Introduction to BDC

Presented to the PRIDE Programs

June 13, 2023

By Ingrid Borecki, PhD

BDC Steering Committee Chair

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

Introductions and Housekeeping

Presenters

Ingrid Borecki
Chair Steering Committee,
Fellows Program lead

Emily Hughes
BDC Powered by PIC-SURE
Harvard Medical School

David Roberson
BDC Powered by Seven Bridges
Seven Bridges/Velsera

Live Support

Amber Voght
User Engagement Specialist,
BDC Coordinating Center

Kaleena Narwani
User Engagement Specialist,
BDC Coordinating Center

Questions before we begin?

Agenda

Topic	Time
Motivation and Value Proposition	20 min
Data	10 min
Discovering Data on Open PIC-SURE Demo	15 min
Performing a GWAS on Seven Bridges Demo	15 min
Fellows Advice	10 min
Next Steps: Join the Community / Pilot Credits	10 min
Q&A	10 min

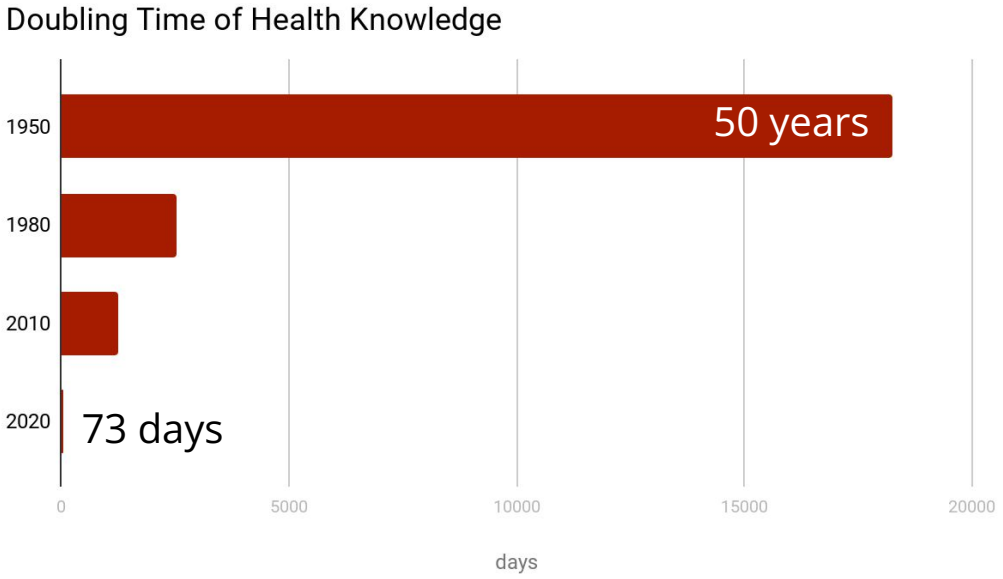
Question for you:

What is your research area?

Introduction to BDC

Ingrid Borecki, BDC Steering Committee Chair

The rate of data generation is accelerating rapidly.

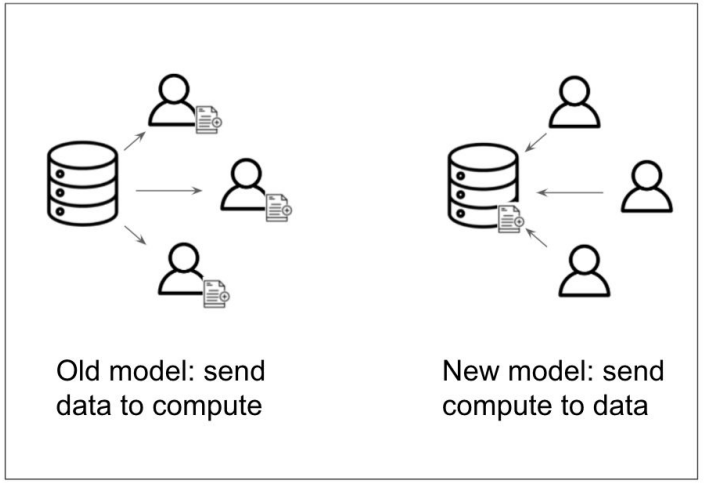


- More biomedical data will be generated this year than all previous years **combined**.
- Diverse data modalities including Health data, Survey, Sequencing, Imaging, Metabolomics, Proteomics, Sensor, E-Phys, Flow Cytometry etc.

Densen, P. Trans Am Clin Climatol Assoc 2011

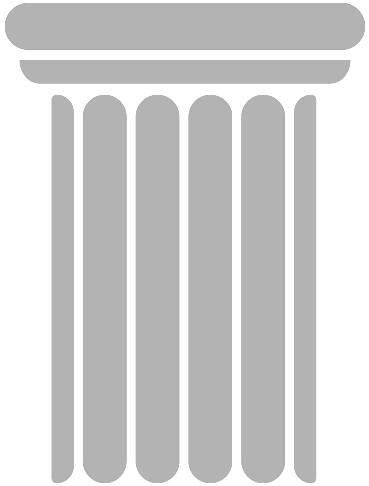
Cloud is the most effective way to store, access and analyze our growing health data.

- 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.

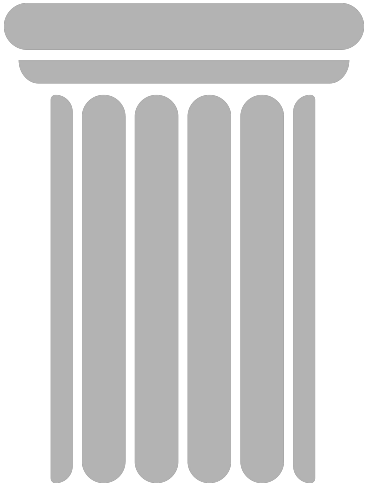


NHLBI BioData Catalyst®

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.

What BDC offers:

Managing the Computing Environment

Elastic Computing

Easier Access to many High Value Datasets

Tooling

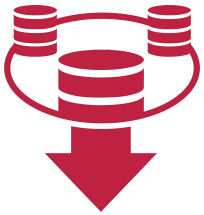
Data Discovery in PIC-SURE

Statistical Analysis Tools (R, SAS)

Other Specialized Workflows

Community and Peer Interactions

The Computing Environment

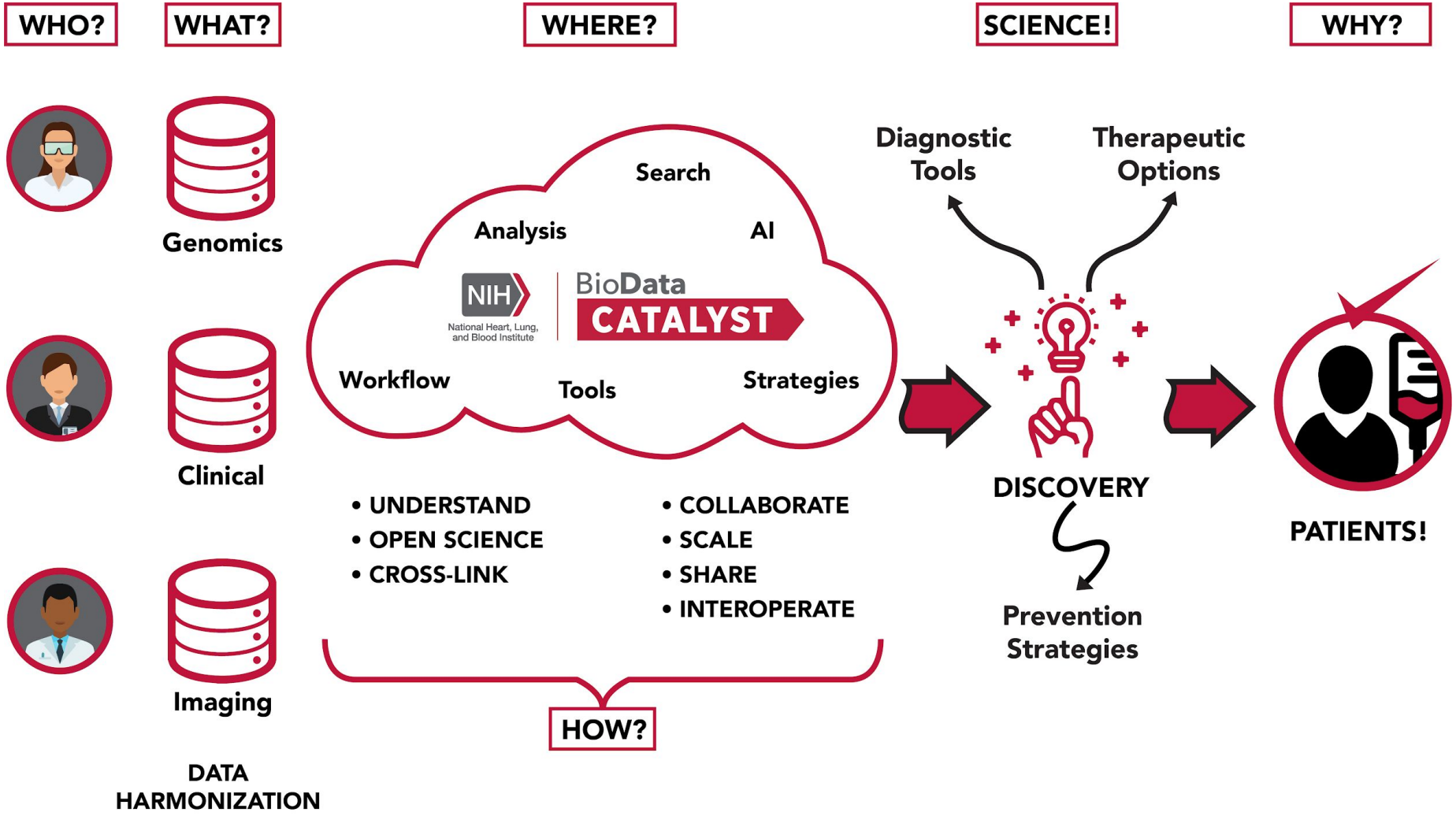


No need to **download** and **manage** (multiple) large datasets

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

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

Help desk and **documentation**

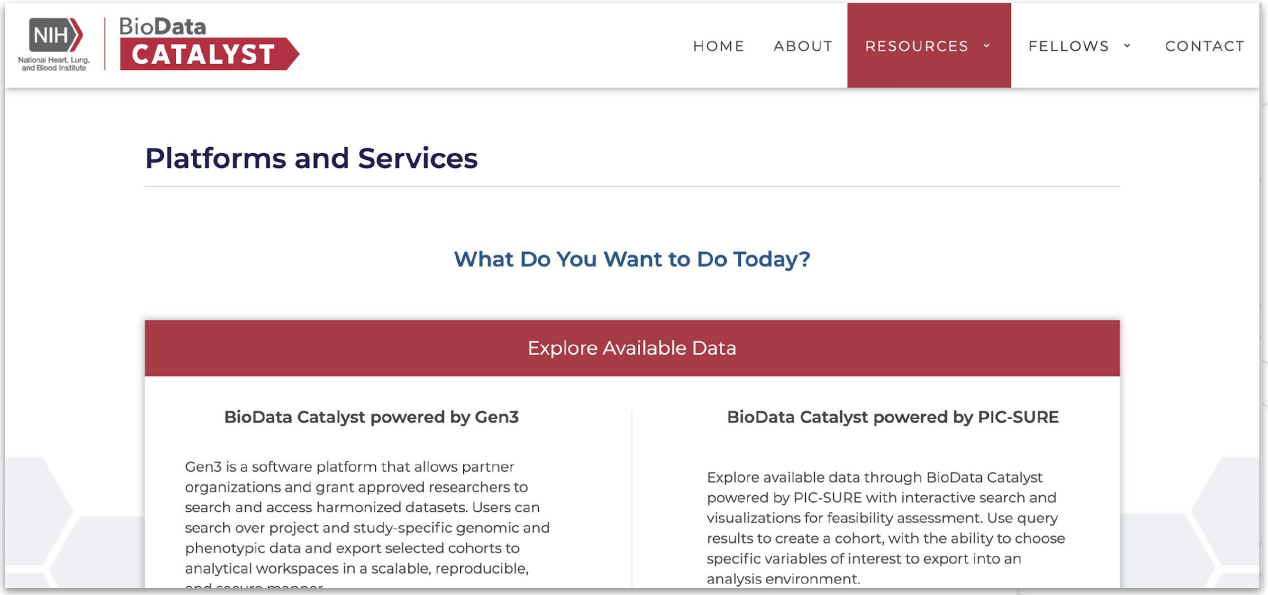

WHO?
WHAT?
WHERE?
SCIENCE!
WHY?
Genomics
Clinical
Imaging
DATA HARMONIZATION
Search
Analysis
AI
NIH
National Heart, Lung, and Blood Institute
BioData CATALYST
Workflow
Tools
Strategies
• UNDERSTAND
• OPEN SCIENCE
• CROSS-LINK
• COLLABORATE
• SCALE
• SHARE
• INTEROPERATE
HOW?
Diagnostic Tools
Therapeutic Options
DISCOVERY
Prevention Strategies
PATIENTS!

Platforms and Services

Explore Data
- PIC-SURE
- Gen3

Analyze Data
- Seven Bridges
- Terra

Community Tools
- Dockstore



Web resource: Services

Data

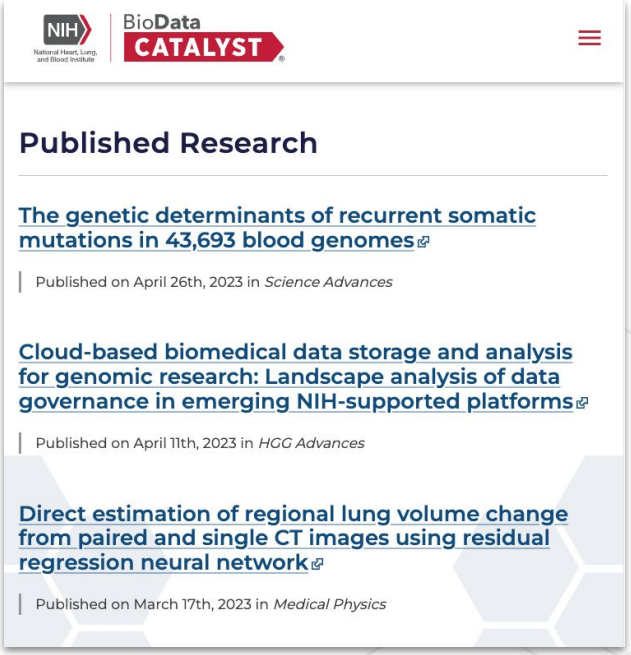
Emily Hughes, PIC-SURE

Question for you:

Do you have a dataset of interest?

Our Researchers are working on…

Sickle Cell Disease
Congenital Heart Disease
Coronary Artery Disease
Asthma
COVID
Obesity
Sleep
Cardiometabolic health
Fibrosis
Genomics
…and so much more!



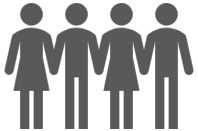
→ View Published Research Using BDC

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

Data Available in BDC

3.42
Petabytes of data

280,000+
Participants



490,000+
Data files

150,000+
Whole genomes

High-Value NHLBI datasets already ingested

TOPMed

COPD

COVID-19

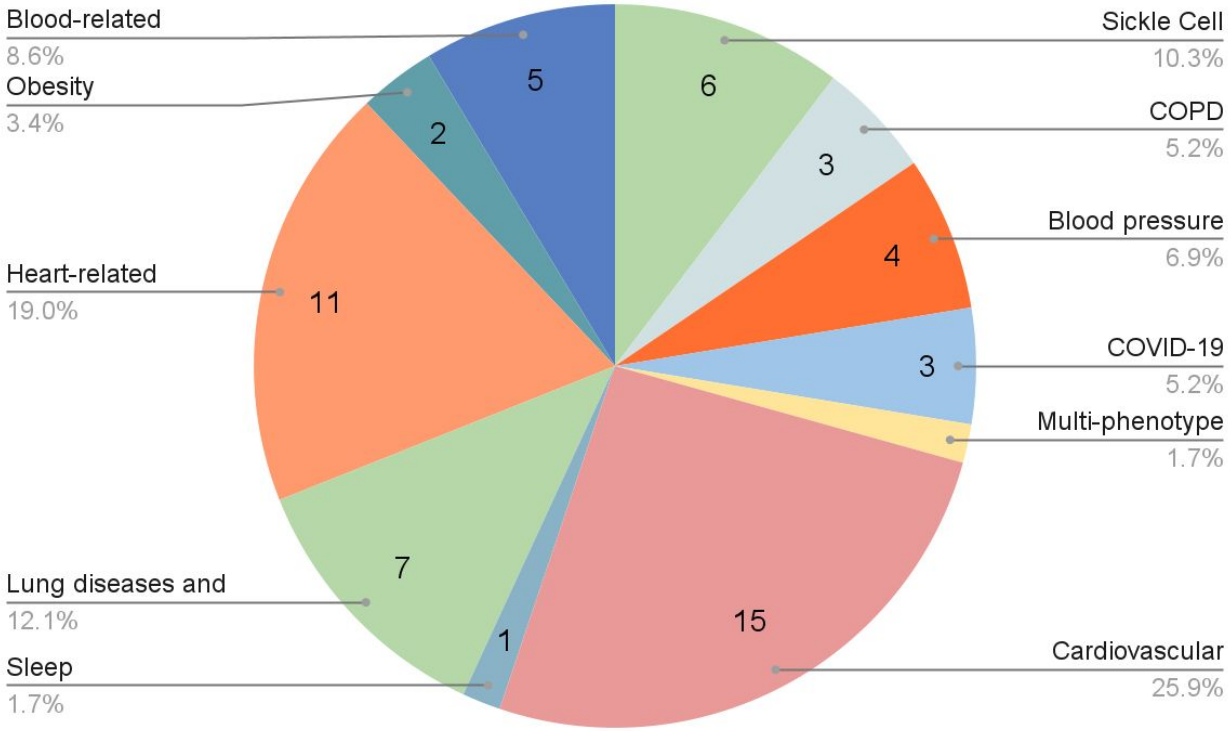
Sickle Cell Disease

1000 Genomes

Bring Your Own Data

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

Data Available in BDC



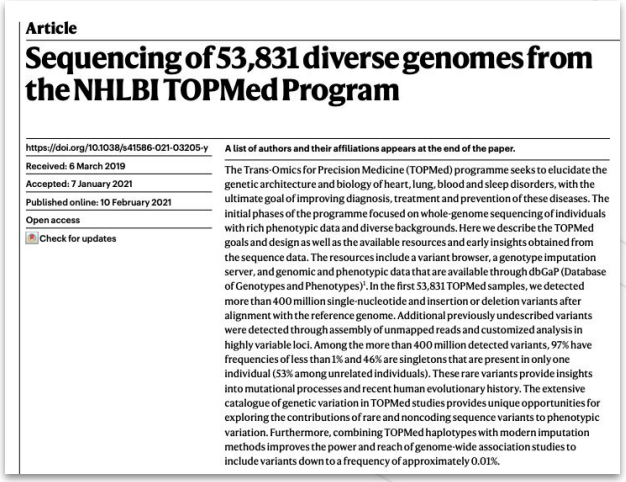
NHLBI's Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing (WGS) Data

TOPMed data has been released to the scientific community

Opportunities for Training and Research:

- >80 studies with study participants from diverse ancestries
- Data access on dbGAP; "General Research Use" or "Health/Medical/Biomedical" consents; No IRB required [Link to dbGap query]

Now, over 150,000 sequenced genomes!



Article

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program

https://doi.org/10.1038/s41586-021-03205-y
Received: 6 March 2019
Accepted: 7 January 2021
Published online: 10 February 2021
Open access
Check for updates

A list of authors and their affiliations appears at the end of the paper.

The Trans-Omics for Precision Medicine (TOPMed) programme seeks to elucidate the genetic architecture and biology of heart, lung, blood and sleep disorders, with the ultimate goal of improving diagnosis, treatment and prevention of these diseases. The initial phases of the programme focused on whole-genome sequencing of individuals with rich phenotypic data and diverse backgrounds. Here we describe the TOPMed goals and design as well as the available resources and early insights obtained from the sequence data. The resources include a variant browser, a genotype imputation server, and genomic and phenotypic data that are available through dbGaP (Database of Genotypes and Phenotypes)[1]. In the first 53,831 TOPMed samples, we detected more than 400 million single-nucleotide and insertion or deletion variants after alignment with the reference genome. Additional previously undescribed variants were detected through assembly of unmapped reads and customized analysis in highly variable loci. Among the more than 400 million detected variants, 97% have frequencies of less than 1% and 46% are singletons that are present in only one individual (53% among unrelated individuals). These rare variants provide insights into mutational processes and recent human evolutionary history. The extensive catalogue of genetic variation in TOPMed studies provides unique opportunities for exploring the contributions of rare and noncoding sequence variants to phenotypic variation. Furthermore, combining TOPMed haplotypes with modern imputation methods improves the power and reach of genome-wide association studies to include variants down to a frequency of approximately 0.01%.

Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299 (2021).

TOPMed WGS Data

Phenotypic

Harmonized data

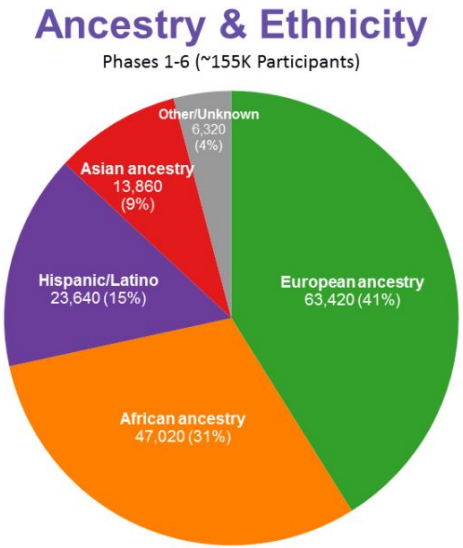
44 high-priority clinical and demographic variables have been harmonized by the TOPMed Data Coordinating Center (DCC).

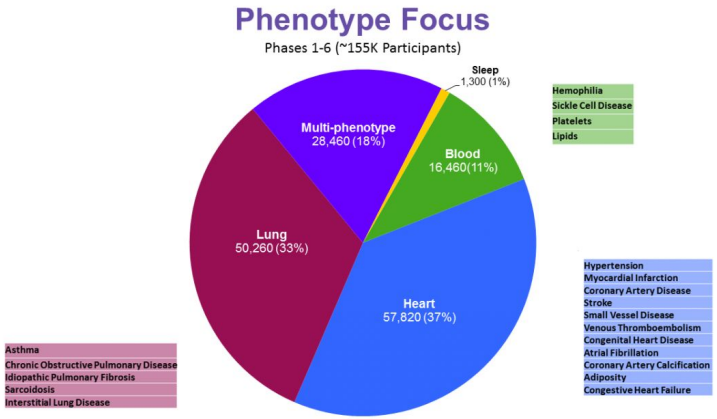
Non-harmonized data

The full set of raw clinical and phenotypic variables for the hosted studies are also available.

Genomic

Genomic data provided by the Trans-Omics for Precision Medicine (TOPMed) program, including CRAM and VCF files.





Studies Available on BDC (TOPMed)

AACAC
ACTIV-4A
ACTIV-4B
AFLMU
AMISH
ARIC
Asthma
AustralianFamilialAF
BAGS
BioMe
BioVU_AF
BostonBrazil_SCD
C3PO
CAMP
CARDIA
CARE_BADGER
CARE_CLIC
CARE_PACT
CARE_TREXA
CATHGEN
CCAF
CCAF_AF
CCDG_PMBB_AF
CFS
ChildrensHS_GAP
ChildrensHS_IGERA
ChildrensHS_MetaAir
CHIRAH
CHS
CIBMTR
CMG_WGS_HMB
COPDGene
CRA
CSSCD
DECAF
DHS
ECLIPSE
EGCUT
EOCOPD
Framingham
GALA
GALAII
GCPD-A
GENAF
GeneSTAR
GENOA
GenSalt
GGAF
GOLDN
HCHS-SOL
HVH
HyperGEN
INSPIRE_AF
IPF
JHS
JHU_AF
LTRC
Mayo_VTE
MESA
MGH_AF
miRhythm
MLOF
MPP
MSH
NSRR-CFS
OMG_SCD
ORCHID
PARTNERS
PCGC
PCGC_CHD
pharmHU
PIMA
PUSH_SCD
RED_CORAL
REDS-III_Brazil_SCD
SAFHS
SAGE
SAPPHIRE_asthma
Sarcoidosis
SARP
SAS
SHARP
SPIROMICS
STOPII
THRV
VAFAR
VU_AF
Walk_PHaSST_GRU
WGHS
WHI

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

Studies Available on BDC (COVID)

AACAC
ACTIV-4A
ACTIV-4B
AFLMU
AMISH
ARIC
Asthma
AustralianFamilialAF
BAGS
BioMe
BioVU_AF
BostonBrazil_SCD
C3PO
CAMP
CARDIA
CARE_BADGER
CARE_CLIC
CARE_PACT
CARE_TREXA
CATHGEN
CCAF
CCAF_AF
CCDG_PMBB_AF
CFS
ChildrensHS_GAP
ChildrensHS_IGERA
ChildrensHS_MetaAir
CHIRAH
CHS
CIBMTR
CMG_WGS_HMB
COPDGene
CRA
CSSCD
DECAF
DHS
ECLIPSE
EGCUT
EOCOPD
FHS
GALA
GALAII
GCPD-A
GENAF
GeneSTAR
GENOA
GenSalt
GGAF
GOLDN
HCHS-SOL
HVH
HyperGEN
INSPIRE_AF
IPF
JHS
JHU_AF
LTRC
Mayo_VTE
MESA
MGH_AF
miRhythm
MLOF
MPP
MSH
NSRR-CFS
OMG_SCD
ORCHID
PARTNERS
PCGC
PCGC_CHD
pharmHU
PIMA
PUSH_SCD
RED_CORAL
REDS-III_Brazil_SCD
SAFHS
SAGE
SAPPHIRE_asthma
Sarcoidosis
SARP
SAS
SHARP
SPIROMICS
STOPII
THRV
VAFAR
VU_AF
Walk_PHaSST_GRU
WGHS
WHI

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

Studies Available on BDC (SCD)

AACAC
ACTIV-4A
ACTIV-4B
AFLMU
AMISH
ARIC
Asthma
AustralianFamilialAF
BAGS
BioMe
BioVU_AF
BostonBrazil_SCD
C3PO
CAMP
CARDIA
CARE_BADGER
CARE_CLIC
CARE_PACT
CARE_TREXA
CATHGEN
CCAF
CCAF_AF
CCDG_PMBB_AF
CFS
ChildrensHS_GAP
ChildrensHS_IGERA
ChildrensHS_MetaAir
CHIRAH
CHS
CIBMTR
CMG_WGS_HMB
COPDGene
CRA
CSSCD
DECAF
DHS
ECLIPSE
EGCUT
EOCOPD
FHS
GALA
GALAII
GCPD-A
GENAF
GeneSTAR
GENOA
GenSalt
GGAF
GOLDN
HCHS-SOL
HVH
HyperGEN
INSPIRE_AF
IPF
JHS
JHU_AF
LTRC
Mayo_VTE
MESA
MGH_AF
miRhythm
MLOF
MPP
MSH
NSRR-CFS
OMG_SCD
ORCHID
PARTNERS
PCGC
PCGC_CHD
pharmHU
PIMA
PUSH_SCD
RED_CORAL
REDS-III_Brazil_SCD
SAFHS
SAGE
SAPPHIRE_asthma
Sarcoidosis
SARP
SAS
SHARP
SPIROMICS
STOPII
THRV
VAFAR
VU_AF
Walk_PHaSST_GRU
WGHS
WHI

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

Studies Available on BDC

Publicly available datasets:

BioLINCC Framingham

BioLINCC CAMP

BioLINCC Digitalis

1000 Genomes

Synthetic tutorial dataset

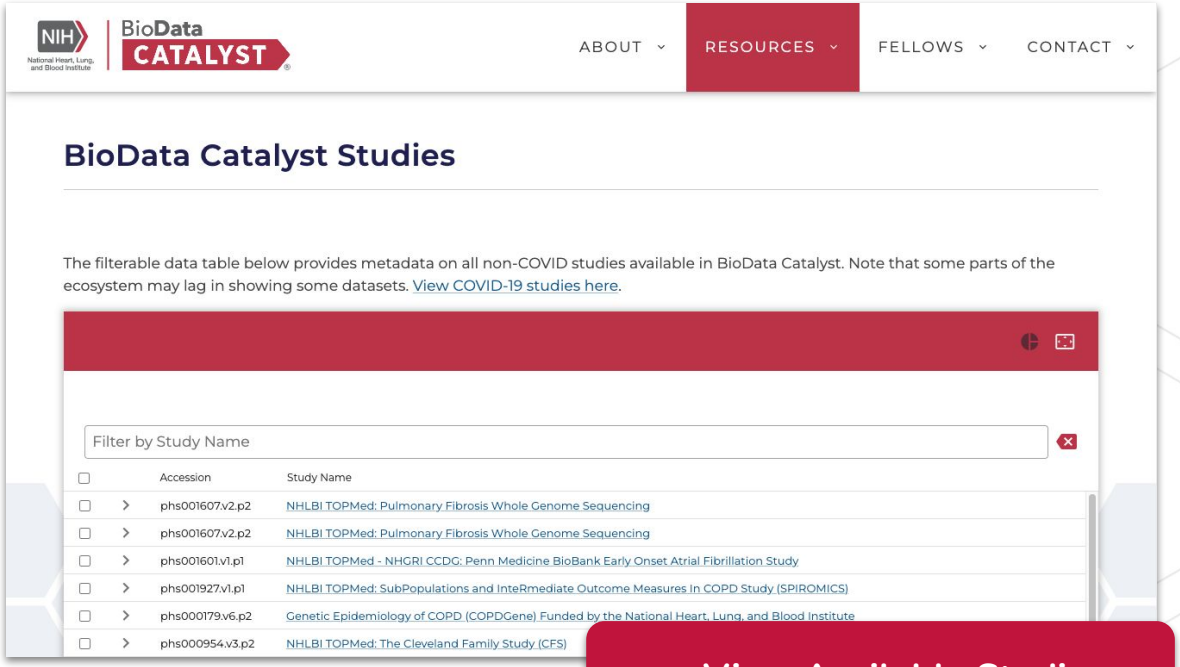
Data Available in BDC

BDC is always ingesting new data

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

Resources → Data

Click "Explore Studies"



→ View Available Studies

Cardiovascular Health Datasets

Dataset Name	Focus
Multi-Ethnic Study of Atherosclerosis (MESA) SHARe - longitudinal	Atherosclerosis, broadly phenotyped
Coronary Artery Risk Development in Young Adults (CARDIA) - longitudinal	
Atherosclerosis Risk in Communities (ARIC) Cohort - longitudinal	
Framingham Heart Study - longitudinal	
Cardiovascular Health Study (CHS) Cohort: an NHLBI-funded observational study of risk factors for cardiovascular disease in adults 65 years or older	Cardiovascular disease
The *Hispanic* Community Health Study / Study of *Latinos* (HCHS/*SOL*)	Diverse pops
Jackson Heart Study (JHS) Cohort - African Americans	Diverse pops

... and more!

https://topmed.nhlbi.nih.gov/group/project-studies?field_is_this_a_value=sub

Open PIC-SURE Demo

Emily Hughes, PIC-SURE

What is Open PIC-SURE?

A component of the BDC ecosystem that allows you to:

- **Search** any clinical variable in the ecosystem
- Build queries by **filtering** on variables
- Retrieve **aggregate counts** based on selected cohort

... all with just an eRA Commons account.

No dbGaP authorization to access data is required!

Demo

Going further with data exploration

Once you are authorized to access datasets, you can use Authorized PIC-SURE to:

- Build queries with both **clinical and genomic variables**
- Explore **participant-level** data
- Easily **export data** to other analysis platforms in BDC
- Learn to build complex queries through **programming languages**, such as R and python
  - Curated coding examples that show how to use PIC-SURE to build queries and conduct simple analyses in Jupyter Notebook and RStudio

Genome Wide Association on BDC Powered by Seven Bridges

David Roberson, Seven Bridges/Velsera

Discussion Time

Prompts

- Are you working locally or in the cloud right now?
- How do you currently learn about new computational resources?

Demo objectives

Key Concepts

- BDC is an integrated environment where data, tools and compute resources are co-located.
- Analysis can be done using Data Studio or by running Apps as scalable cloud jobs (tasks)

Demo

- ❑ Create a project and invite a team member
- ❑ Add GWAS “Apps” from the Public Apps Gallery
- ❑ Launch and use a Data Studio

Seven Bridges workspace environment

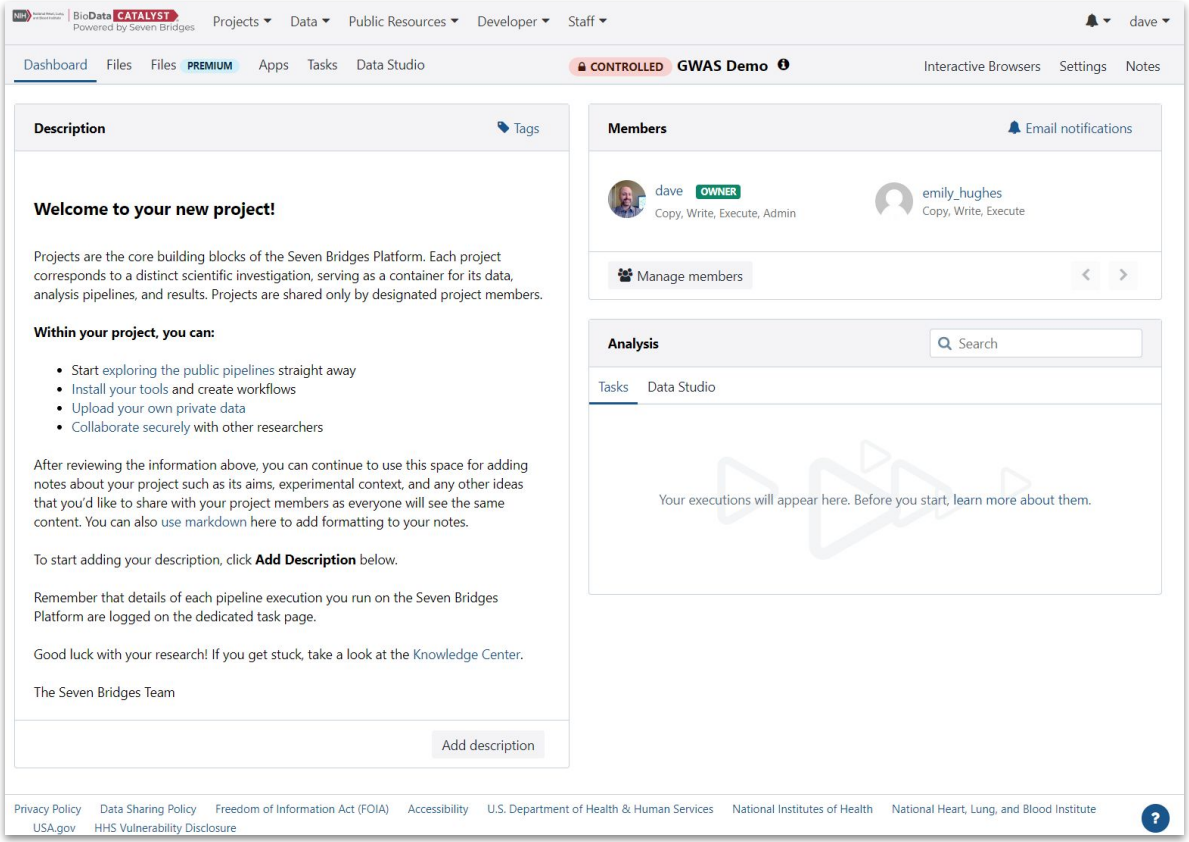
Private, secure workspaces with the option to collaborate

Set up analyses with visual user interface or API

Jupyterlab Notebooks, RStudio and SAS Studio

Compute on AWS or Google

Hundreds of hosted pipelines



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

Seven Bridges Demo

Demo objectives

Key Concepts

- BDC is an integrated environment where data, tools and compute resources are co-located.
- Analysis can be done using Data Studio or by running Apps as scalable cloud jobs (tasks)

Demo

✓ Create a project and invite a team member

✓ Add GWAS “Apps” from the Public Apps Gallery

✓ Launch and use a Data Studio

Advice

Kaleena Narwani, BDC Coordinating Center

What helped you accomplish your research?

Computation and Storage

- High parallelization of each of the steps of my research- **what would take nearly 5 days on local takes less than six hours on platform**
- Computation power and storage space
- Data resources and storage

Help and Support

- Utilize the help and resources provided rather than staying stuck
- Evaluate the apps available in the Public Repository before assuming you need to build one
- Reach out to the help team and don't feel frustrated
- User-friendly App with documentation and access to help desk

What helped you accomplish your research?

Start small and stay organized

- Create full workflows **incrementally** - it is easier to figure out errors with smaller steps
- **Start small** when implementing/testing your workflow: small dummy dataset, testing locally, start with one sample, ...
- Use data tables, tag your workflows, document early, write verbose notebooks.

Community and Collaboration

- Close connected research community
- Collaborators from multiple institutions
- Sharing results and workflows easily to collaborators (& eventually the general field)
- Engage with the research community.

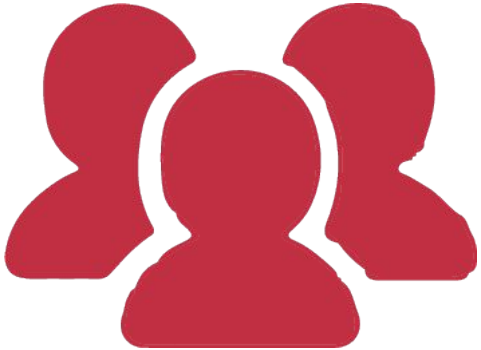
Next Steps:

Join the Community | Request Pilot Funding | Learn More

Kaleena Narwani, BDC Coordinating Center

Community engagement and support

*BDC is a people-centric endeavor. We are building a **community of practice** working to collaboratively solve technical and scientific challenges.*



- User-driven, vibrant community
- Peer-to-peer mentoring
- Expert support available
- Community Forum
- Community Hours and Showcases

→ Join the BDC Community

Existing Content & Future Events

Monthly Community Hours

Researcher Showcase with B2B and PCGC/CDDRC Fellow, Andrew Blair

Wednesday, June 21 at 1 pm ET

https://bit.ly/BDC-June

Sign up now!

View past materials on our forum

- Curated notes, slides, and recordings on a **variety** of topics, including:
  - Exploring and Accessing Data
  - Tour of Analysis Workspaces
  - Interactive Analysis
  - Cloud Costs
  - Reproducible Research Methods
  - Researcher Showcases
  - And more !

You can also find **recordings** on our YouTube channel

What are Cloud Credits?

Users are not charged for the storage of hosted datasets; however, if hosted data is used in analyses, users incur costs for computation and storage of derived results.

BDC users who upload/import their own data to the system incur storage costs for these uploaded files as well.

Web resource: Cloud Costs and Credits

Cloud Credits Workflow

1

Sign up for the community

Sign up at biodatacatalyst.nhlbi.nih.gov/contact/ecosystem

2

Sign up for a workspace

Seven Bridges and/or Terra

3

Apply for Pilot Credits

Fill out the Cloud Credits Request form.

Use all credits on a single platform, or split.

4

Apply for additional credits or pay yourself

Cover costs after pilot funding has been exceeded.

Potential Exception: Research in the heart, lung, blood, and sleep fields

Web Form

After signing up for a workspace, fill out the **Cloud Credits request form** for free cloud credits

https://biodatacatalyst.nhlbi.nih.gov/resources/cloud-credits/

Learning Resources

BDC website: https://biodatacatalyst.nhlbi.nih.gov/

Learn: https://biodatacatalyst.nhlbi.nih.gov/resources/learn

Documentation: https://bdcatalyst.gitbook.io/biodata-catalyst-documentation/

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

Our Gitbook documentation includes:

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

Getting Started

1. Join the Community: https://biodatacatalyst.nhlbi.nih.gov/contact/ecosystem
   - You'll need an eRA Commons ID for login credential
2. Asset exploration (pre-dbGaP)
3. Request Cloud Credits!
   - Initial credits are granted to conduct data discovery and preliminary analysis
   - Further computational costs should be budgeted in grant
4. Start your project: dbGaP authorizations and formal onboarding
5. Documentation

Questions?

Thank you!

Contact us with questions

- Ingrid Borecki, BDC Steering Committee Chair and Fellows Program Lead: iborecki28@gmail.com
- Emily Hughes, PIC-SURE: emily_hughes@hms.harvard.edu
- Dave Roberson, Velsera: david.roberson@velsera.com
- Amber Voght, BDC Coordinating Center: alvoght@renci.org
- **Kaleena Narwani, BDC Coordinating Center: knarwani@renci.org**
- BDC help desk – fast response times on any question: https://biodatacatalyst.nhlbi.nih.gov/contact/

Open Discussion

Open Discussion Topics

Are you interested in future sessions?

What would you like to hear about next?

- Finding Data (Study, Variable, and Variant Search)
- Bring Your Own Data
- Tools and Workflows on the Ecosystem
- Cloud Credits and Cost
- Researchers and their experience
- Other topics?

What part of your research are you currently working on?

EXTRA SLIDES

COPDGene Image data

The COPDGene® Study is one of the largest studies ever to investigate the underlying genetic factors of Chronic Obstructive Pulmonary Disease or COPD.

Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE): ECLIPSE was a longitudinal observational study of 2164 COPD subjects and a smaller number of smoking controls (337) and nonsmoking controls (245) followed regularly for three years, with three chest CT scans (at baseline, one year, and three years). [dbGAP; disease specific research only; no IRB]

COVID data sets from the PETAL Network

The NHLBI is also leveraging BioData Catalyst ecosystem to help coordinate various data management needs of many of the COVID-19 efforts.

Outcomes Related to COVID-19 Treated With Hydroxychloroquine Among In-patients With Symptomatic Disease (ORCHID) Study: A randomized placebo controlled study of hydroxychloroquine called ORCHID stopped enrolling new patients on June 19th based on the fourth scheduled interim analysis showing no evidence of benefit or harm. [ORCHID Study Page; dbGAP; HMB consent; no IRB]

Additional PETAL Network data sets will be available soon.

https://petalnet.org/; Accessed 5/24/2021

NHLBI Sickle Cell Disease Studies

Study data sets will be available soon!

- Multicenter Study of Hydroxyurea (MSH)
- Optimizing Primary Stroke Prevention in Children with Sickle Cell Anemia (STOP II) [dbGAP; GRU consent; no IRB]
- Cooperative Study of Sickle Cell Disease (CSSCD) [dbGAP; GRU consent; no IRB]
- Hematopoietic Cell Transplant for Sickle Cell Disease (HCT for SCD) [dbGAP; GRU consent; no IRB]

TOPMed WGS Data

Phenotypic

Harmonized data

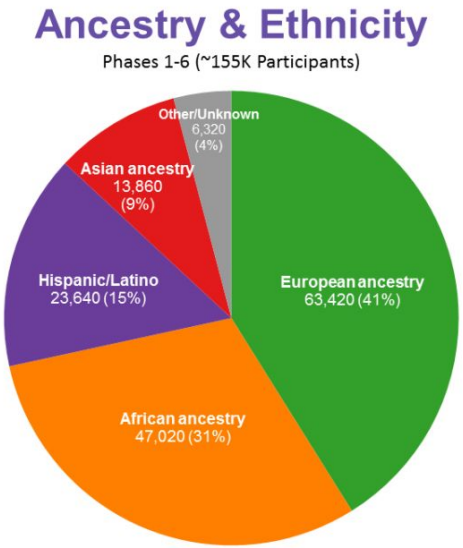
44 high-priority clinical and demographic variables have been harmonized by the TOPMed Data Coordinating Center (DCC) in order to facilitate cross-study analysis.

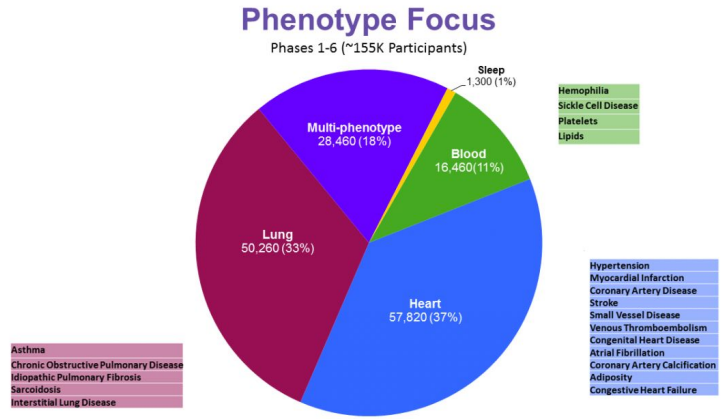
Non-harmonized data

The full set of raw clinical and phenotypic variables for the hosted studies are also available on the Gen3 platform. Exploration and search is available via the Gen3 search engine (under the "Files" tab) and in the PIC-SURE API.

Genomic

Genomic data provided by the Trans-Omics for Precision Medicine (TOPMed) program, including CRAM and VCF files. These files are available in the Gen3 Exploration page.





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

Bring-Your-Own Data

- To support **flexibility and analysis**, we allow researchers to bring their own data and workflows into the ecosystem.
- Users can upload data for which they have the appropriate approval, provided that they do not violate the terms of their Data Use Agreements, Limitations, or IRB policies and guidelines.

Web resource: Bring Your Own Data

Requesting Access to Data

Components of a Data Access Request (DAR) in dbGAP

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

Resources and template language are available for submitting a dbGaP Data Access Request

- Contact our help desk
- View documentation