lock parameters appropriately



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.



Create new analysis

Analysis name: My first analysis

Environment

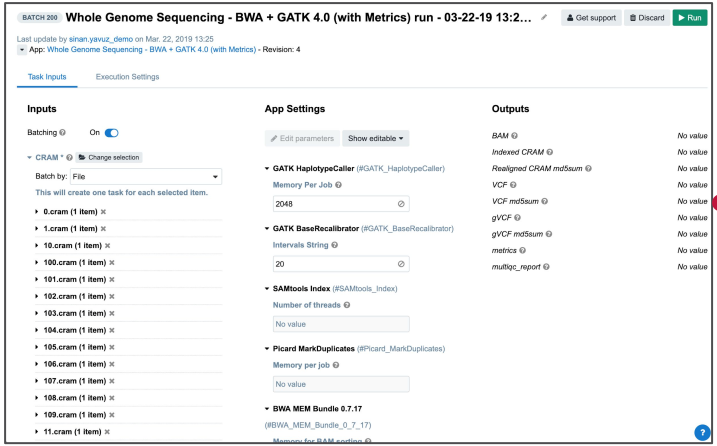
- JupyterLab: Web-based UI for Project Jupyter
- RStudio: IDE for R
- SAS Studio BETA: Analytics and data management platform

Environment setup: SAS Data Science

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 200 **Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 03-22-19 13:2...**

Last update by sinan.yavuz_demo on Mar. 22, 2019 13:25

App: Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) - Revision: 4

Inputs

Batching: On

CRAM * — Change selection

Batch by: File

This will create one task for each selected item.

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

App Settings

GATK HaplotypeCaller (#GATK_HaplotypeCaller)
Memory Per Job: 2048

GATK BaseRecalibrator (#GATK_BaseRecalibrator)
Intervals String: 20

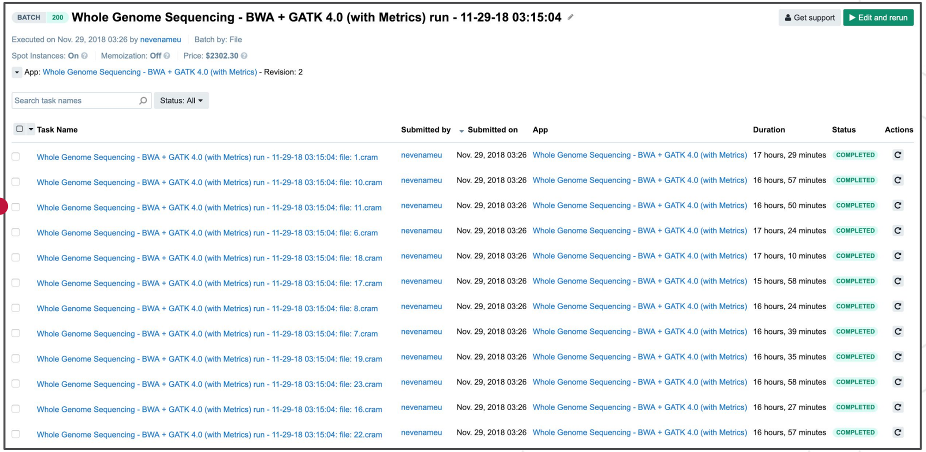
SAMtools Index (#SAMtools_Index)
Number of threads: No value

Picard MarkDuplicates (#Picard_MarkDuplicates)
Memory per job: No value

BWA MEM Bundle 0.7.17 (#BWA_MEM_Bundle_0_7_17)
Memory for BAM sorting

Outputs

BAM	No value
Indexed CRAM	No value
Realigned CRAM md5sum	No value
VCF	No value
VCF md5sum	No value
gVCF	No value
gVCF md5sum	No value
metrics	No value
multiqc_report	No value



BATCH 200 **Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04**

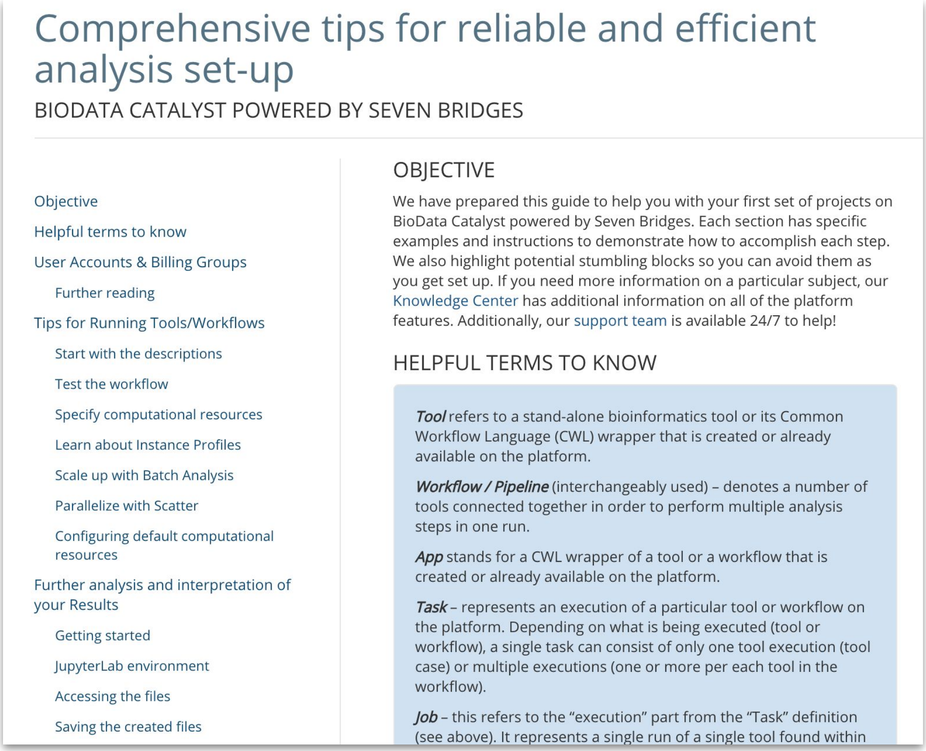
Executed on Nov. 29, 2018 03:26 by nevenameu | Batch by: File

Spot Instances: **On** | Memoization: **Off** | Price: **$2302.30**

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

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

Detailed documentation and tutorials



Comprehensive tips for reliable and efficient analysis set-up

BIODATA CATALYST POWERED BY SEVEN BRIDGES

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

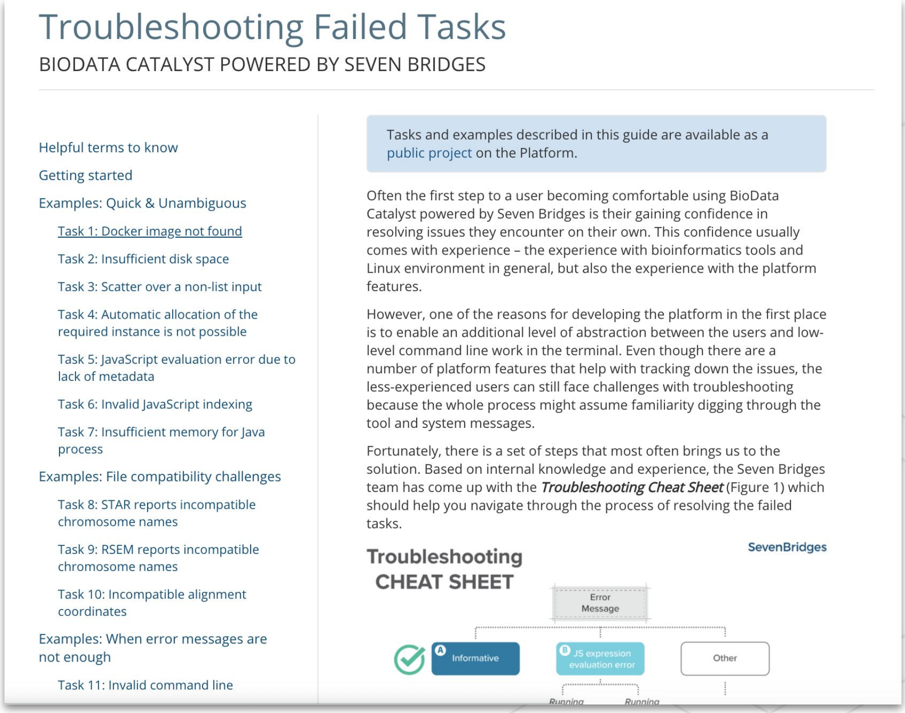
Tool refers to a stand-alone bioinformatics tool or its Common Workflow Language (CWL) wrapper that is created or already available on the platform.

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

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.

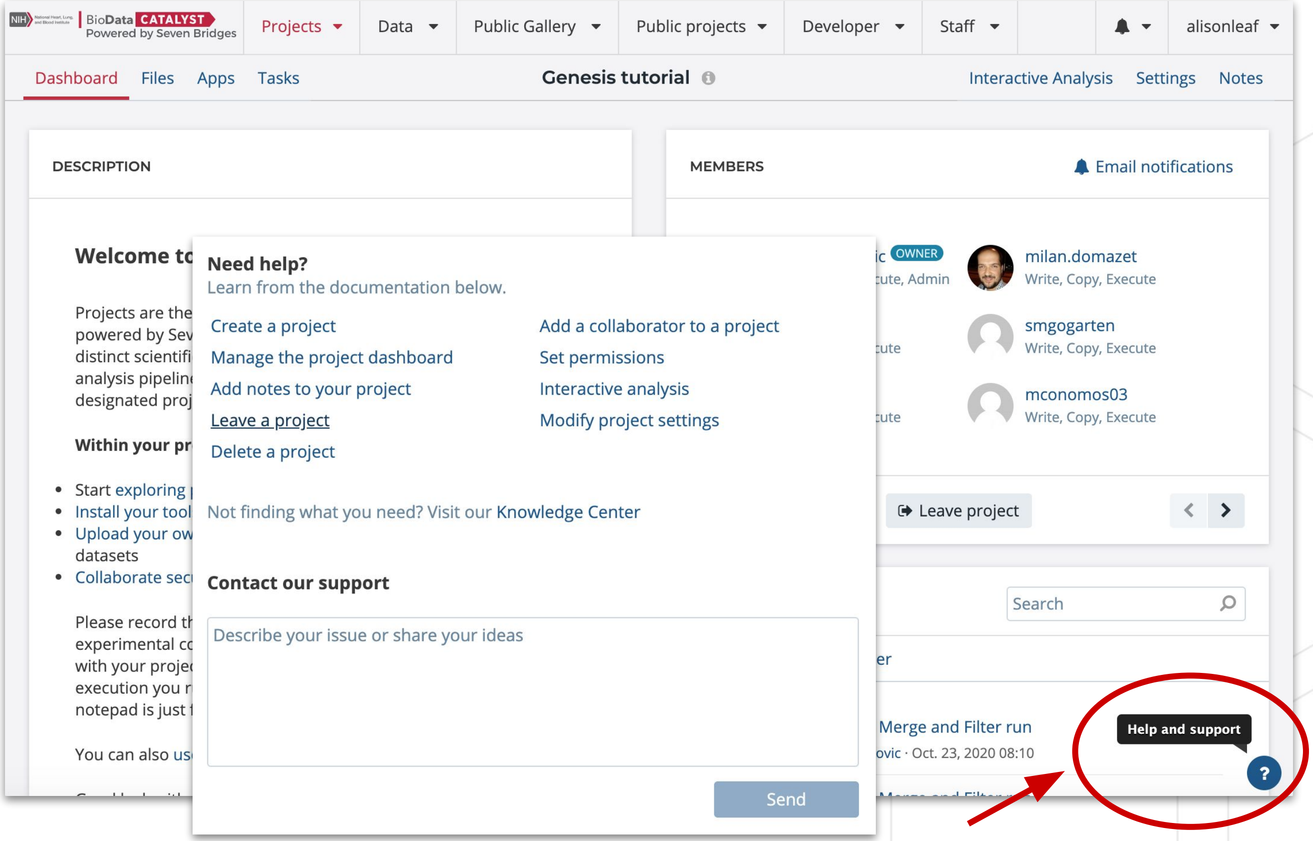
However, one of the reasons for developing the platform in the first place is to enable an additional level of abstraction between the users and low-level command line work in the terminal. Even though there are a number of platform features that help with tracking down the issues, the less-experienced 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.

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.



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
- Add a collaborator to a project
- Set permissions
- Interactive analysis
- Modify project settings

Not finding what you need? Visit our Knowledge Center

Contact our support

Describe your issue or share your ideas

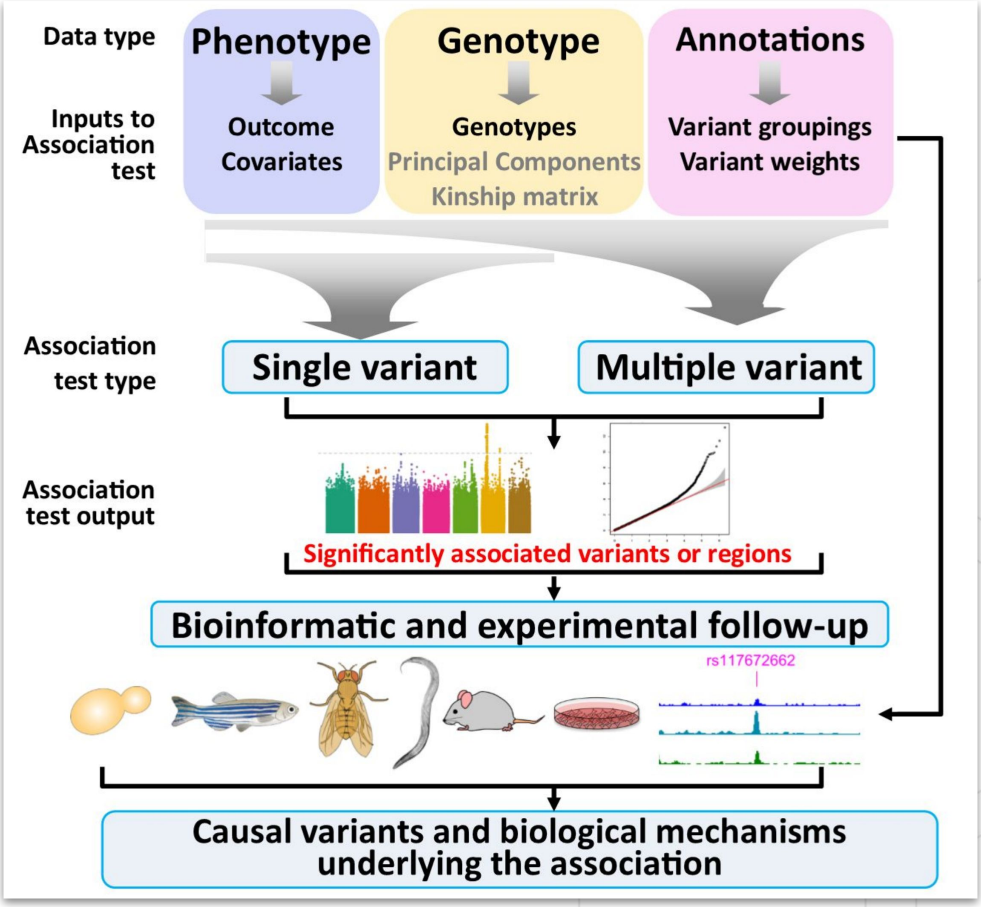
Send

Whole person: integrating social, environmental, and genetic factors

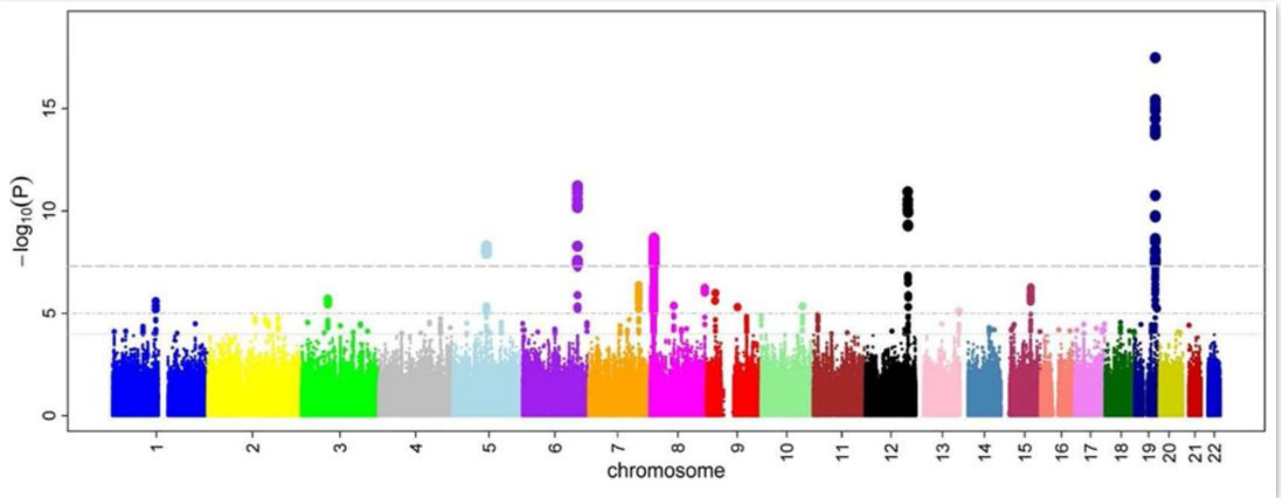
GWAS is a great place to start

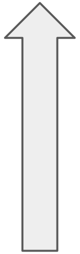
Genome wide association study

- Method that helps scientists identify genes associated with a particular disease or trait
- Used frequently on BDC
- Identifies signals of significance but further experimental follow up required to understand the functional biology



Manhattan plot is a main GWAS output





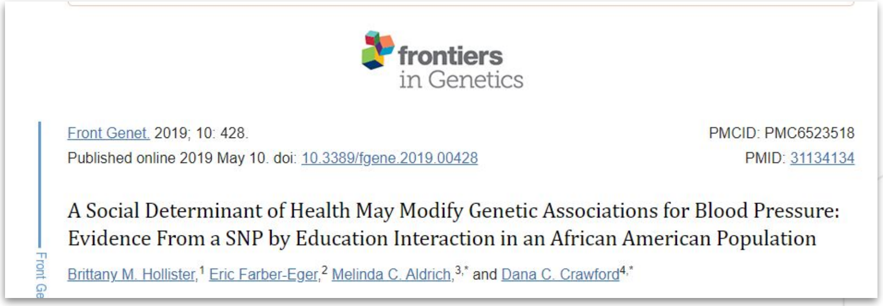
P-value gets smaller as “height” increases

—

SNP Significance gets higher as “height” increases

Interactions between Genetics and SDoH

- **Social determinants of health**
  - Income and social protection
  - Education
  - Unemployment and job insecurity
  - Working life conditions
  - Food insecurity
  - Housing, basic amenities and the environment
  - Early childhood development
  - Social inclusion and non-discrimination
  - Structural conflict
  - Access to affordable health services of decent quality



frontiers in Genetics

Front Genet. 2019; 10: 428.
Published online 2019 May 10. doi: 10.3389/fgene.2019.00428

PMCID: PMC6523518
PMID: 31134134

A Social Determinant of Health May Modify Genetic Associations for Blood Pressure: Evidence From a SNP by Education Interaction in an African American Population

Brittany M. Hollister,[1] Eric Farber-Eger,[2] Melinda C. Aldrich,[3,*] and Dana C. Crawford[4,*]

Link to pubmed entry

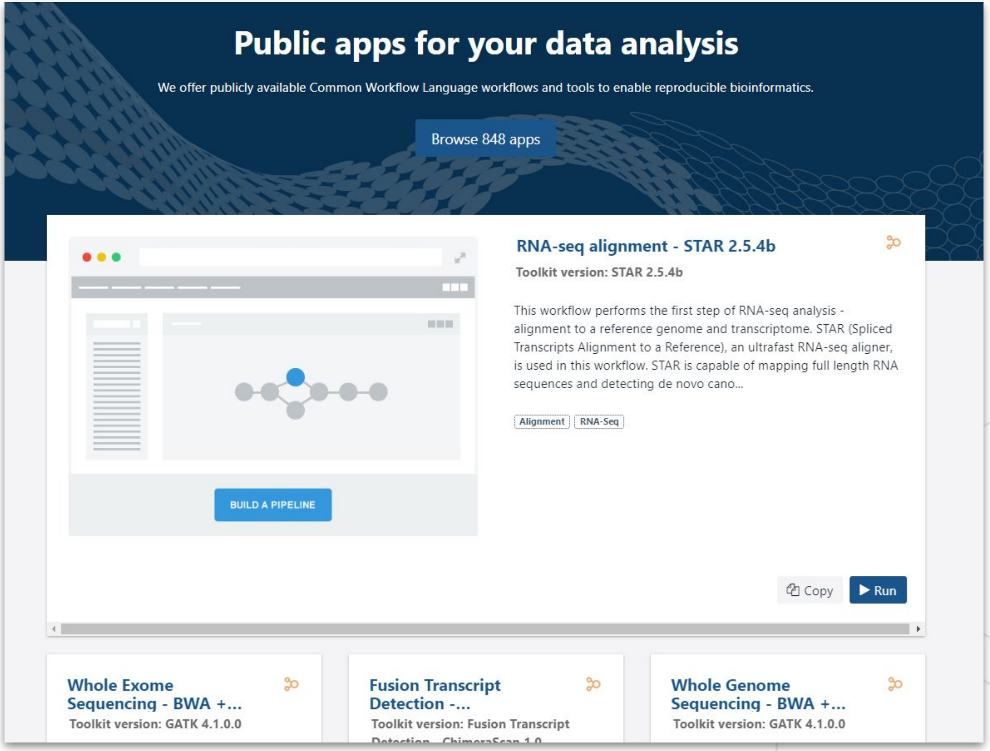
SDoH can be used as covariates in GWAS?

Run association pipelines out of the box

- **GENESIS**
- Plink
- EPACTS
- STAAR

No login required!

Publish your apps to share with the world!



Public apps for your data analysis

We offer publicly available Common Workflow Language workflows and tools to enable reproducible bioinformatics.

RNA-seq alignment - STAR 2.5.4b

Toolkit version: STAR 2.5.4b

This workflow performs the first step of RNA-seq analysis - alignment to a reference genome and transcriptome. STAR (Spliced Transcripts Alignment to a Reference), an ultrafast RNA-seq aligner, is used in this workflow. STAR is capable of mapping full length RNA sequences and detecting de novo cano...

Whole Exome Sequencing - BWA +...

Toolkit version: GATK 4.1.0.0

Fusion Transcript Detection -...

Toolkit version: Fusion Transcript Detection - ChimeraScan 1.0

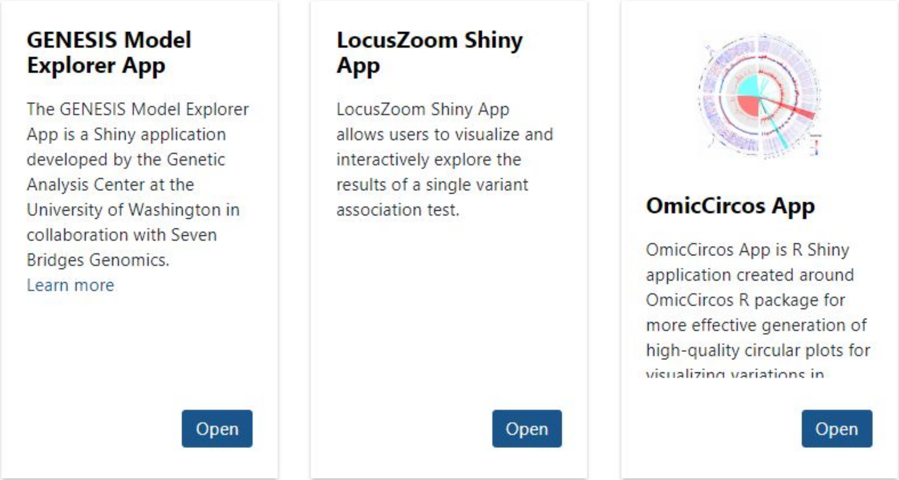
Whole Genome Sequencing - BWA +...

Toolkit version: GATK 4.1.0.0

https://platform.sb.biodatacatalyst.nhlbi.nih.gov/public/apps

GWAS Additional Features

- Annotation Explorer
  - Use for prep for Aggregate test
  - Post GWAS annotations
- Aggregate and Sliding window association test apps
- Data Overview feature
  - Explore variant frequency across TOPMed Freeze8 studies on GRCh38
- Study Variable Explorer
  - Search for annotated variables across studies and manually annotate and compare



GENESIS Model Explorer App

The GENESIS Model Explorer App is a Shiny application developed by the Genetic Analysis Center at the University of Washington in collaboration with Seven Bridges Genomics.

Learn more

Open

LocusZoom Shiny App

LocusZoom Shiny App allows users to visualize and interactively explore the results of a single variant association test.

Open

OmicCircos App

OmicCircos App is R Shiny application created around OmicCircos R package for more effective generation of high-quality circular plots for visualizing variations in

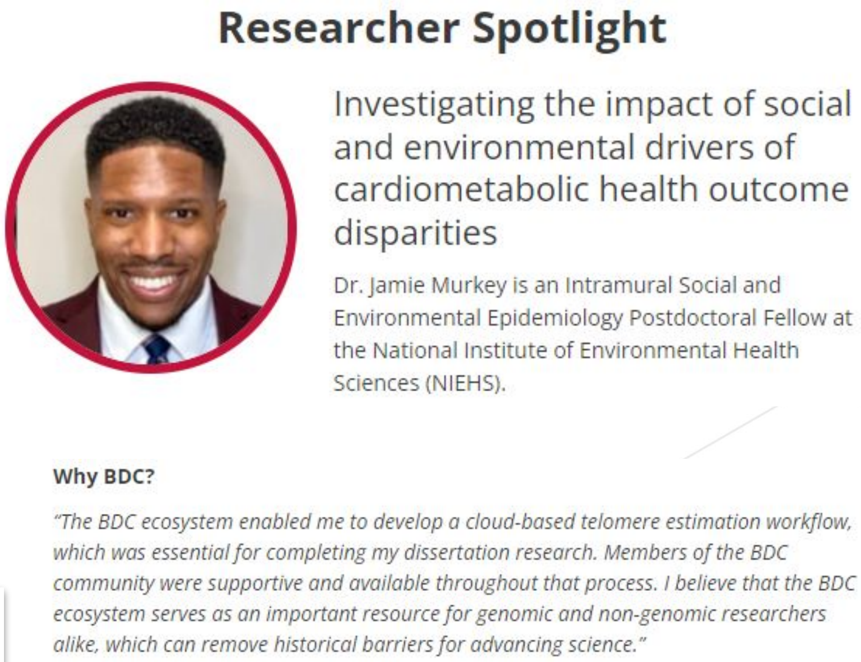
Open

Researcher Spotlight: Dr. Jamie Murkey

Primary question: *What role do social factors play in influencing the pathophysiology of cardiovascular disease (CVD) and racial/ethnic disparities in CVD events?*

Aim 1: Perform a comparative assessment between two bioinformatics tools to develop and test a cloud-based telomere length estimation workflow.

He developed **TeleGen**, a bioinformatics workflow that generated telomere length data for 100% of TOPMed MESA Participants



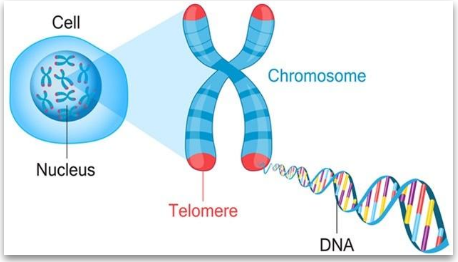
Researcher Spotlight

Investigating the impact of social and environmental drivers of cardiometabolic health outcome disparities

Dr. Jamie Murkey is an Intramural Social and Environmental Epidemiology Postdoctoral Fellow at the National Institute of Environmental Health Sciences (NIEHS).

Why BDC?

"The BDC ecosystem enabled me to develop a cloud-based telomere estimation workflow, which was essential for completing my dissertation research. Members of the BDC community were supportive and available throughout that process. I believe that the BDC ecosystem serves as an important resource for genomic and non-genomic researchers alike, which can remove historical barriers for advancing science."



[Linkedin Profile](#)

Live Demo: Seven Bridges

Breakout room

In your breakout rooms:

1. **Try:** Create your own project/sandbox on BDC-Seven Bridges
    a. Name project in this format - **NDSA_First_Last**
    b. Add your TA to the project
    c. Copy files, apps and JupyterLab from the main NDSA Workshop project
2. **Discuss:** What apps or software environments would you like to run on Seven Bridges for your **research** or to help with **teaching**

Feel free to ask TAs any questions.

Breakout rooms - converge

- What was the experience like getting your sandbox project setup?
- What software would you need for your particular use cases?