

# BioData Catalyst Workshop, Day One

Thursday, November 17th at 11 am ET

**We will get started shortly.**



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

Interact with us on our forum during today's workshop:

<https://bit.ly/BDC-Howard-Workshop>

# BioData Catalyst Workshop, Day One

Thursday, November 17th at 11 am ET

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



National Heart, Lung,  
and Blood Institute

**BioData**

**CATALYST**

Interact with us on our forum during today's workshop:

<https://bit.ly/BDC-Howard-Workshop>

# Statement of Conduct

The BioData Catalyst Consortium is dedicated to **providing a harassment-free experience for everyone**, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, or religion (or lack thereof). We do not tolerate harassment of community members in any form. Sexual language and imagery is generally not appropriate for any venue, including meetings, presentations, or discussions.

Web Resource: [Statement of Conduct](#)

# Agenda

## Day One: Thursday, November 17th

Topic	Time
<a href="#">Introductions and Housekeeping</a>	5 min
<a href="#">What is BioData Catalyst?</a>	15 min
<a href="#">Researcher Presentation and Q&amp;A: Dr. Fayuan Wen</a>	30 min
<b>Break - 20 min</b>	
<a href="#">Interactive Demo: Finding and Using NHLBI Hosted Data</a>	1 hr
<a href="#">Bring Your Own Data</a>	20 min
<a href="#">Overview of the BioData Catalyst Ecosystem</a>	10 min
Q&A	20 min

## Day Two: Friday, November 18th

Topic	Time
<a href="#">Tools, Workflows, and Interactive Analysis</a>	10 min
<a href="#">Understanding, Estimating, and Managing Cloud Costs</a>	15 min
<a href="#">Running a GWAS on BioData Catalyst Powered by Seven Bridges</a> , Part 1	1 hour
<b>Break - 20 min</b>	
<a href="#">Running a GWAS on BioData Catalyst Powered by Seven Bridges</a> , Part 2	1 hour
Q&A	30 min



# Introductions and Housekeeping



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Meet Your Hosts



**Kat Thayer**

*BioData Catalyst Powered by Terra  
Broad Institute*



**Emily Hughes**

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



**Ingrid Borecki**

*Chair Steering Committee,  
Fellows Program lead*



**Dave Roberson**

*BioData Catalyst Powered by Seven Bridges  
Seven Bridges*

**Thank you to our guest researcher**



**Fayuan Wen**

*Post-doc Associate  
Adjunct Lecturer  
Howard University*

# Have a question during the workshop?

Ask questions **at any time** for live support: <https://bit.ly/BDC-Howard-Workshop>

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



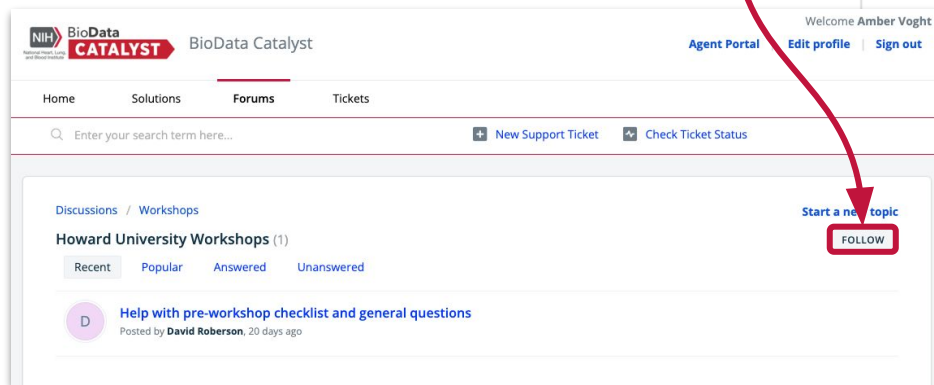
**Dave Roberson**

Community Engagement Specialist,  
Seven Bridges



**Amber Voght**

User Engagement Specialist,  
BioData Catalyst Coordinating Center



# Questions before we begin?

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

# What is BioData Catalyst?

Ingrid Borecki



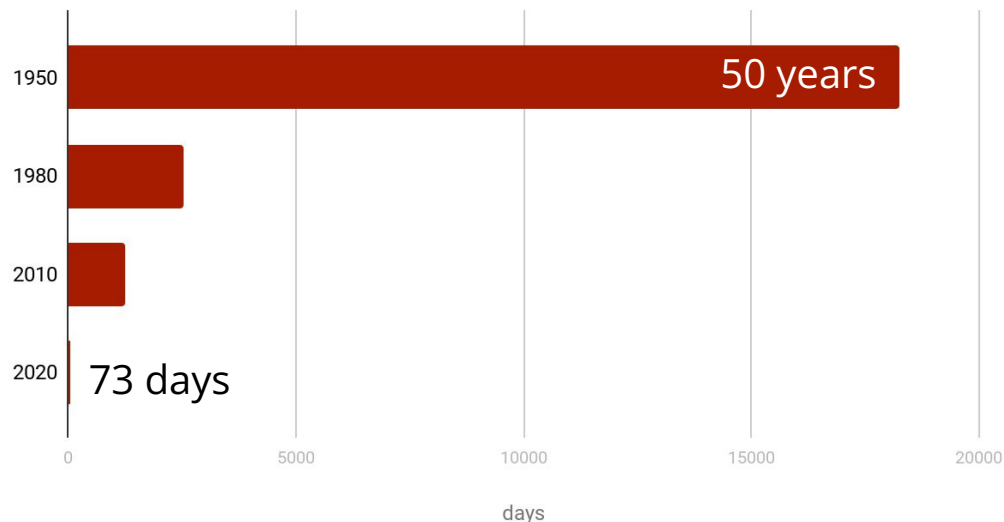
# Let's talk about:

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



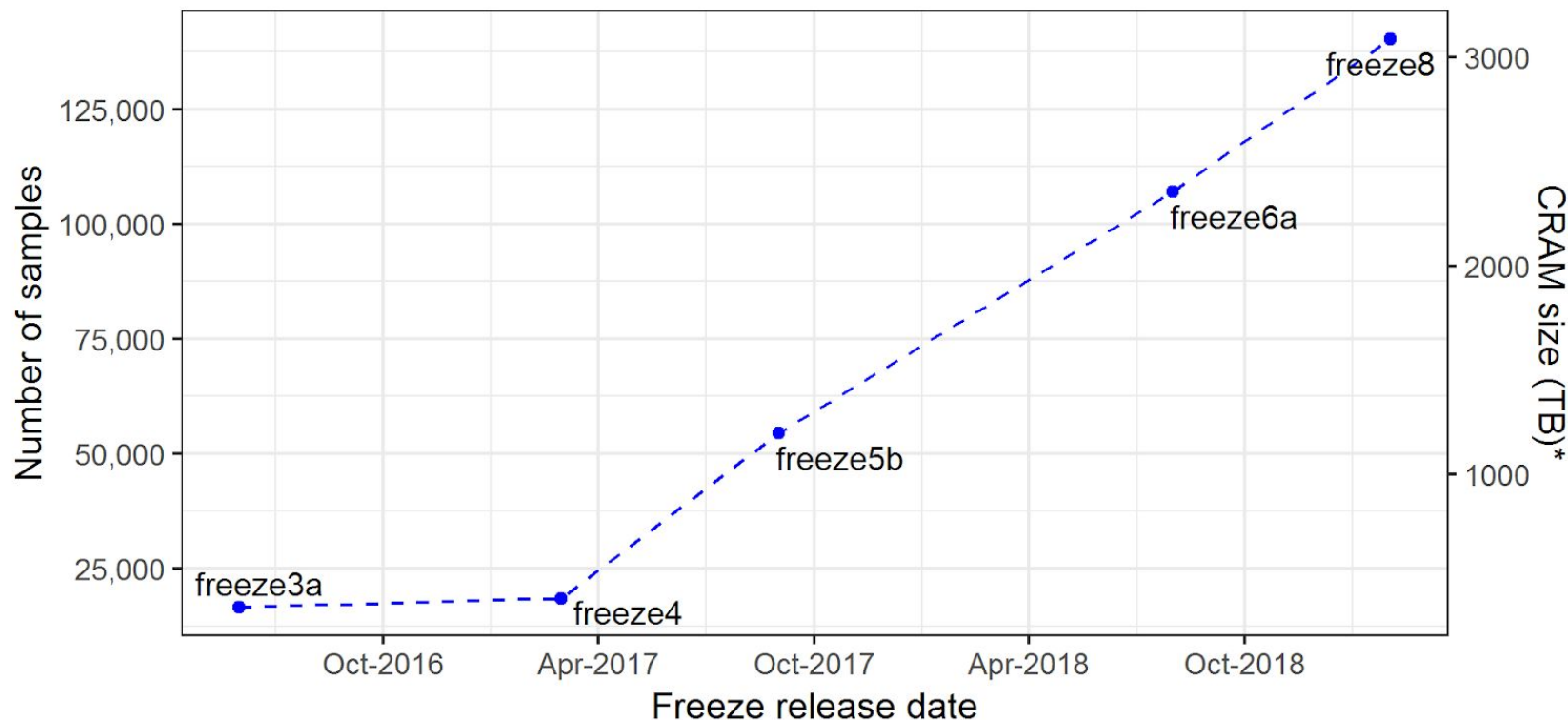
# The rate of data generation is accelerating rapidly

Doubling Time of Health Knowledge



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

## Growth in TOPMed Genome Sequence Data

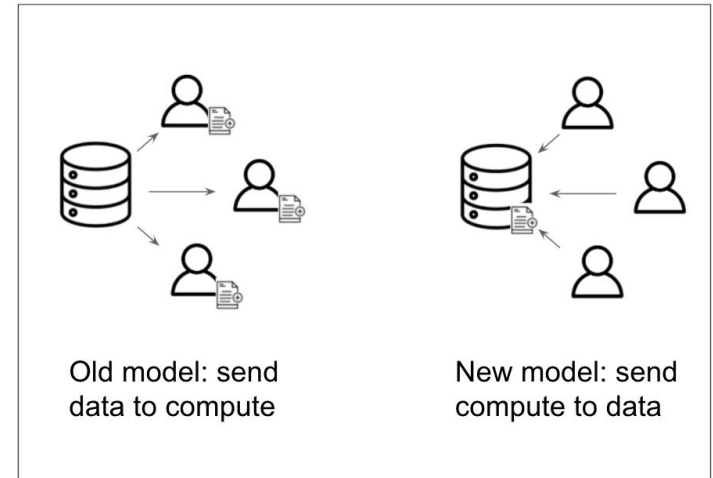


\*Based on average size of 22 GB for 1 DNA sample



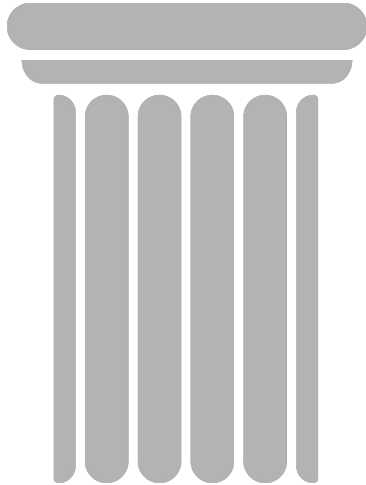
# Using the Cloud to store and analyze growing health data

- 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.

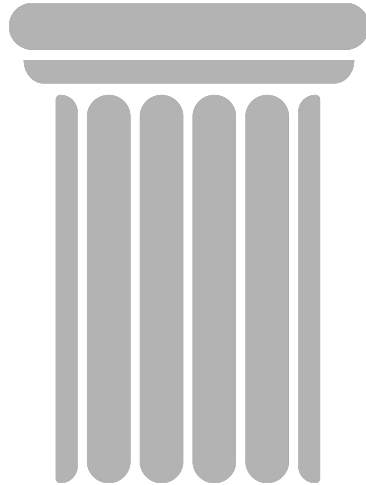


# NHLBI BioData Catalyst

## 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.

**WHO?**



**WHAT?**



Genomics



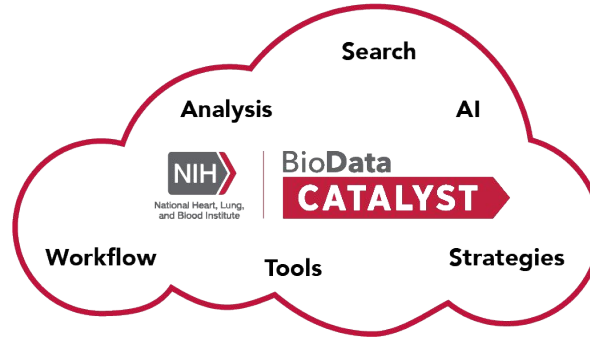
Clinical



Imagery

DATA  
HARMONIZATION

**WHERE?**

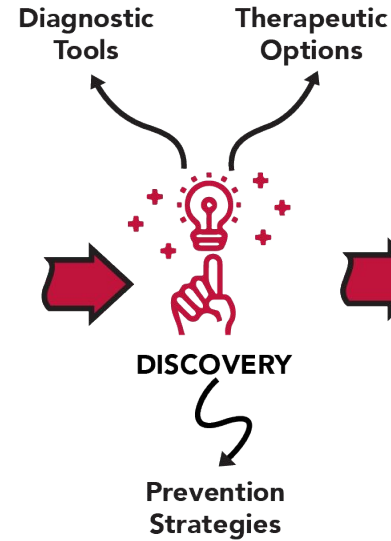


- UNDERSTAND
- OPEN SCIENCE
- CROSS-LINK

- COLLABORATE
- SCALE
- SHARE
- INTEROPERATE

**HOW?**

**SCIENCE!**



**WHY?**



**PATIENTS!**

# What BioData Catalyst offers



## Managing the Computing Environment

Elastic Computing



## Easier Access to many High Value Datasets



## Tooling

Data Discovery

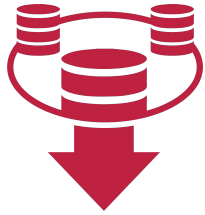
Statistical Analysis  
Tools (R, SAS)

Other Specialized  
Workflows



## Community and Peer Interactions

# The Computing Environment



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



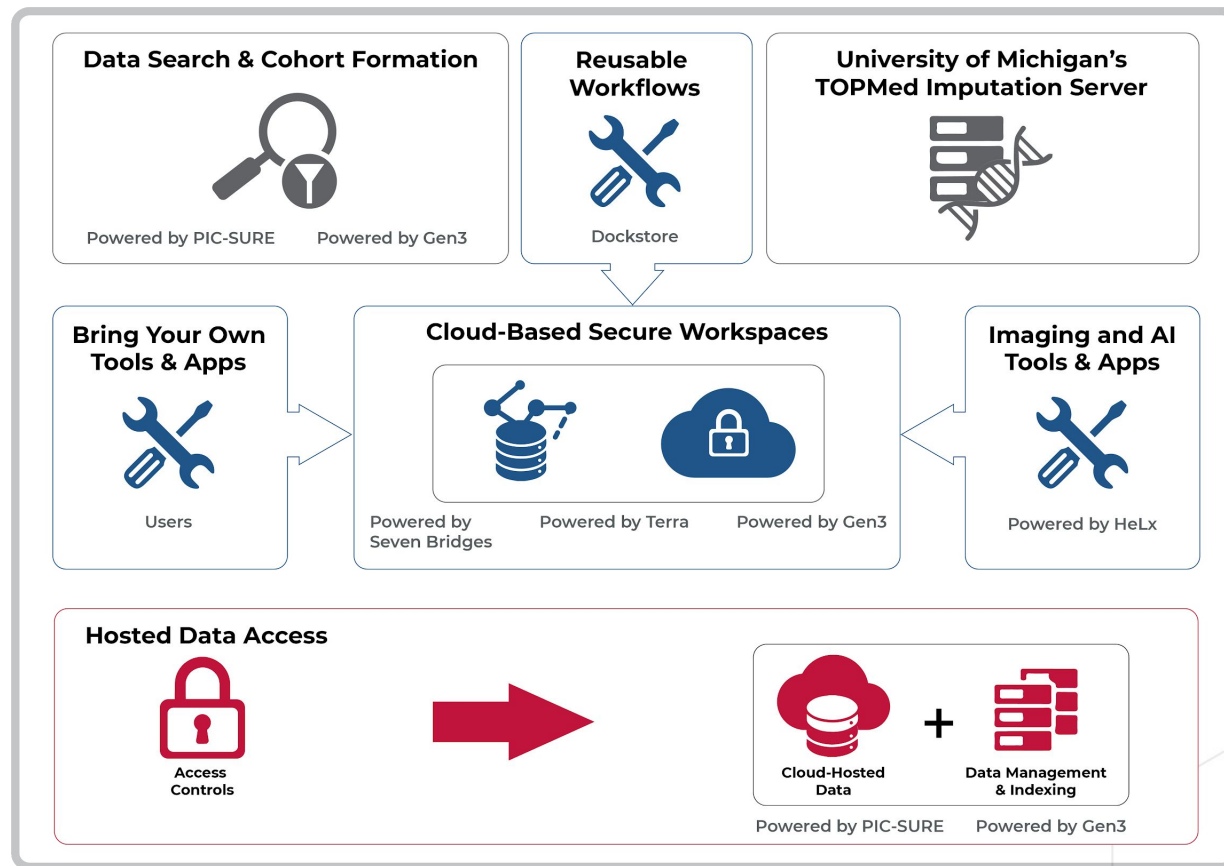
No **computer**  
**system** to  
**manage**



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



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



# Let's talk about:

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



# Platforms and Services

## Explore Data

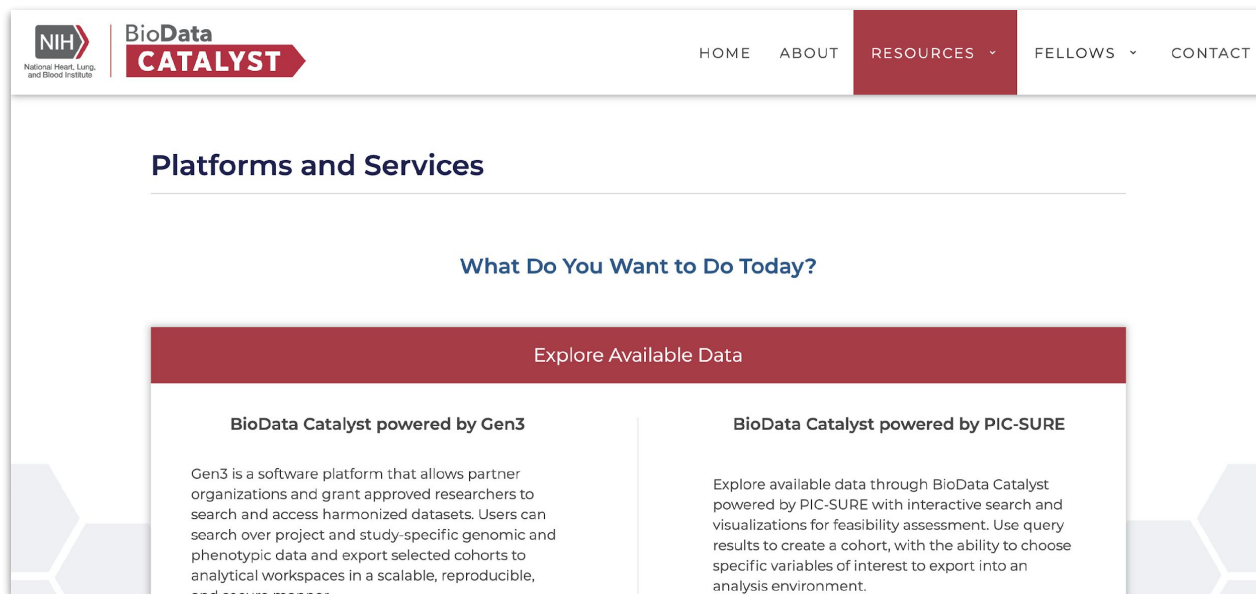
- PIC-SURE
- Gen3

## Analyze Data

- Seven Bridges
- Terra

## Community Tools

- Dockstore



Web resource: [Services](#)



# Learning Resources

BDCatalyst website <https://biodatacatalyst.nhlbi.nih.gov/>

**Web Resource:** [Learn](#)

**Documentation Resource:** [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



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

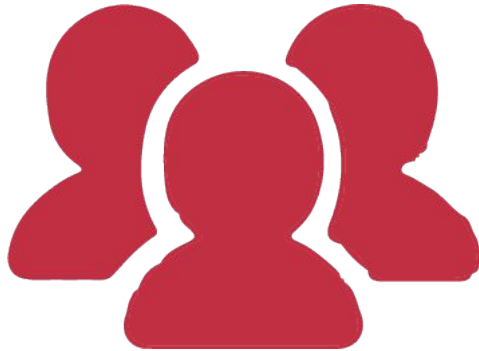
# Let's talk about:

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



# Community engagement and support

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



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

# Community Hours

Performing a GWAS on *BioData*  
*Catalyst Powered by Seven Bridges*

Wednesday, November 30 at 1 pm ET

[bit.ly/BDC-GWAS-Community-Hours](https://bit.ly/BDC-GWAS-Community-Hours)

**Sign up now!**

View [past materials](#) on our forum

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



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

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

**Join the NHLBI BioData  
Catalyst Community**

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

# Questions?

**Next up:** Researcher Presentation: Dr. Fayuan Wen

# Researcher Presentation:

Fayuan Wen, Howard University



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Association Study of Iron Overload in Sickle Cell Disease Population Using NHLBI WGS from TOPMed

Presenter: Fayuan Wen, PhD

Postdoctoral Associate, Center for Sickle Cell Disease

Adjunct Lecturer, Biology Department

Howard University



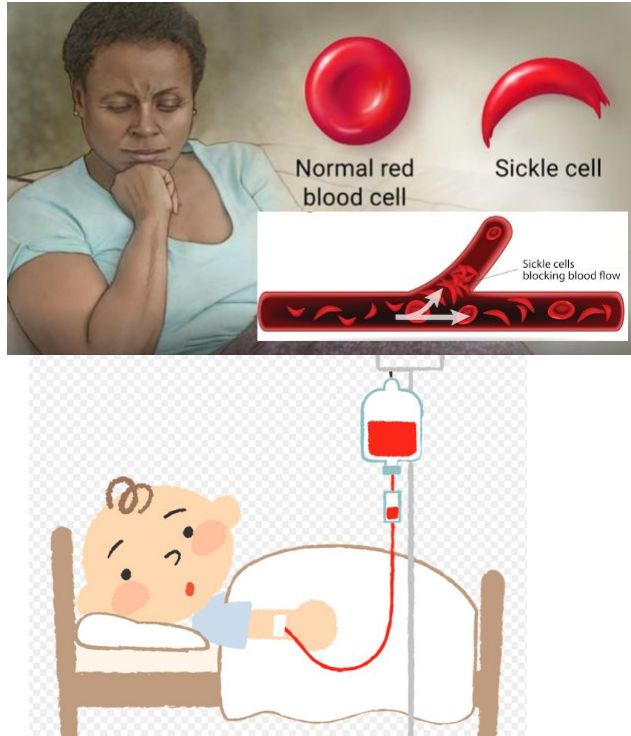
National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

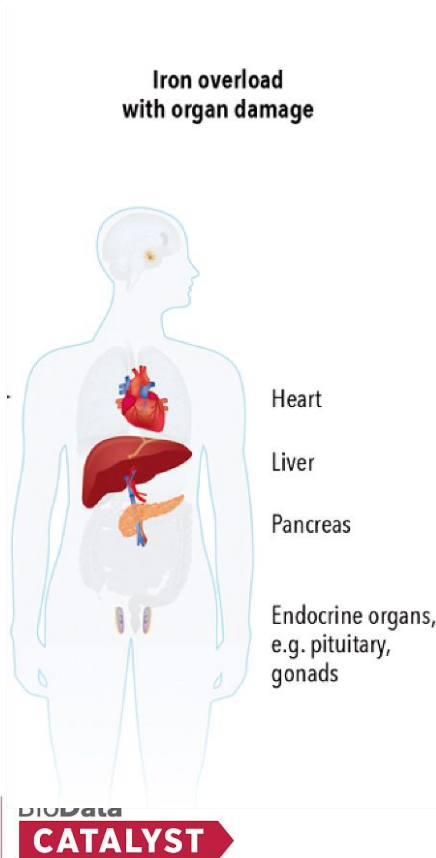


# Introduction



- Sickle Cell Disease (SCD) is a hereditary disorder caused by a mutation in HBB gene.
- Blood transfusion is an effective and proven treatment for some severe complications of sickle cell disease.
- It may help prevent primary and secondary stroke in children who have sickle cell disease.

# Introduction



- Iron overload is a well-known complication in SCD resulting from long-term blood transfusions and hemolysis.
- Underlying genetic factors may contribute to phenotypic diversity of iron overload in SCD.
- More sensitive biomarkers for iron overload outcome prediction that can be applied at early clinical stages would be of direct clinical benefit to the SCD population, especially African Americans.

# Method

- Genome Wide Association Analysis using GENESIS on NHLBI BioData Catalyst platform powered by Seven Bridges
- Phenotype and Genotype Data from TOPMed cohorts:
  - Howard PUSH-SCD(N=370),
  - OMG SCD(N=636),
  - Walk PHaSST SCD(N=381)
  - REDS-III Brazil SCD(N=2620) cohorts
- Variables including ferritin, transferrin, iron overload status, sex, age, race and life-time blood transfusion from four cohorts were harmonized and used for GWAS.

# Workflow and Tools used on Seven Bridges

## Data Access

Proposal to  
TOPMed, IRB, Data  
access request to  
dbGaP

Download phenotype  
and genotype data  
from Exchange Area

Decrypted, Decompressed and  
uploaded to Seven Bridges platform.  
sratool kit 2.9.6 vdb-decrypt  
biodatacatalyst-uploader.sh

## Analysis

Merge and filter the four cohorts' VCF files, then  
converted to GDS format

Bcftools merge  
Bcftools filter: 'AC>4 &&QUAL>10'  
GENESIS VCF to GDS

GENESIS Single Variant  
Association Testing on  
Seven Bridges

Phenotype data harmonization,  
formatted it as  
AnnotatedDataFrame using  
interactive analysis R Studio

Run Null model using  
downloaded TopMed freeze8  
Kinship and PCA files  
GENESIS Null Model

Result visualization  
GENESIS LocusZoom  
Annotation  
TOPMed Annotation  
Explorer

# GENESIS Single Variant Association Test on Seven Bridges

**COMPLETED** 20201214FourCohortsFerritinLOGGENESISSingleVariantAssociation

Executed on Dec. 14, 2020 09:58 by fayuan

Spot Instances: **On** Memoization (WorkReuse): **On** Price: **\$3.90** Duration: **38 minutes**

App: GENESIS Single Variant Association Testing - Revision: 1

**Inputs**

- GDS files**
  - chr10FourCohortsAc4Qual1.filtered.gds
  - chr11FourCohortsAc4Qual1.filtered.gds
  - chr12FourCohortsAc4Qual1.filtered.gds
  - chr13FourCohortsAc4Qual1.filtered.gds
  - chr14FourCohortsAc4Qual1.filtered.gds
  - ...and 18 more items
- Known hits file**
  - No files selected
- Null model file**
  - 20200904ferritinLOGNullmodel\_null\_model\_invnorm.RData
- Phenotype file**
  - 20200904ferritinLOGNullmodel\_phenotypes.RData
- Variant Include Files**
  - No files selected

**App Settings**

define\_segments.R (#define\_segments\_r)

Genome build (Genome build) hg38

assoc\_single.R (#assoc\_single\_r)

CPU 4

Genome build (Genome build) hg38

Output prefix (Output prefix) 20201214FourCohortsFerritinLOG

memory GB 32

assoc\_plots.R (#assoc\_plots\_r)

Output prefix (Plots prefix) 20201214FourCohortsFerritinLOG

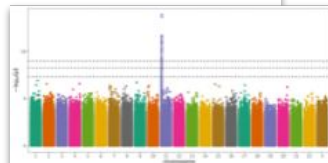
**Outputs**

**Association test plots**

- 20201214FourCohortsFerritinLOG\_manh.png
- 20201214FourCohortsFerritinLOG\_manh\_truncated.png
- 20201214FourCohortsFerritinLOG\_qq.png
- 20201214FourCohortsFerritinLOG\_qq\_bychr.png
- 20201214FourCohortsFerritinLOG\_qq\_truncated.png
- ...and 1 more item

**Association test results**

- 20201214FourCohortsFerritinLOG\_chr1.RData
- 20201214FourCohortsFerritinLOG\_chr2.RData
- 20201214FourCohortsFerritinLOG\_chr3.RData
- 20201214FourCohortsFerritinLOG\_chr4.RData
- 20201214FourCohortsFerritinLOG\_chr5.RData
- ...and 18 more items



# Result visualization

Interactive R to prepare locus file for GENESIS LocusZoom

```

CheckRdataforDec14GENESIS

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

.Rhistory Untitled1 chr11locus3 chr11locus2 assoc

Source on Save Run Source

1
2 setwd("/sbgenomics/project-files")
3 assoc=load("/sbgenomics/project-files/20201214FourCohortsFerritinLOG_chr11.RData")
4 load("/sbgenomics/project-files/20201214FourCohortsFerritinLOG_chr11.RData")
5 View(assoc)
6 library(tidyverse)
7 chr11locus2 = subset(assoc, Score.pval < 5e-8)
8 View(chr11locus2)
9 write.table(chr11locus2, file= "chr11locus2", sep= "\t", row.names = TRUE, col.names = NA)
10
11 #Subset the significant variants
12
13 chr11locus3 = filter(assoc, Score.pval < 5e-8) %>% select (variant.id, chr, pos, Score.pval)
14
15 #Filter the results
  
```

20201214FourCohortsFerritinLOG\_chr11.RData

#	variant.id	chr	pos	allelicIndex	nobs	freq	MAC	Score	Score.SE	Score.Stat	Score.pval	Ext	Ext.SE	PVE
1	24036289	11	102993	1	3410	0.008211437	56	-10.0863363	16.459693	-0.612802199	0.54000714	-0.0372104754	0.06075447	1.306442e-04
2	24036477	11	136798	1	3410	0.0014662757	10	-6.6424502	6.851970	-0.877130800	0.48831726	-0.0887604460	0.14185828	1.930867e-04
3	24036510	11	139612	1	3410	0.0008797654	6	7.4686349	5.307531	1.407176851	0.15937495	0.2851283282	0.18841152	5.834257e-04
4	24036558	11	121159	1	3410	0.0180351906	123	31.6194804	24.439475	1.293787214	0.19573883	0.0529384207	0.04091741	4.951895e-04
5	24036652	11	126183	1	3410	0.0045454545	31	-15.1874070	12.703170	-1.195560420	0.23186814	-0.0941151259	0.07872051	4.211446e-04
6	24037013	11	133738	1	3410	0.0029125513	20	-5.4852965	9.749215	-0.562639833	0.57388016	-0.0577112977	0.10272736	9.327153e-05
7	24037096	11	135572	1	3410	0.0008797654	6	-6.4108370	5.316980	-1.205728908	0.22792203	-0.2267694833	0.18807667	4.281389e-04
8	24037238	11	144932	1	3410	0.0008797654	6	-2.8786843	4.861025	-0.582198082	0.55371868	-0.1218255425	0.20571794	1.031286e-04
9	24037337	11	148591	1	3410	0.0193548387	132	15.1015935	24.607130	0.613708043	0.53940827	0.0249402529	0.04063863	1.100776e-04
10	24037606	11	161165	1	3410	0.0032218065	22	-12.7818340	10.036592	-1.276066131	0.20193214	-0.1273952370	0.09383435	4.797726e-04
11	24037723	11	168166	1	3410	0.0018651026	40	4.3296063	13.572442	0.318999941	0.74972656	0.0235035029	0.07367871	2.998261e-05
12	24037726	11	168189	1	3410	0.0121700880	83	18.9983867	19.501318	0.974210392	0.32995209	0.0489561308	0.05127859	2.790364e-04
13	24037825	11	172743	1	3410	0.0205278592	140	-0.4804732	25.389460	-0.018136193	0.88553005	-0.0007143276	0.03918842	8.891478e-08
14	24038147	11	176614	1	3410	0.0126099937	86	-32.5308575	19.555075	-1.863550645	0.09620226	-0.0816700215	0.05113762	8.153803e-04

Showing 1 to 16 of 1,519,326 entries, 14 total columns

	variant.id	chr	pos	Score.pval
column 5: metaflags				
2	24149360	11	4932190	4.449242e-08
3	24149363	11	4991804	7.480210e-10
3	24149363	11	4991953	1.663038e-08
4	24149370	11	4962597	2.772388e-08
5	24149513	11	5001177	4.627301e-09
6	24149522	11	5000634	1.546289e-08
7	24149524	11	5000806	1.546289e-08
8	24149553	11	5003197	1.463931e-08
9	24149602	11	5008473	1.009856e-08
10	24149624	11	5009356	6.164773e-09
11	24149647	11	5009867	1.029186e-08
12	24149684	11	5011706	6.157468e-10
13	24149738	11	5016258	1.052182e-08
14	24149821	11	5021245	1.407528e-08

Showing 1 to 16 of 81 entries, 4 total columns

variant.id	chr	pop
24148506	11	topmed
24149360	11	topmed
24149363	11	topmed
24149370	11	topmed
24149513	11	topmed
24149522	11	topmed
24149524	11	topmed
24149553	11	topmed



## Result visualization: GENESIS LocusZoom

COMPLETED

20210223-2nd\_no\_segmentChr11GENESIS LocusZoom run - 02-16-21 22:30:07

Get support

View stats & logs

Edit and rerun

Executed on Feb. 23, 2021 16:58 by **fayuan**

Spot Instances: **On** Memoization (WorkReuse): **Off** Price: **\$0.10** Duration: **22 minutes**

App: **GENESIS LocusZoom** - Revision: 0

Inputs

Association results files

chr11.RData

Database bundle

No files selected

Database directory

data

GDS files

chr11.gds

LD sample include

No files selected

Locus file

chr11variantlist.txt

Track files

No files selected

variant.id

chr

pop

24148506

11

topned

24149360

11

topned

24149363

11

topned

24149370

11

topned

24149513

11

topned

24149522

11

topned

24149524

11

topned

24149553

11

topned

App Settings

Show non-default

Genome build

hg38

Locus type

variant

Number of CPUs

4

Output prefix

20210217Chr11LocusZoom

Significance line

5e-8

Outputs

Locuszoom plots

ldl \_1\_20210217Chr11LocusZoom\_var24148506\_id\_TOPMED\_chr11\_49...

ldl \_1\_20210217Chr11LocusZoom\_var24149360\_id\_TOPMED\_chr11\_49...

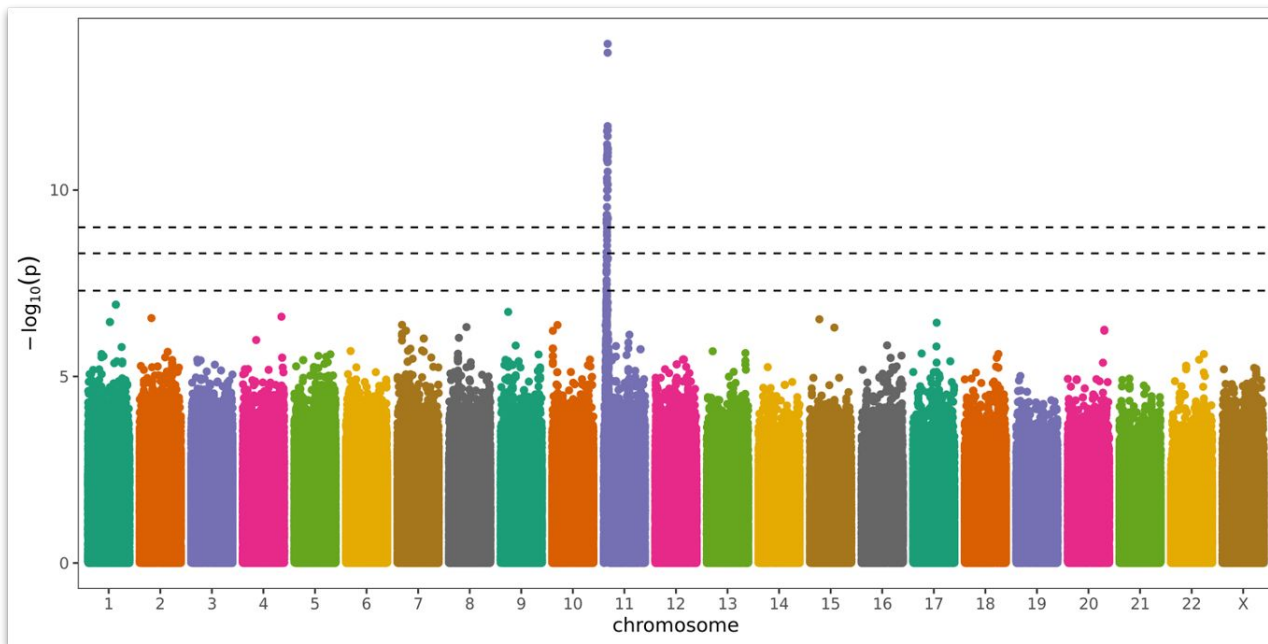
ldl 20210217Chr11LocusZoom\_var24149363\_id\_TOPMED\_chr11\_4991...

ldl 20210217Chr11LocusZoom\_var24149370\_id\_TOPMED\_chr11\_4992...

ldl 20210217Chr11LocusZoom\_var24149513\_id\_TOPMED\_chr11\_5000...

...and 75 more items

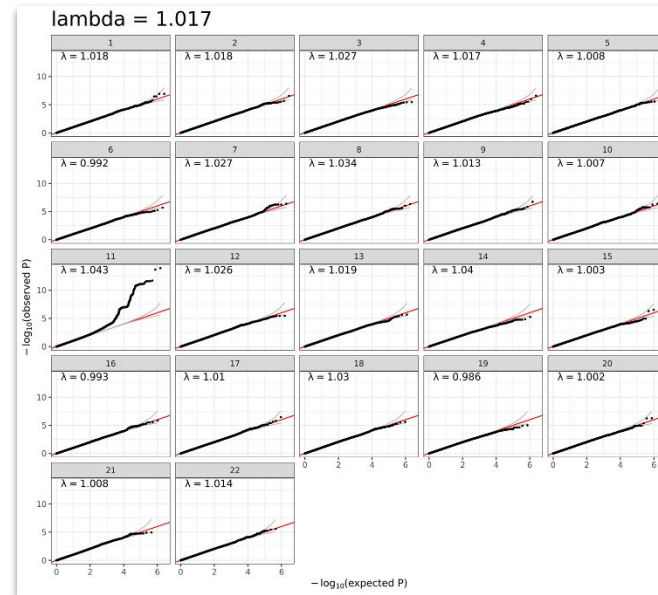
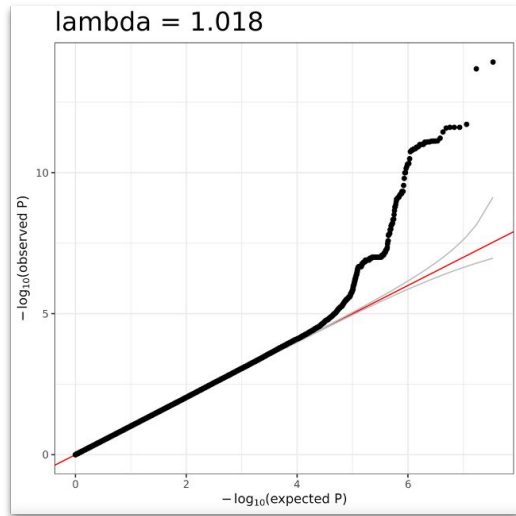
# GWAS: Manhattan plot of Ferritin



Manhattan plot showing association P values of the WGS association analyses with serum ferritin. Use log-transformed ferritin value as response variable, co-variants: age, gender, race, life-time blood transfusion. Generated by GENESIS pipeline.



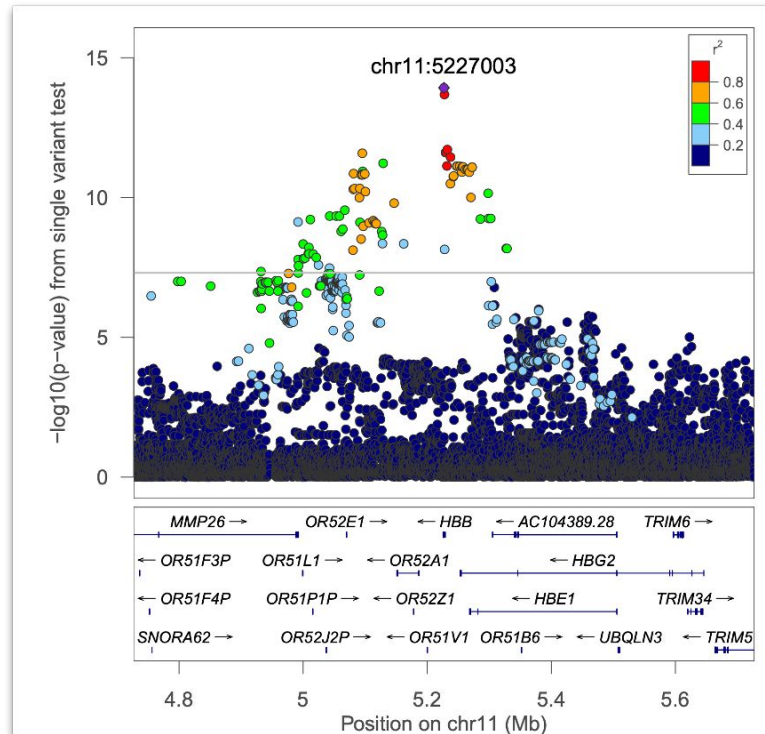
# Quantile-quantile plots



Generated by GENESIS pipeline.

The black dots represented the distribution of observed ordered  $-\log_{10}$  (P values) against the theoretical model distribution of expected ordered  $-\log_{10}$  P values. The grey line represents the theoretical model distribution of expected  $-\log_{10}$  P values under the null distribution.

# Regional Association Plot



Variants are plotted with their  $-\log_{10}(\text{p-value})$  on the left y-axis and the genomic position on the x-axis.

Estimated recombination rates on the right y-axis, are plotted to reflect the local linkage disequilibrium structure.

# The top 20 variants from WGS association results for Ferritin

Chr:POS	Variant	REF	ALT	MAF	MAC	Pvalue	VEP_ensembl_Gene_Name
11:5227003	rs34889882	CAG	C	0.10997067	750	1.19817E-14	HBB RF00621 AC104389.6
11:5227003	rs33930165	C	T	0.10997067	750	1.19817E-14	HBB AC104389.6 RF00621
11:5227851	rs74904621	T	A	0.10923754	745	2.07958E-14	HBB HBD AC104389.6 RF00621
11:5232478	rs112645511	G	A	0.0973607	664	1.93229E-12	HBB HBD
11:5229600	rs112829078	C	T	0.09824047	670	2.46689E-12	HBB HBD AC104389.6 RF00621
11:5229917	rs112987542	A	G	0.09824047	670	2.46689E-12	HBB HBD AC104389.6
11:5230700	rs111600160	C	G/A	0.09824047	670	2.46689E-12	HBB HBD
11:5095453	rs75601248	C	G	0.09824047	670	2.62378E-12	OR52E3P
11:5237598	rs75319671	T	C	0.09809384	669	3.61242E-12	HBD HBBP1
11:5128986	rs59398809	A	G	0.13709677	935	5.98533E-12	OR52A4P OR52A5
11:5231804	rs113427639	A	G	0.09956012	679	7.44867E-12	HBB HBD
11:5247539	rs112190925	C	T	0.09193548	627	7.61582E-12	HBG1 AC104389.5 HBBP1 BGLT3 HBD
11:5248550	rs554228323	C	CT	0.09193548	627	7.61582E-12	HBG2 HBD HBG1 AC104389.5 BGLT3
11:5252310	rs113854910	C	T	0.09120235	622	7.67233E-12	HBG2 HBG1
11:5256748	rs112346661	C	T	0.09222874	629	8.07223E-12	HBG2 AC104389.5 AC104389.4
11:5257903	rs112768836	C	T	0.09222874	629	8.07223E-12	HBG2 AC104389.5 AC104389.4
11:5272370	rs112638028	A	G	0.0898827	613	8.28011E-12	HBG2 HBE1 AC104389.4
11:5272416	rs112780476	T	C	0.0898827	613	8.28011E-12	HBG2 HBE1 AC104389.4
11:5261492	rs138048054	A	G	0.09208211	628	9.92788E-12	HBG2 AC104389.4
11:5264248	rs111781852	G	A	0.09208211	628	9.92788E-12	HBE1 HBG2 AC104389.4
11:5265961	rs111934403	C	T	0.09208211	628	9.92788E-12	HBE1 HBG2 AC104389.4

The significant 81 variants with  $p < 5E-8$  were annotated using TOPMed Annotation Explorer. The top 20 variants were shown in the table.

# Gene Functional Annotation

Among the 31 genes, 16 were functional annotated using Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8.

Molecular Function GO Term	Genes
GO:0004984~olfactory receptor activity	OR52A4P, OR52J3, OR51B2, OR52E2, OR52E1, OR51B5, OR51B4, OR52A1, OR51L1, OR52A5
GO:0005344~oxygen transporter activity	HBG2, HBG1, HBE1, HBB, HBD
GO:0004930~G-protein coupled receptor activity	OR52A4P, OR52J3, OR51B2, OR52E2, OR52E1, OR51B5, OR51B4, OR52A1, OR51L1, OR52A5
GO:0019825~oxygen binding	HBG2, HBG1, HBE1, HBB, HBD
GO:0020037~heme binding	HBG2, HBG1, HBE1, HBB, HBD
GO:0005506~iron ion binding	HBG2, HBG1, HBE1, HBB, HBD

Cellular Component GO Term	Genes
GO:0005833~hemoglobin complex	HBG2, HBG1, HBE1, HBB, HBD
GO:0072562~blood microparticle	HBG2, HBE1, HBB, HBD
GO:0005886~plasma membrane	OR52A4P, OR52J3, OR51B2, OR52E2, OR52E1, OR51B5, OR51B4, OR52A1, OR51L1, OR52A5
GO:0016021~integral component of membrane	OR52A4P, OR52J3, OR51B2, OR52E2, OR52E1, HBB, OR51B5, OR51B4, OR52A1, OR51L1, OR52A5

Biological Process GO Term	Genes
GO:0050911~detection of chemical stimulus involved in sensory perception of smell	OR52A4P, OR52J3, OR51B2, OR52E2, OR52E1, OR51B5, OR51B4, OR52A1, OR51L1, OR52A5
GO:0015671~oxygen transport	HBG2, HBG1, HBE1, HBB, HBD
GO:0007186~G-protein coupled receptor signaling pathway	OR52A4P, OR52J3, OR51B2, OR52E2, OR52E1, OR51B5, OR51B4, OR52A1, OR51L1, OR52A5
GO:0007165~signal transduction	OR52A4P, OR52J3, OR51B2, OR52E2, OR52E1, OR51B5, OR51B4, OR52A1, OR51L1, OR52A5
GO:0007596~blood coagulation	HBG2, HBG1, HBE1, HBB, HBD
GO:0007608~sensory perception of smell	OR51B5, OR51B4, OR52A1, OR52A5
GO:0051291~protein heterooligomerization	HBE1, HBB

# Future Directions

Further analysis will be expanded to other related phenotype such as transferrin, iron-overload status on the NHLBI BioData Catalyst platform.

Determine the pathways and functional relationships of iron overload related genes.

Structural variants?

The results from this study point to novel gene variants that might contribute to iron overload in SCD patients and serve as new biomarkers. Our findings will be useful for the future treatment of SCD patients and design of novel SCD therapeutics.

# Acknowledgments

This work was supported by **NIH Research Grants** (1P50HL118006, 1R01HL125005 and 1OT3HL147154)

**NHLBI BioData Catalyst**, NHLBI BioData Catalyst Powered by **Seven Bridges**, NHLBI Trans-Omics for Precision Medicine (**TOPMed**), **dbGaP**.

**Supervisors** at the Center for Sickle Cell Disease: Sergei Nekhai, James Taylor, Juan Salomon-Andonie

**Colleagues:** Angela Rock, Gulriz Kurban, Xiaomei Niu, Songping Wang

**Collaborators:** Victor R. Gordeuk, Mark Gladwin, Xu Zhang, Seyed Mehdi Nouraie, Yingze Zhang, Allison Ashley-Koch, Marilyn J. Telen, Brian Custer, Shannon Kelly, Carla Luana Dinardo, Ester Sabino, Quenna Wong

# Questions?

# Interactive Demo: Finding and Using NHLBI Hosted Data

Emily Hughes, PIC-SURE



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**



# Data available in BioData Catalyst

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

# Data available in BioData Catalyst

3.42  
Petabytes of  
data



490,000+  
Data files



280,000+  
Participants



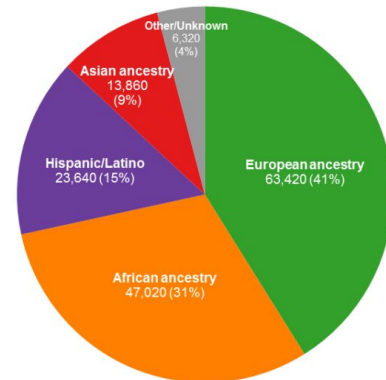
150,000+  
Whole genomes



## TOPMed Data

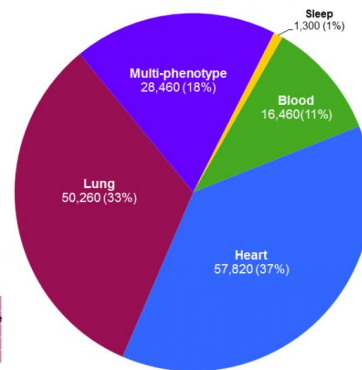
### Ancestry & Ethnicity

Phases 1-6 (~155K Participants)



### Phenotype Focus

Phases 1-6 (~155K Participants)



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

Hemophilia  
Sickle Cell Disease  
Platelets  
Lipids

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

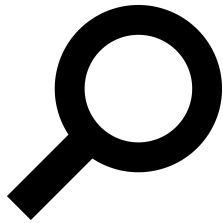
# Check Access to Data

## Three main ways to check your access to data:

1. BioData Catalyst website
  - **Demo:** About BioData Catalyst Dataset, <https://biodatacatalyst.nhlbi.nih.gov/resources/data>
2. *BioData Catalyst Powered by Gen3*
  - **Demo:** Exploring files on Gen3, <https://gen3.biodatacatalyst.nhlbi.nih.gov/explorer>
3. *BioData Catalyst Powered by PIC-SURE Data Access Dashboard*

# Empowering researchers to access data

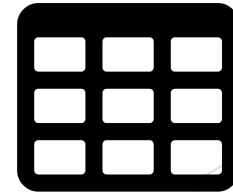
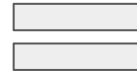
*BioData Catalyst Powered by PIC-SURE* facilitates approachable research for all skill levels.



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

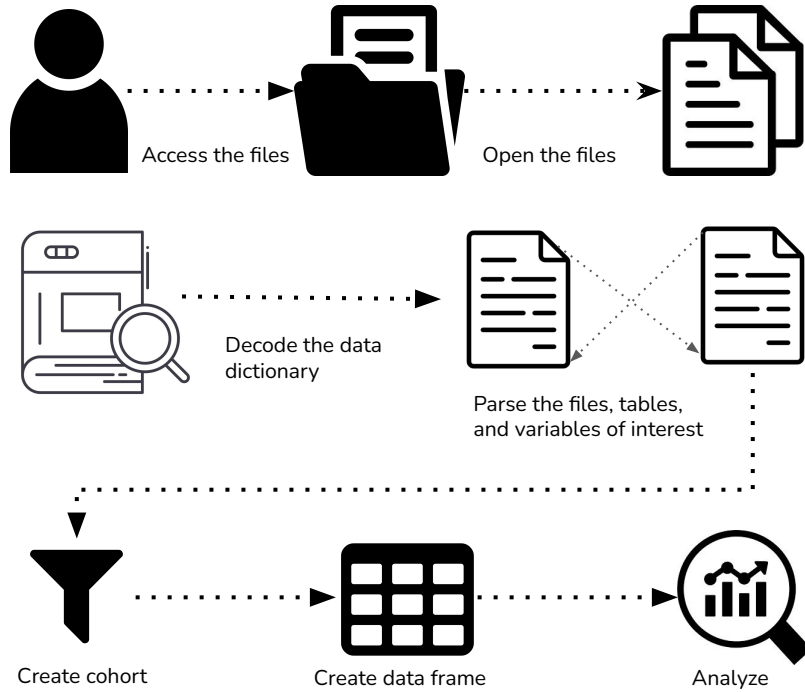
# *BioData Catalyst Powered by PIC-SURE*

Patient  
Information  
Commons  
-  
Standard  
Unification of  
Research  
Elements

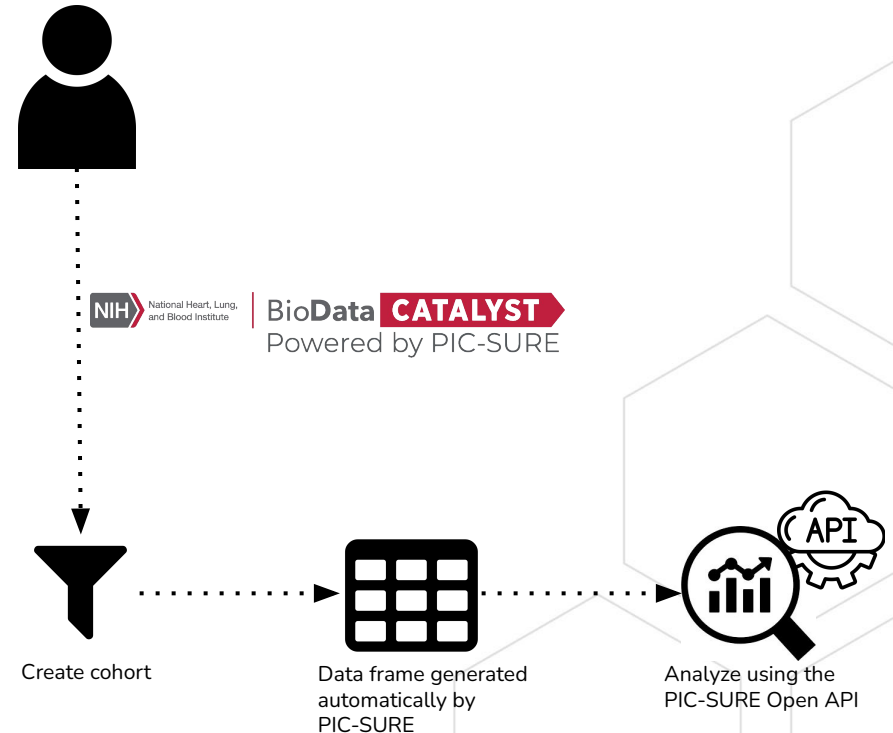
- Allows for searching and exporting data at the **variable** and **variant** level
- Integrates clinical and genomic datasets across BioData Catalyst
- UI allows users to search available data using queries to build cohorts
- Results can be exported via the API for analysis

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

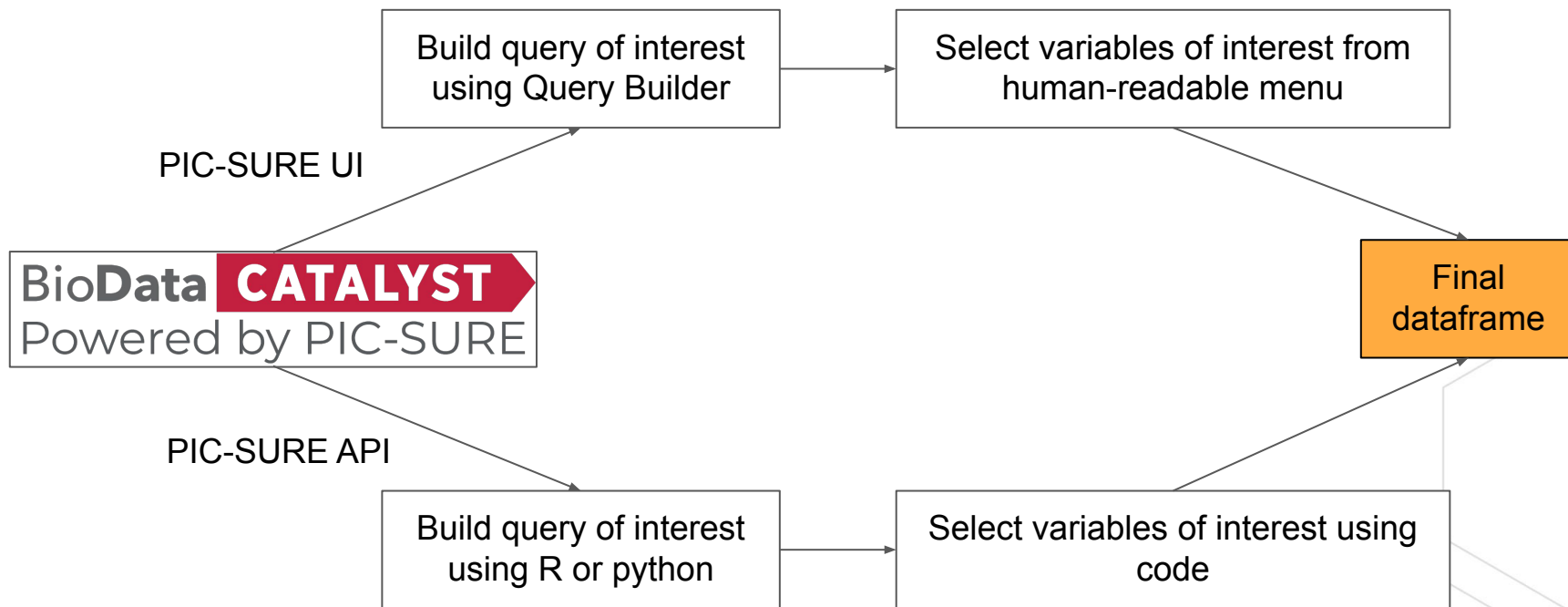
# Traditional Workflow



# PIC-SURE Workflow



# PIC-SURE workflow - 2 options



# Open vs Authorized Access

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



# PIC-SURE Data Access Dashboard

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

## Demo

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

**BioData CATALYST**  
Powered by PIC-SURE

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

**Authorized Access**  
 Explore Now  
 32 Studies  
 236,969 Participants

**Open Access**  
 Explore Now  
 64 Studies  
 286,180 Participants

Authorized Phenotypic and Genomic Datasets

- Aggregate Counts
- Participant-level Data
- Visualizations
- R and Python API Access

No Authorization Required

- All Phenotypic Datasets Available in PIC-SURE
- Aggregate Counts Only
- R and Python API Access

**Data Access Table**  
View all studies and associated information available in PIC-SURE separated by consent group. Check personalized authorization for available studies in the "Access" column.

(Current TOPMed data is Freeze5a)  
P = Phenotype, G = Genomic, P/G = Phenotype/Genomic, n/a = Not Applicable

Search:

Access	Abbreviation	Name	Study Focus	Study Design	Clinical Variables	Participants with Phenotypes	Samples Sequenced	Additional Information	Consents	Accession
<a href="#">Request</a>	ACTIV4a	A Multicenter, Adaptive, Randomized Controlled Platform Trial of the Safety and Efficacy of Antithrombotic Strategies in Hospitalized Adults with COVID-19 (ACTIV4a)	COVID-19	Interventional	N/A	1,083	N/A		General Research Use (GRU)	phs002694.v1.p1
<a href="#">Request</a>	ACTIV4b	COVID-19 Positive Outpatient Thrombosis Prevention in Adults Aged 40-80	COVID-19	Interventional	N/A	657	N/A		General Research Use (GRU)	phs002710.v1.p1
<a href="#">Granted</a>	AMISH	NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish	Cardiovascular Disease	Family/Twin/Trios	76	1,123	N/A		Health/Medical/Biomedical (IRB, MDS) (HMB-IRB-MDS)	phs000956.v4.p1

# PIC-SURE Open Access

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

## Demo

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

The screenshot displays the PIC-SURE Open Access search interface. The search bar at the top contains the term "epilepsy". Below the search bar, the interface is divided into several sections:

- Filter Search Results by Study Tags:** This section allows refining results using study filters. It shows buttons for CARDIA (22), CHS (4), CISCOD (2), FHS (1), CFS (1), and PCDC (1).
- Filter Search Results by Variable Tags:** This section allows refining results using tag-based filters. It shows buttons for ALLERGIC (28), ALLERGIES (28), ANEMIA (28), ATTACKS (28), BROKEN (28), CANCER (28), DAMAGE (28), FOOD (28), HEADACHE (28), HEADACHES (28), LEGS (28), and LUNG (28).
- Search Results:** This section displays a table of search results. The table has columns for Study, Variable Name, Variable Description, and Actions. It shows 31 variables matching the search. The first 10 entries are displayed, with a "Showing 1 to 10 of 29 entries" indicator at the bottom.
- Results Panel:** This panel provides a summary of the search results. It shows the total number of participants (180475 ±3) and lists the added variable filters. The filters include:
  - Filter on variable height\_baseline\_1: Study: HRMN (DCC Harmonized data set), Value: Include only participants with values greater than 18.
  - Filter on variable age\_at\_height\_baseline\_1: Study: HRMN (DCC Harmonized data set), Value: Include only participants with values greater than 18.
  - Filter on variable annotated\_sex\_1: Study: HRMN (DCC Harmonized data set), Value: Include only participants with values in [Female].
- Filtered Results by Study:** This section shows the number of participants for each study. It lists FHS (8028 total participants), HMB-IRB-MDS (6878 participants), HMB-IRB-NPU-MDS (1150 participants), COPDGENE (4847 total participants), HMB (4725 participants), and DS-CS (122 participants).

# PIC-SURE Authorized Access

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

## Demo

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

Search: epilepsy Search Genomic Filtering Reset

Filter Search Results by Study Tags

Refine results using study filters:

CARDIA (22) CHS (4)  
phs000285 phs000287

FHS (1) CFS (1)  
phs000007 phs000284

Filter Search Results by Variable Tags

Refine results using tag based filters. Showing 12 out of 7403 tags that have variables which match your search term:

ACID (26) AFTER (26)  
ALLERGY (26)  
BAD (26) BONE (26)  
CELL (26)  
CERVICAL (26)  
CHRONIC (26)  
COLD (26)  
CRAMPS (26)  
EYES (26) FEET (26)

View more tags

Search Results: 28 variables match your search. [Learn More](#)

Show 10 entries

Study	Variable Name	Variable Description	Actions
CHS	SYMPT109	OTHER SYMPTOM 1	▼ ↗
FHS	G3A660	1ST NP BIO PARENT - IF PARENT NOT LIVING, CAUSE OF DEATH	▼ ↗
CFS	OSMCSP	95 Specify other significant medical condition	▼ ↗
CARDIA	E08EPIAG	AGE DIAGNOSED-EPILEPSY. Q 15	▼ ↗
CARDIA	F08EPIAG	AGE DIAGNOSED-EPILEPSY. Q 15	▼ ↗
CARDIA	D08EPIAG	AGE FIRST DIAGNOSED WITH EPILEPSY. Q 15	▼ ↗
CARDIA	B09RESO	B09RESO	▼ ↗
CARDIA	F31ERRSN	CONDITION CAUSING VISIT TO ER. Q 3c	▼ ↗
CARDIA	A09MJ1DN	DESC OF 1ST MAJ HEALTH PROB. Q 1.13C	▼ ↗
CARDIA	A09NERDN	DESC OF NERVOUS DISORDER. Q 1.10B	▼ ↗

Showing 1 to 10 of 28 entries Previous 1 2 3 Next

**Results Panel**

Data Summary [What is this?](#)

**178401**  
Total Participants  
8 Variables

Tool Suite [What is this?](#)

Select and Package Data Variable Distributions

Added Variable Filters  
Active filters applied to your cohort.

Filter on variable **height\_baseline\_1**  
Study: HRMN (DCC Harmonized data set)  
Value: Include only participants with values

Filter on variable **age\_at\_height\_baseline\_1**  
Study: HRMN (DCC Harmonized data set)  
Value: Include only participants with values greater than 18

Filter on variable **annotated\_sex\_1**  
Study: HRMN (DCC Harmonized data set)  
Value: Include only participants with

# PIC-SURE Application Programming Interface (API)

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

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

## Introduction to the PIC-SURE API

This is a tutorial notebook aimed to get the user quickly up and running with the PIC-SURE API.

### PIC-SURE python API

#### What is PIC-SURE?

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

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

## Using PIC-SURE to build a query and retrieve data

You can also use the PIC-SURE API to build a query and retrieve data. With this functionality, you can filter based on specific variables, add others, and export the data as a dataframe into this notebook.

The first step to this is setting up the `authQuery`.

```
1: authPicSure = bdc.useAuthPicSure()
   authQuery_categorical_example = authPicSure.query()
```

### Build a query with a categorical variable

Let's practice building a query by filtering on variables. First, let's select a categorical variable to use. We can identify one using the `is_categorical` column of the variable dataframe.

```
1: i = 0
   categories = {}
   while len(categories) == 0 or len(categories) > 8:
       categorical_var_info = my_variables_df[my_variables_df.is_categorical == True]
```

# PIC-SURE API

## BioData Catalyst Powered by Terra

**BioData CATALYST** Powered by Terra WORKSPACES

Workspaces > biodata-catalyst/BioD...

**ABOUT THE WORKSPACE**

### BioData Catalyst Python PIC-SURE API examples

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

### PIC-SURE API Overview

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

### Workspace information

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

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

**OWNERS**

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

**TAGS**

Add a tag

No tags yet

**Google Bucket**

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

## BioData Catalyst Powered by Seven Bridges

**BioData CATALYST** Powered by Seven Bridges

Projects > Data > Public Gallery > Public projects > Developer > ernhughes1

**Dashboard** Files Apps Tasks

**PIC-SURE API** Interactive Analysis

### DESCRIPTION

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

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

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

### Important notes

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

### ANALYSES

Search

Tasks Data Cruncher

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

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

# Export into workspace

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

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



# Questions?

**Next up:** Bring Your Own Data

# Bring Your Own Data

Dave Roberson, Community Engagement Specialist  
at Seven Bridges



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**



# Bring-Your-Own Data

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

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

# Seven Bridges workspace environment

Private, secure workspaces with the option to collaborate

Set up analyses with visual user interface or API

Jupyterlab Notebooks and RStudio

Compute on AWS or Google

Hundreds of hosted CWL pipelines

The screenshot displays the Seven Bridges workspace environment interface. At the top, a navigation bar includes the NIH BioData CATALYST logo, a 'Projects' dropdown menu, and various other navigation options like 'Data', 'Public Gallery', 'Public projects', 'Automations', 'Developer', 'Staff', and a user profile 'alisonleaf'. Below this, a secondary bar shows 'Dashboard', 'Files', 'Apps', and 'Tasks' tabs, with 'Alison\_test\_GWAS' selected. The main content area is divided into two columns. The left column, titled 'DESCRIPTION', contains a 'Welcome to your new project!' message, explaining that projects are core building blocks of the platform. It lists actions users can take: exploring public datasets, installing tools, uploading private data, and collaborating. It also includes a note about recording project details and a link to the Knowledge Center. The right column, titled 'MEMBERS', shows a list of project members: alisonleaf (OWNER), dave, milan.domazet, and boris.majic, each with their roles. Below the members list is a 'Manage members' button. Further down, the 'ANALYSES' section shows a list of completed tasks, including 'GENESIS Null Model run' and 'GENESIS VCF to GDS run', both submitted by alisonleaf on Jan. 17, 2020.

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

Tasks Data Cruncher

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

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

# Work alone in a private project

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

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

The screenshot shows the BioData Catalyst web interface for a project named 'test project'. The top navigation bar includes the NIH BioData Catalyst logo, a 'Powered by Seven Bridges' tagline, and a series of dropdown menus: Projects, Data, Public Gallery, Public projects, Developer, and Staff. A user profile for 'alisonleaf' is visible in the top right corner. Below the navigation bar, the interface is divided into three main sections. The left section, titled 'DESCRIPTION', contains a 'Welcome to your new project!' message, a paragraph explaining that projects are the core building blocks of the platform, and a list of actions users can take within a project: exploring public datasets, installing tools, uploading private data, and collaborating securely. It also includes a note about recording project details and a tip about using markdown. The middle section, titled 'MEMBERS', shows the user 'alisonleaf' as the 'OWNER' with permissions to 'Write, Copy, Execute, Admin'. It includes a message about teamwork ('Don't work alone. The best research happens in teams.') and a button to 'Invite new members'. The right section, titled 'ANALYSES', has a search bar and a list of analyses, with 'Tasks' and 'Data Cruncher' visible. A footer note says 'Your executions will appear here.'

BioData CATALYST  
Powered by Seven Bridges

Projects Data Public Gallery Public projects Developer Staff

Dashboard Files Apps Tasks test project Interactive Analysis Settings Notes

DESCRIPTION Tags

**Welcome to your new project!**

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

**Within your project, you can:**

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

Please record the details of your project here, such as its aims, experimental context, and any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the platform are logged on the task page. This notepad is just for your own notes.

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

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

MEMBERS Email notifications

alisonleaf OWNER  
Write, Copy, Execute, Admin

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

Invite new members

Share your tools, data, and ideas with collaborators

ANALYSES Search

Tasks Data Cruncher

Your executions will appear here.

# Collaborate in shared projects by adding members

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

Users can be added/deleted via GUI and public API

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

The screenshot displays the BioData CATALYST web application interface. The top navigation bar includes the NIH logo, 'BioData CATALYST Powered by Seven Bridges', and various menu items like 'Projects', 'Data', 'Public Gallery', 'Public projects', 'Automations', 'Developer', 'Staff', and a user profile 'alisonleaf'. Below this, a secondary navigation bar shows 'Dashboard', 'Files', 'Apps', and 'Tasks'. The main content area is titled 'Alison\_test\_GWAS' and includes an 'Interactive Analysis' section. The 'DESCRIPTION' tab is active, showing a 'Welcome to your new project!' message and a list of actions users can perform within the project. A 'MEMBERS' section on the right lists three members: alisonleaf (OWNER), dave, and milan.domazet, each with their role and a 'Manage members' button. A 'Manage members' modal is open in the foreground, showing a list of members with their roles and a 'Permissions' section. The modal also includes an 'Invite new members' section with a search bar and an 'Invite' button.

**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 your project members. Remember to run on the platform for your own notes.

You can also use markdown to format your text.

Good luck with your research! Visit the [Knowledge Center](#) for more information.

**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**

**Tasks** **Data Cruncher**

**Manage members**

**1 member** **Permissions** [\(Learn more\)](#)

- alisonleaf** OWNER  
Joined on July 2, 2020 10:04

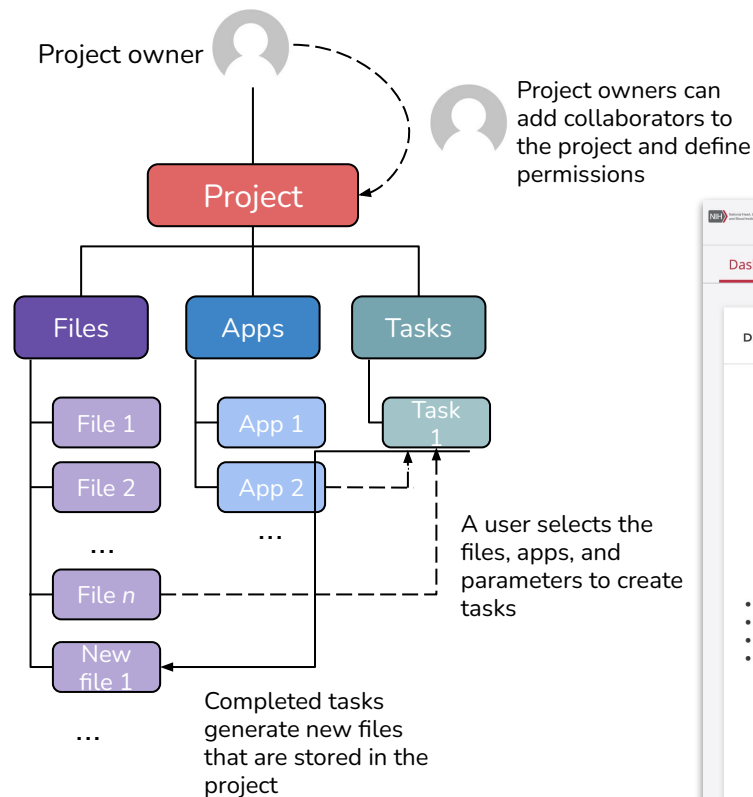
You cannot edit your own permissions.

**Invite new members**

Write, Copy, Execute [Invite](#)

**sara\_seepo**  
Sara Seepo

# Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions

**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

# Conveniently bring in your own data

## Data Tools

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

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

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

Download ▾

### Seven Bridges File System (SBFS) BETA

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

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

Download ▾

### Upload files via the API

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

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



Drag & drop files from your computer or

[Browse files](#)

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

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

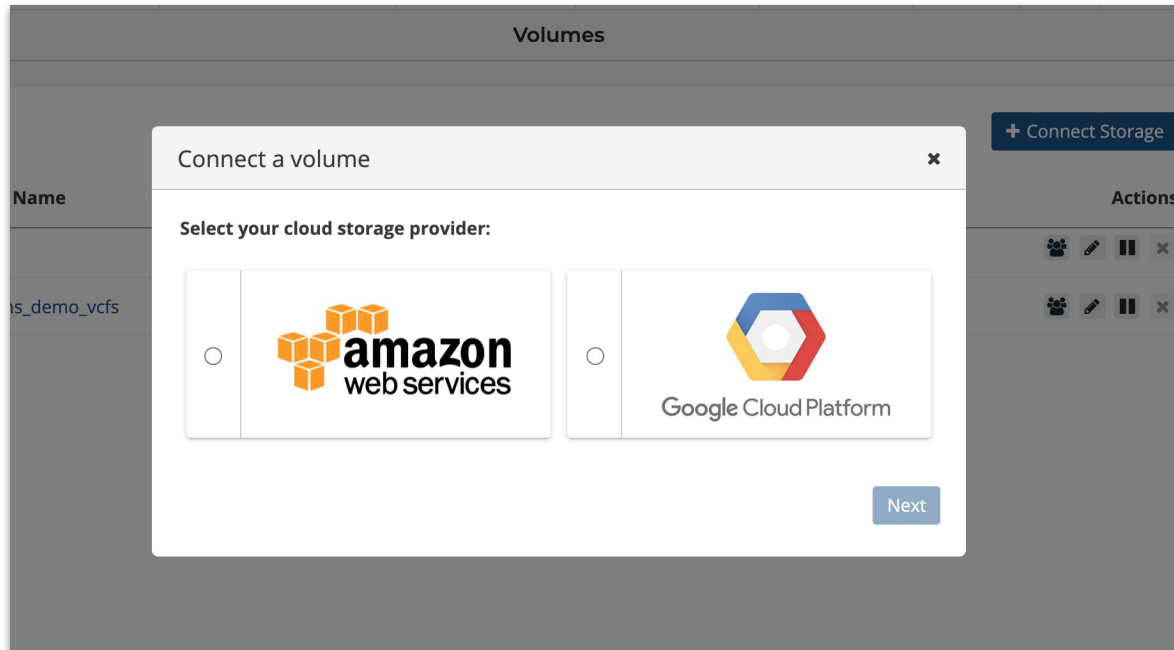
Paste the link of the file(s) you want to import

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

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

Import

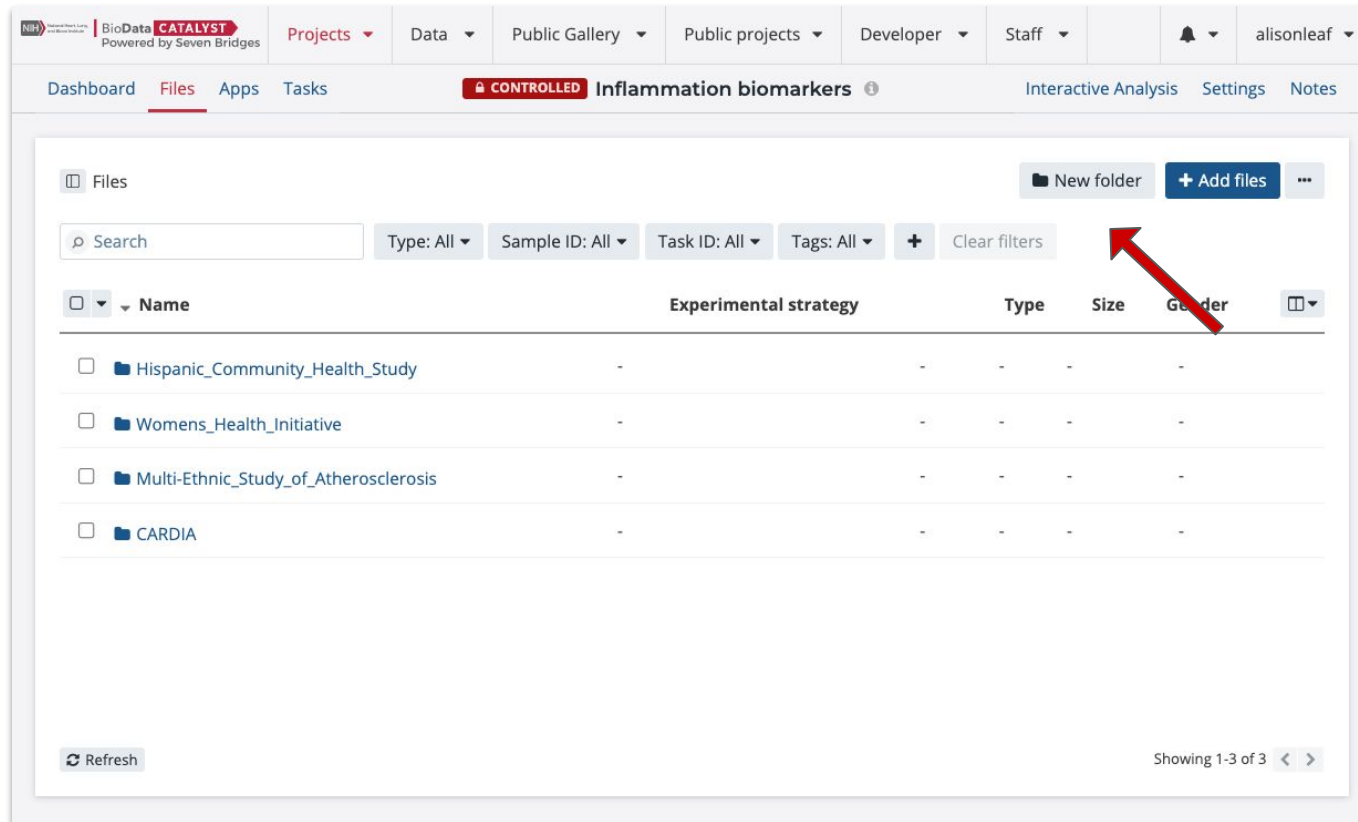
# Connect private cloud storage directly to platform



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

# Organize and manage files within projects

Nested folder structure for organizing files



The screenshot displays the BioData CATALYST interface for the 'Inflammation biomarkers' project. The 'Files' tab is active, showing a list of folders. A red arrow points to the 'New folder' button in the top right corner of the file list area.

Files

Search

Type: All Sample ID: All Task ID: All Tags: All Clear filters

Name	Experimental strategy	Type	Size	Gender
Hispanic_Community_Health_Study	-	-	-	-
Womens_Health_Initiative	-	-	-	-
Multi-Ethnic_Study_of_Atherosclerosis	-	-	-	-
CARDIA	-	-	-	-

Refresh

Showing 1-3 of 3



# Questions?

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

# Overview of the BioData Catalyst Ecosystem

Kat Thayer



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# BioData Catalyst is an ecosystem of platforms

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

## Explore Available Data

### BioData Catalyst Powered by Gen3

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

[Launch](#)

[Documentation](#) 

[Learn](#)

### BioData Catalyst Powered by PIC-SURE

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

[Launch](#)

[Documentation](#) 

[Learn](#)

# BioData Catalyst is an ecosystem of platforms

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

## Analyze Data in Cloud-based Shared Workspaces

### BioData Catalyst Powered by Seven Bridges

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

[Launch](#)

[Documentation](#)

[Learn](#)

### BioData Catalyst Powered by Terra

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

[Launch](#)

[Documentation](#)

[Learn](#)

# BioData Catalyst is an ecosystem of platforms

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

## Use Community Tools on Controlled-access Datasets

### Dockstore

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

[Launch](#)

[Documentation](#) 

[Learn](#)

## Imputation Server

### Access the Imputation Server

#### Imputation Server developed by the University of Michigan

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

[Launch](#)

[Documentation](#) 

# Introduction to *BioData Catalyst Powered by Gen3*

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

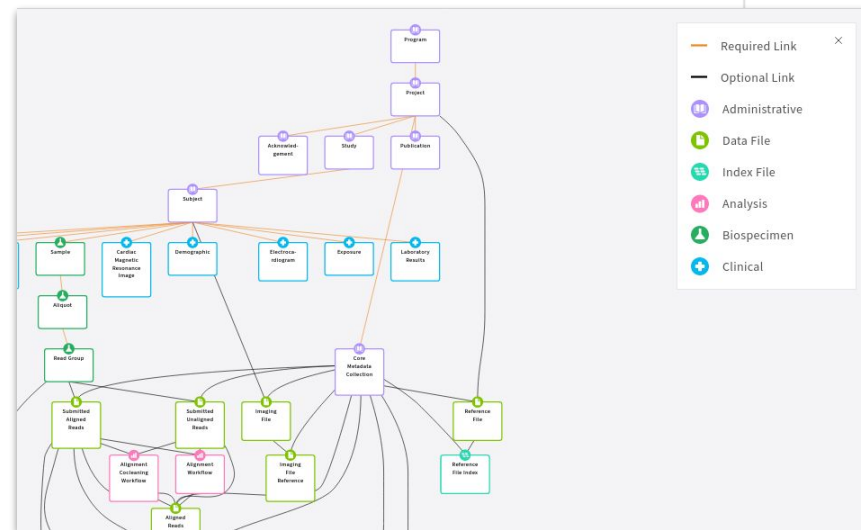
- creates pointers to data files and links them to metadata (**file information**) .

## Indexing data files

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

## Graph Model

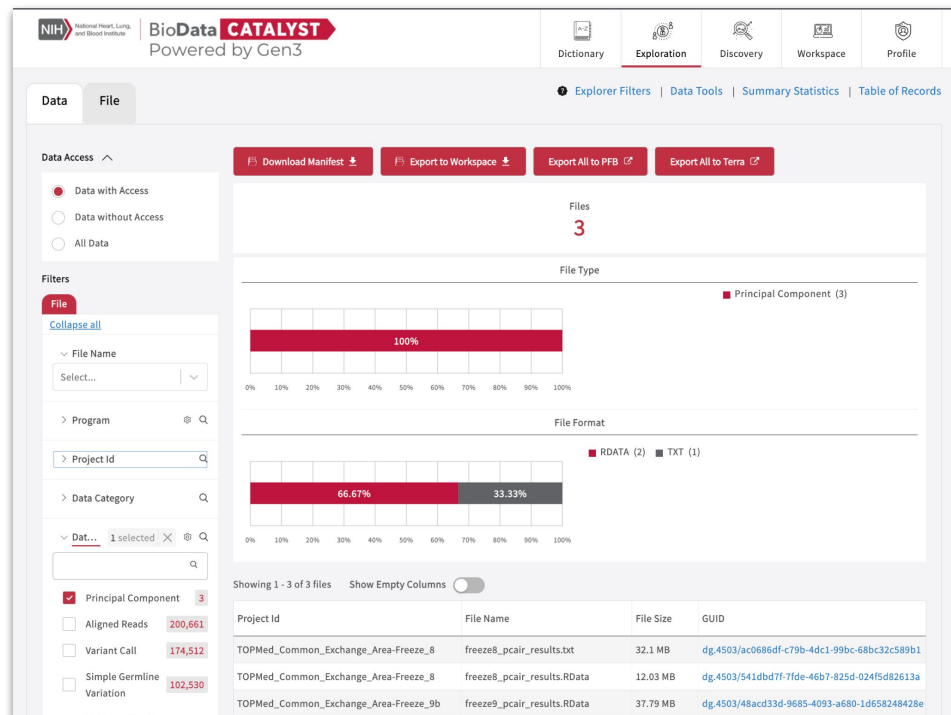
- The ability to relate metadata (**file information**) via nodes and edges
- Allows for linkage between data files and clinical information



# Introduction to *BioData Catalyst Powered by Gen3*

## Exploration

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



# About *BioData Catalyst Powered by Terra*

Terra is a scalable platform for biomedical research

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



# Terra differentiators

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



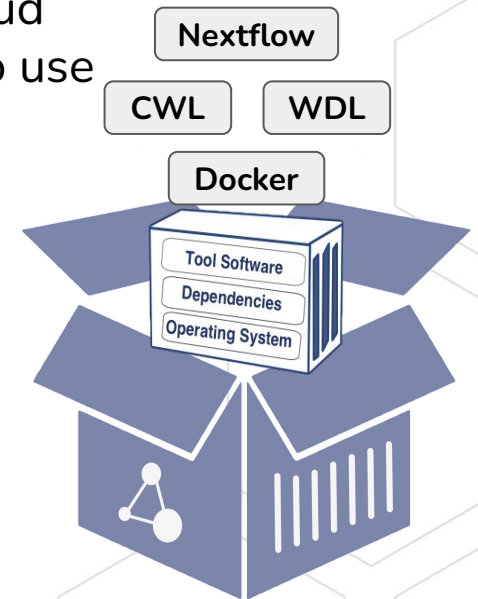
# Introduction to Dockstore

“an app store for bioinformatics”

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

## Advantages

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



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

# Publish your workflows on Dockstore!

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

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

**Workflows & Tools**

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

**About the Collection**

**Viral NGS Workflows**

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

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

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

**Overview of analytical workflows available**

```

graph TD
    FASTQ --> ImportFASTQ[Import FASTQ Data]
    SRAID[SRA ID] --> ImportSRA[Import SRA Data]
    ImportFASTQ --> uBAM[uBAM]
    ImportSRA --> uBAM
    uBAM --> Kraken[Kraken Classifier]
    uBAM --> ViralAssembly[Viral Assembly]
    Kraken --> FASTA[FASTA]
    ViralAssembly --> FASTA
    FASTA --> BuildAugur[Build Augur Tree]
    BuildAugur --> Nextstrain[Nextstrain]
    uBAM -.-> MapRef[Map to reference sequence]
    MapRef -.-> Scaffold[Scaffolding]
    Scaffold -.-> Refine[Refining to remove partner sequences & low quality reads]
    Refine -.-> RefineAssembly[Refine assembly]
  
```

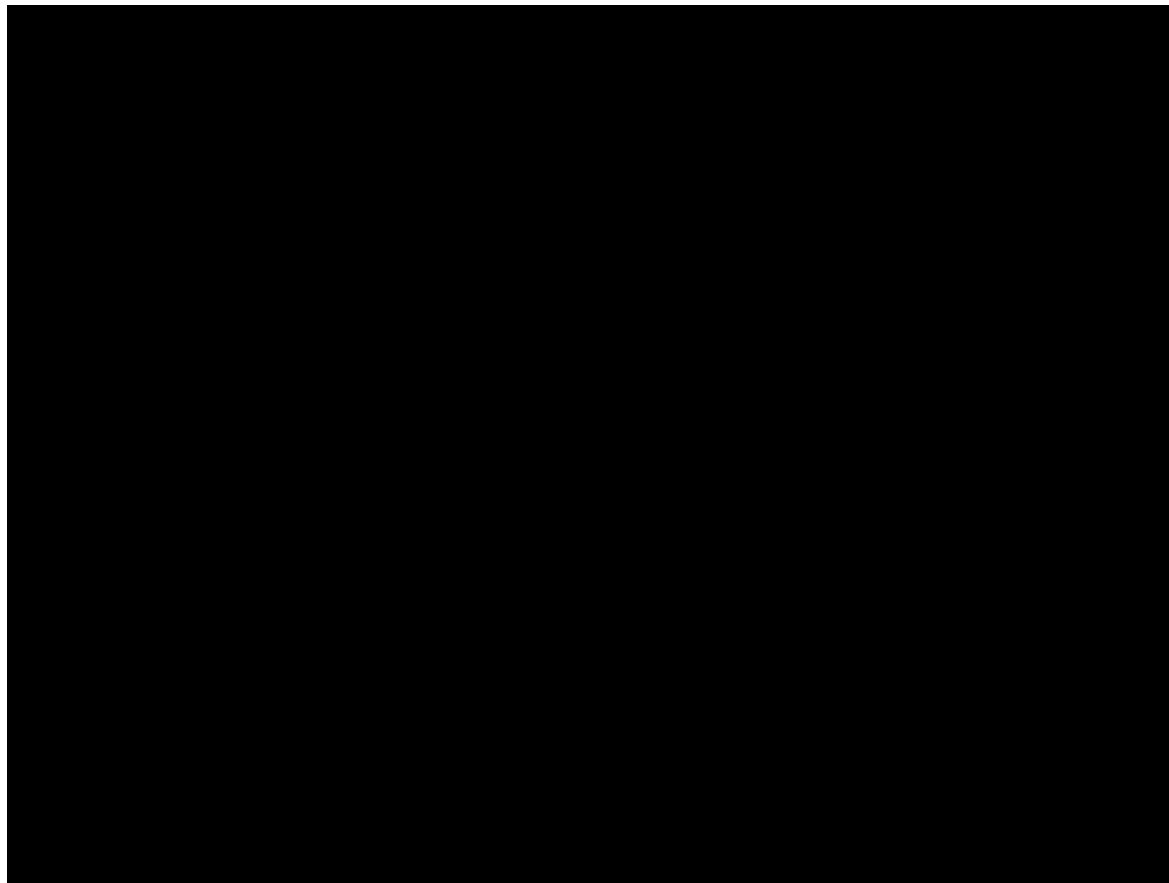
**Tutorials**

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

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

[github.com/broadinstitute/viral-pipelines/beast\\_gpu](https://github.com/broadinstitute/viral-pipelines/beast_gpu)

**We conducted all analyses using viral-ngs 2.0.21 on the Terra platform (app.terra.bio). All of the workflows named below are publicly available via the Dockstore Tool Registry Service (<https://dockstore.org/organizations/BroadInstitute/collections/pgs>). Code is also archived at [doi:10.5281/zenodo.4306358](https://doi.org/10.5281/zenodo.4306358) and [doi:10.5281/zenodo.4306362](https://doi.org/10.5281/zenodo.4306362). Briefly, samples**



Import **xvcfmerge** workflow from Dockstore

# Additional Information

## Useful links

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

## Accessing genomic data via the GA4GH DRS standard

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

## Workspace tutorial on Gen3 data

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

# Questions?

# End of Day One Material



# BioData Catalyst Workshop, Day Two

Friday, November 18th at 11 am ET

**We will get started shortly.**



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

Interact with us on our forum during today's workshop:

<https://bit.ly/BDC-Howard-Workshop>

# BioData Catalyst Workshop, Day Two

Friday, November 18th at 11 am ET

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



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

Interact with us on our forum during today's workshop:

<https://bit.ly/BDC-Howard-Workshop>

# Statement of Conduct

The BioData Catalyst Consortium is dedicated to **providing a harassment-free experience for everyone**, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, or religion (or lack thereof). We do not tolerate harassment of community members in any form. Sexual language and imagery is generally not appropriate for any venue, including meetings, presentations, or discussions.

Resource: [Statement of Conduct](#)

# Agenda

## Day One: Thursday, November 17th

Topic	Time
<a href="#">Introductions and Housekeeping</a>	5 min
<a href="#">What is BioData Catalyst?</a>	15 min
<a href="#">Researcher Presentation and Q&amp;A: Dr. Fayuan Wen</a>	30 min
<b>Break - 20 min</b>	
<a href="#">Interactive Demo: Finding and Using NHLBI Hosted Data</a>	1 hr
<a href="#">Bring Your Own Data</a>	20 min
<a href="#">Overview of the BioData Catalyst Ecosystem</a>	10 min
Q&A	20 min

## Day Two: Friday, November 18th

Topic	Time
<a href="#">Tools, Workflows, and Interactive Analysis</a>	10 min
<a href="#">Understanding, Estimating, and Managing Cloud Costs</a>	15 min
<a href="#">Running a GWAS on BioData Catalyst Powered by Seven Bridges</a> , Part 1	1 hr
<b>Break - 20 min</b>	
<a href="#">Running a GWAS on BioData Catalyst Powered by Seven Bridges</a> , Part 2	1 hr
Q&A	30 min

# Tools, Workflows, and Interactive Analysis

Dave Roberson, Community Engagement Specialist  
at Seven Bridges

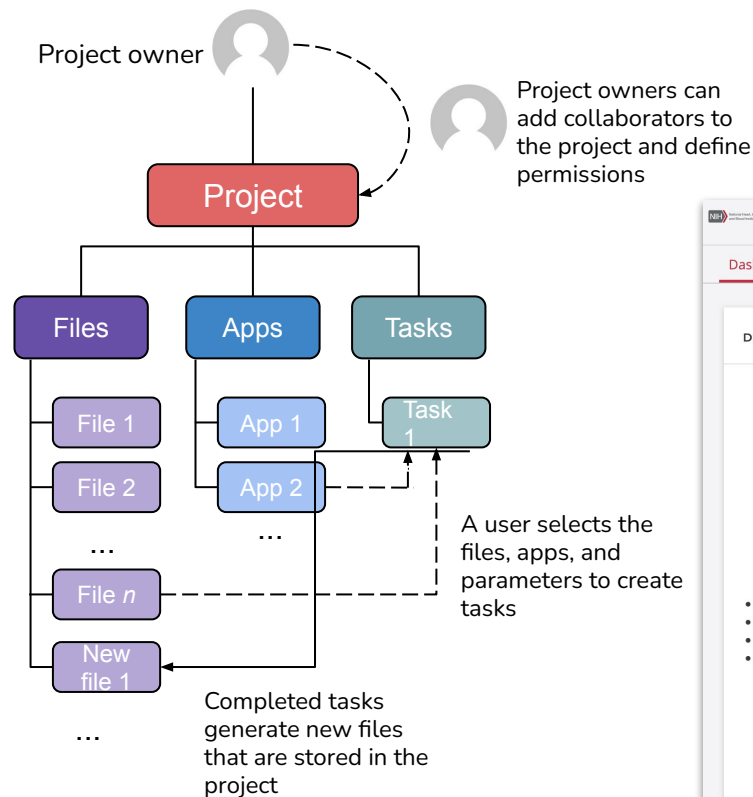


National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions

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

# 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, the 'Environment setup' section shows a dropdown menu currently set to 'SAS Data Science'. Navigation buttons 'Previous' and 'Next' are located at the bottom right of the dialog.

# User friendly workflow editor enables reproducibility by default

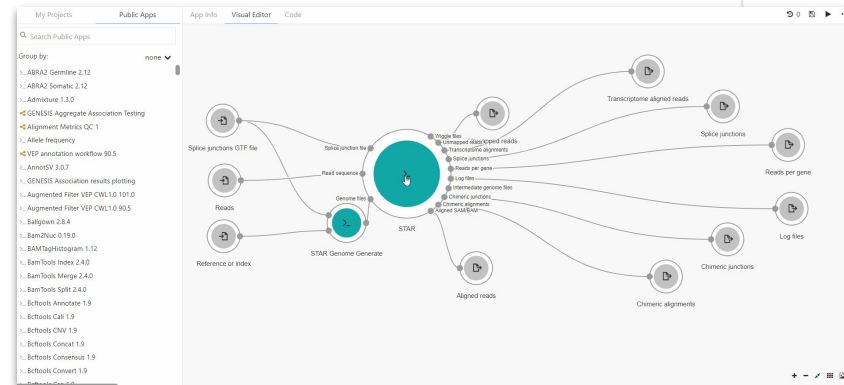
Common Workflow Language enables **portability, reproducibility, and scalability**

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

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

Create your own tools with our CWL Tool Editor

Expose or lock parameters appropriately

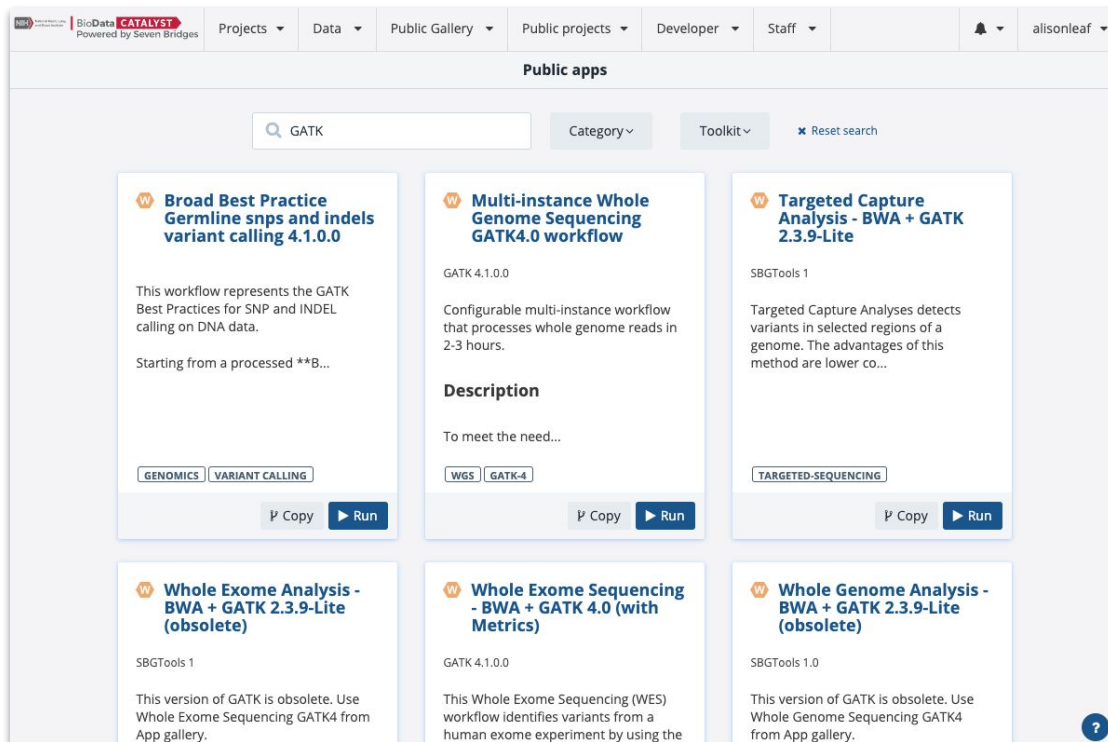




# Find the tools you need in the Public Apps Gallery

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

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



# Run association pipelines out of the box

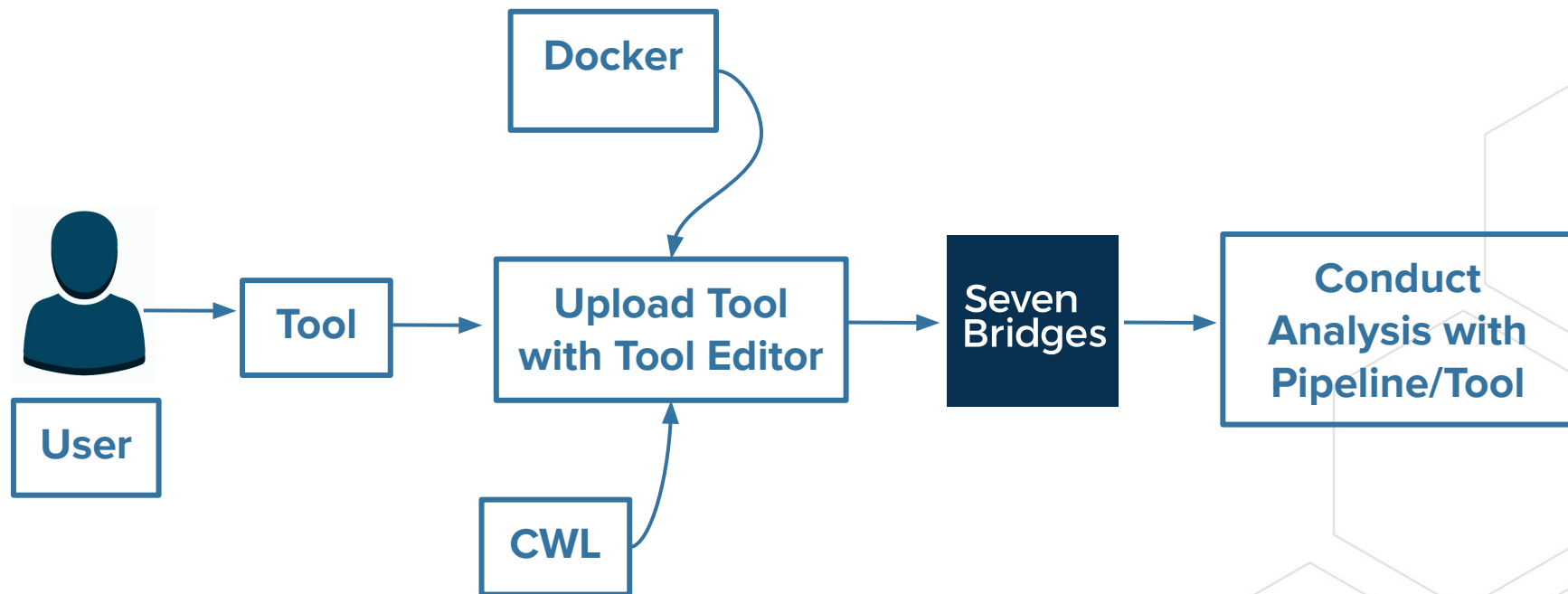
- GENESIS
- Plink
- EPACTS
- STAAR (coming soon)

The screenshot displays the BioData CATALYST Public apps interface. The top navigation bar includes the NIH logo, BioData CATALYST branding, and a series of dropdown menus for Projects, Data, Public Gallery, Public projects, Developer, and Staff. A user profile for 'alisonleaf' is visible on the right. The main section is titled 'Public apps' and features a search bar with 'GENESIS' entered, along with 'Category' and 'Toolkit' dropdowns and a 'Reset search' button. Below the search bar, six app cards are displayed in a grid:

- GENESIS Aggregate Association Testing**: Describes an 'Aggregate Association Testing workflow' that runs aggregate association tests using Burden, SKAT [1], fastSKAT [2], and SMM... It includes tags for GWAS and CWL1.0, and buttons for 'Copy' and 'Run'.
- GENESIS Null Model**: Describes a 'Null Model workflow' that fits regression or mixed effects models under the null hypothesis of no genotype effects. It includes tags for GWAS, CWL1.0, and GENOMICS, and buttons for 'Copy' and 'Run'.
- GENESIS Single Variant Association Testing**: Describes a 'Single Variant workflow' that runs single-variant association tests, consisting of several steps to define segments. It includes tags for GWAS, CWL1.0, and GENOMICS, and buttons for 'Copy' and 'Run'.
- GENESIS Sliding Window Association Testing**: Describes a 'Sliding Window Association Testing workflow' that runs sliding-window association tests, using Burden, SKAT [1], etc. It includes a 'Run' button.
- GENESIS VCF to GDS**: Describes a 'VCF to GDS workflow' that converts VCF or BCF files into Genomic Data Structure (GDS) format. It includes a 'Run' button.
- Fusion Transcript Detection - ChimeraScan**: Describes 'Fusion Transcript Detection - ChimeraScan 1.0', which detects and identifies fusion transcripts from paired-end RNA-Seq data using...

A help icon (?) is located in the bottom right corner of the interface.

# Bringing custom tools to the platform



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

Only one input per task can be selected for batching.

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

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

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

Task Inputs Execution Settings

**Inputs**

Batching ☒ On [Change selection](#)

Batch by: File

This will create one task for each selected item.

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

**App Settings**

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

- GATK HaplotypeCaller** (RGATK\_HaplotypeCaller)  
Memory Per Job
- GATK BaseRecalibrator** (RGATK\_BaseRecalibrator)  
Intervals String
- SAMtools Index** (ISAMtools\_Index)  
Number of threads
- Picard MarkDuplicates** (RPicard\_MarkDuplicates)  
Memory per job
- BWA MEM Bundle 0.7.17**  
(BWA\_MEM\_Bundle\_0\_7\_17)  
[Memory for BWA memtool](#)

**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 nevernameu Batch by: File  
Spot Instances: On ☐ Memorization: Off ☐ Price: \$2392.30 ☐

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

Search task names  Status: All

Task Name	Submitted by	Submitted on	App	Duration	Status	Actions
<input type="checkbox"/> Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics) run - 11-29-18 03:15:04: file: 1.cram	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	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	nevernameu	Nov. 29, 2018 03:26	Whole Genome Sequencing - BWA + GATK 4.0 (with Metrics)	16 hours, 57 minutes	COMPLETED	<a href="#">C</a>

# Detailed documentation and tutorials

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

BIODATA CATALYST POWERED BY SEVEN BRIDGES

### Objective

### Helpful terms to know

### User Accounts & Billing Groups

#### Further reading

### Tips for Running Tools/Workflows

#### Start with the descriptions

#### Test the workflow

#### Specify computational resources

#### Learn about Instance Profiles

#### Scale up with Batch Analysis

#### Parallelize with Scatter

#### Configuring default computational resources

### Further analysis and interpretation of your Results

#### Getting started

#### JupyterLab environment

#### Accessing the files

#### Saving the created files

## OBJECTIVE

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

## HELPFUL TERMS TO KNOW

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

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

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

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

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

## Troubleshooting Failed Tasks

BIODATA CATALYST POWERED BY SEVEN BRIDGES

### Helpful terms to know

### Getting started

### Examples: Quick & Unambiguous

#### Task 1: Docker image not found

#### Task 2: Insufficient disk space

#### Task 3: Scatter over a non-list input

#### Task 4: Automatic allocation of the required instance is not possible

#### Task 5: JavaScript evaluation error due to lack of metadata

#### Task 6: Invalid JavaScript indexing

#### Task 7: Insufficient memory for Java process

### Examples: File compatibility challenges

#### Task 8: STAR reports incompatible chromosome names

#### Task 9: RSEM reports incompatible chromosome names

#### Task 10: Incompatible alignment coordinates

Examples: When error messages are not enough

#### Task 11: Invalid command line

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

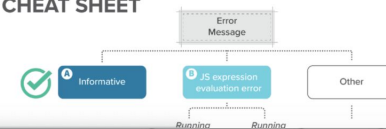
Often the first step to a user becoming comfortable using BioData Catalyst powered by Seven Bridges is their gaining confidence in resolving issues they encounter on their own. This confidence usually comes with experience – the experience with bioinformatics tools and Linux environment in general, but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

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

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

## Troubleshooting CHEAT SHEET

SevenBridges



Visit the Knowledge Center

# Getting Help - Contacting Support from the platform

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

The screenshot displays the BioData CATALYST platform interface. At the top, there is a navigation bar with tabs for Projects, Data, Public Gallery, Public projects, Developer, and Staff. Below this, a secondary bar shows 'Dashboard', 'Files', 'Apps', and 'Tasks'. The main content area is titled 'Genesis tutorial' and includes sections for 'DESCRIPTION' and 'MEMBERS'. A modal window titled 'Need help?' is open, providing links to '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'. It also includes a 'Contact our support' section with a text input field 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. The button features a question mark icon and the text 'Help and support'.

# Questions?

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

# Understanding, Estimating, and Managing Cloud Costs

Dave Roberson, Community Engagement Specialist  
at Seven Bridges



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**



# Agenda

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

# What are cloud costs?

Three categories of costs:

- Storage
- Compute
- Egress

Web resource: [Cloud Costs and Credits](#)

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

# Estimating Cloud Costs

# Estimating Cloud Costs

Researchers incur fees for:

- Data Storage
- Computing / Analysis
- Egress charges

## Estimate and Manage Your Cloud Costs

### Overview

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

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

### Background

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

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

[View Cloud Cost Guide](#)

# Estimating Cloud Costs

## Data Storage

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

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

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

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

# Estimating Cloud Costs

## Computing / Analysis

Compute costs vary and depend on a range of factors:

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

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

# Estimating Cloud Costs

## Egress Charges

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

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

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

# Planning costs for GWAS pipelines

## GENESIS Benchmarking Guide

### Introduction

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

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

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

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

View GENESIS Guide and Benchmarking

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



# Managing Cloud Costs

# Tasks have detailed credit usage information

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

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

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

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

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

**Inputs** [?](#)

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

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

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

**App Settings**

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

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

**Outputs** [?](#)

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

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

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

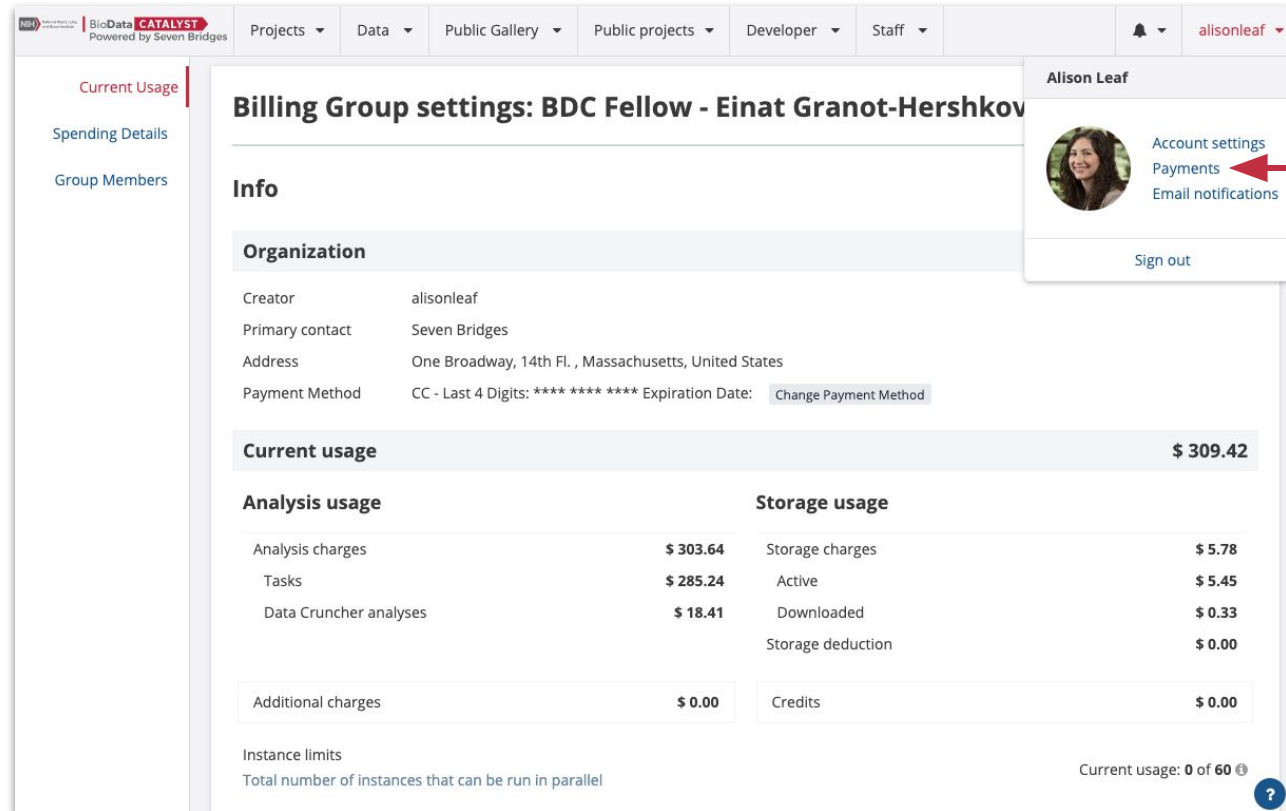
- demo\_height\_chr1.RData
- demo\_height\_chr2.RData
- demo\_height\_chr3.RData
- demo\_height\_chr4.RData
- demo\_height\_chr5.RData
- ...and 17 more items

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

- demo\_height\_report.html

# Track costs on platform payments page

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



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

**Info**

**Organization**

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

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

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

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

Current usage: 0 of 60 ?

**Alison Leaf**

- Account settings
- Payments
- Email notifications

[Sign out](#)

# Funding Cloud Costs

# Try out the ecosystem with Pilot Credits

If you don't already have CWL tools or WDL tools and are flexible about which BioData Catalyst workspace to use, **we recommend trying both** to make an informed decision about which platform is the best fit for you.

BioData Catalyst users may request one of the following: \*

\$500 in initial pilot cloud credits to begin a project or explore the ecosystem

Select your preferred analysis platform \* (or choose to explore both)

✓ Select One

\$500 on Seven Bridges

\$500 on Terra

\$250 each on both Seven Bridges and Terra

# Cloud Credits Workflow

1

**Sign up for the community**

Sign up at  
[biodatacatalyst.nhlbi.nih.gov/contact/ecosystem](https://biodatacatalyst.nhlbi.nih.gov/contact/ecosystem)

2

**Sign up for a workspace**

Seven Bridges and/or  
Terra

3

**Apply for Pilot Credits**

Fill out the [Cloud Credits Request form](#).

Use all credits on a single platform, or split.

4

**Apply for additional credits or pay yourself**

Cover costs after pilot funding has been exceeded.

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

# Requesting grant funding for BioData Catalyst

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

## Writing BioData Catalyst into a Grant Proposal

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

### Writing BioData Catalyst into your proposal's budget

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

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

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

**View BioData Catalyst  
Grant Guide**

# Questions?

## Take a moment to request pilot funds

**Next up:** Running a GWAS on BioData Catalyst



# Running a GWAS on BioData Catalyst

Dave Roberson, Seven Bridges



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# The TOPMed analysis Pipeline & GENESIS R/Bioconductor package

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

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

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

# The TOPMed analysis Pipeline & GENESIS R/Bioconductor package

## Genetic Association in two steps

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

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

The screenshot shows the Bioinformatics journal article page. The header includes the journal name 'Bioinformatics' and the ISCB logo. The navigation bar contains links for Issues, Advance articles, Submit, Purchase, Alerts, and About, along with a search bar. The article title is 'Genetic association testing using the GENESIS R/Bioconductor package'. The authors listed are Stephanie M Gogarten, Tamar Sofer, Han Chen, Chaoyu Yu, Jennifer A Brody, Timothy A Thornton, Kenneth M Rice, and Matthew P Conomos. The article is from Volume 35, Issue 24, published on 15 December 2019. The article content section lists: Abstract, 1 Introduction, 2 Genetic association testing, 3 Sparse GRM/KM for efficient computation, 4 Discussion, Funding, References, and Supplementary data. The abstract summary states: 'The Genomic Data Storage (GDS) format provides efficient storage and retrieval of genotypes measured by microarrays and sequencing. We developed GENESIS to perform various single- and aggregate-variant association tests using genotype data stored in GDS format. GENESIS implements highly flexible mixed models, allowing for different link functions, multiple variance components and phenotypic heteroskedasticity. GENESIS integrates cohesively with other R/Bioconductor packages to build a complete genomic analysis workflow entirely within the R environment.'

Bioinformatics

ISSUES Advance articles Submit Purchase Alerts About All Bioinformatics Search Advanced Search

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

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

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

Published: 22 July 2019 Article history ▼

PDF Split View Cite Permissions Share ▼

**Article Contents**

Abstract

1 Introduction

2 Genetic association testing

3 Sparse GRM/KM for efficient computation

4 Discussion

Funding

References

Supplementary data

**Abstract**

**Summary**

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

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

# GENESIS Workflows on BioData Catalyst

**Tutorial project on *BioData Catalyst Powered by Seven Bridges*:**

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

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

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

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

# Learning objectives

## Part 1: Getting Started

Link hosted files

Create a project

Launch Data Cruncher

## Part 2: Interactive Analysis

Work in a Seven Bridges  
interactive environment to:

- Convert VCF files
- Explore data
- Harmonize phenotypes

## Part 3: Tools/Workflows

Use CWL apps to:

- Fit a Null Model
- Run a single variant association test
- Monitor task progress

# 30 MIN BREAK

We will reconvene at 12:45 pm ET

**UP NEXT: GENESIS Workflows**



National Heart, Lung,  
and Blood Institute

BioData

**CATALYST**

# Questions?

# Feedback on Workshop



# Thank you for joining us

Join the Community

Interact with the forum



Subscribe to our [YouTube channel](#)

Register for Nov 30 Community Hours