BioData Catalyst Workshop, Day One

Thursday, November 17th at 11 am ET

We will get started shortly.

NIH 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.

NIH 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

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

Agenda

Day One: Thursday, November 17th

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

Day Two: Friday, November 18th

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

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

Introductions and Housekeeping

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

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

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

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

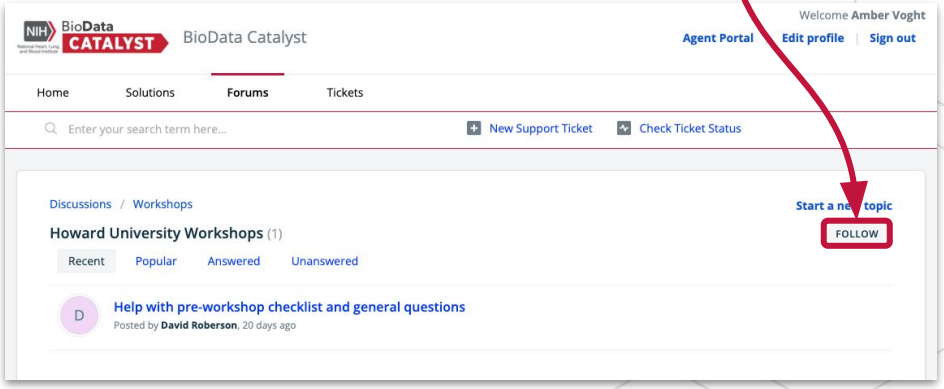
Have a question during the workshop?

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



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

Questions before we begin?

Next up: What is BioData Catalyst?

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

What is BioData Catalyst?

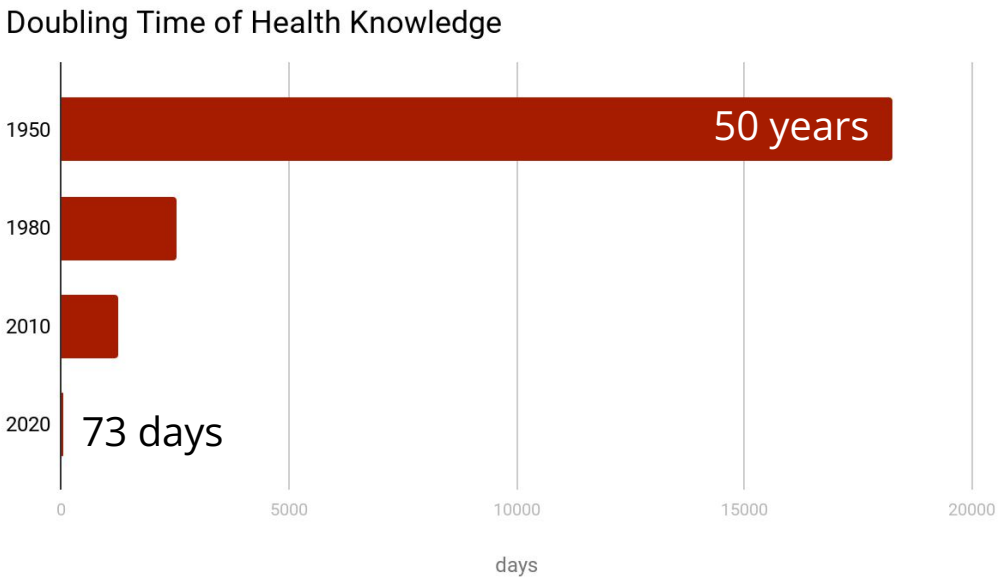
Ingrid Borecki

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

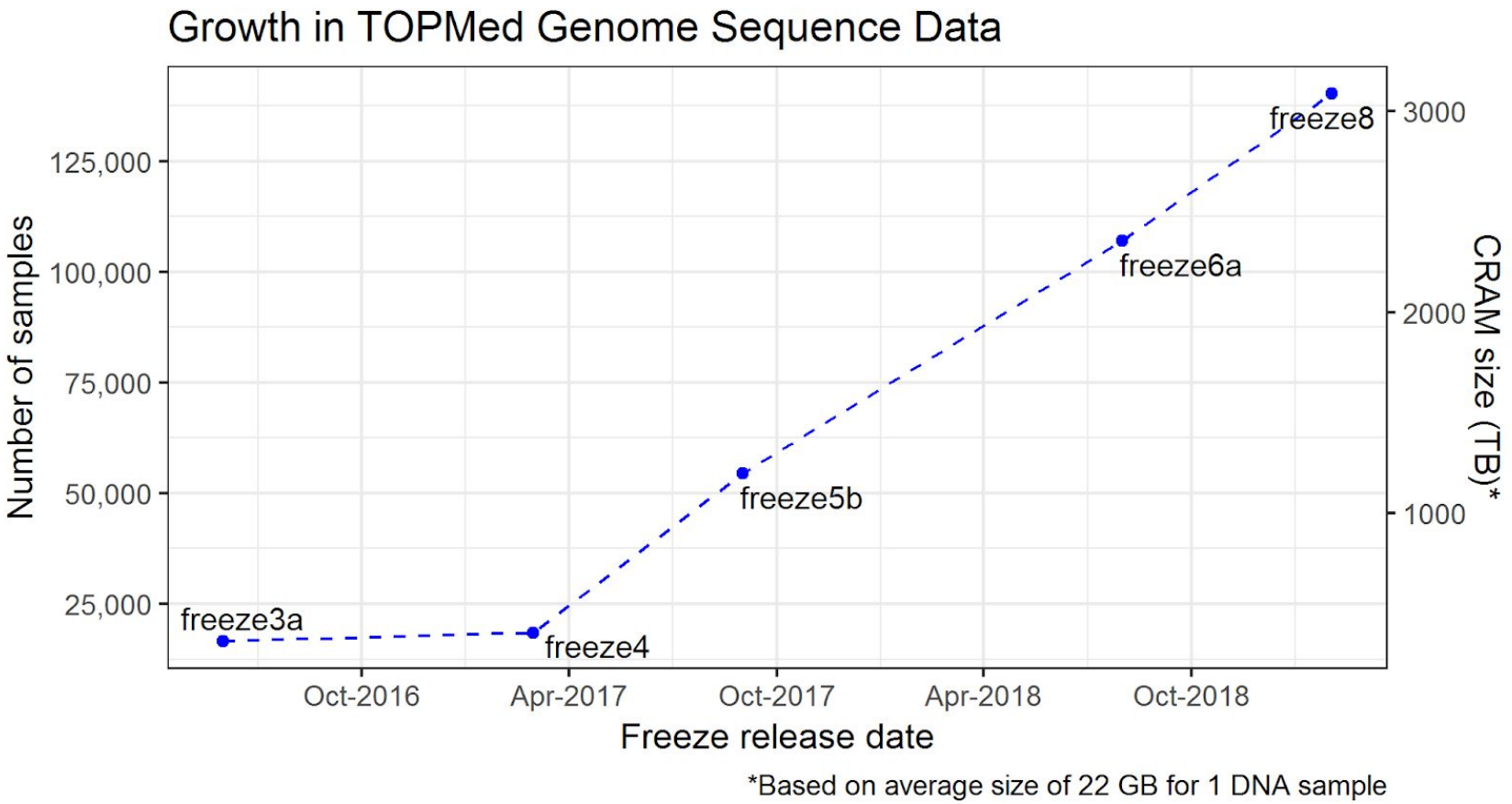
Let's talk about:

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

The rate of data generation is accelerating rapidly

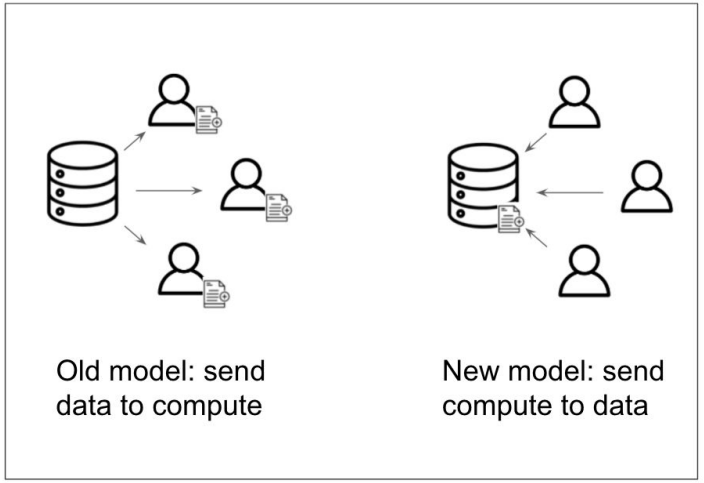


- More biomedical data will be generated this year than all previous years **combined**
- Diverse data modalities including Health data, Survey, Sequencing, Imaging, Metabolomics, Proteomics, Sensor, E-Phys, Flow Cytometry, and so on


Growth in TOPMed Genome Sequence Data
Number of samples
125,000
100,000
75,000
50,000
25,000
CRAM size (TB)*
3000
2000
1000
freeze3a
freeze4
freeze5b
freeze6a
freeze8
Oct-2016
Apr-2017
Oct-2017
Apr-2018
Oct-2018
Freeze release date
*Based on average size of 22 GB for 1 DNA sample

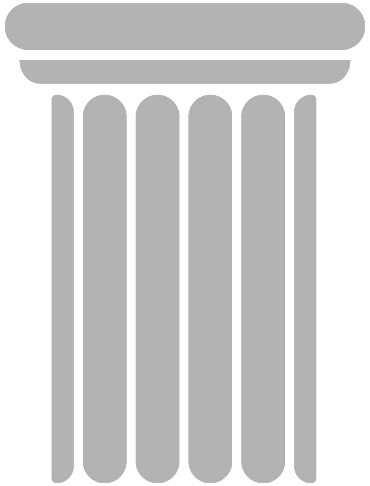
Using the Cloud to store and analyze growing health data

- 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.

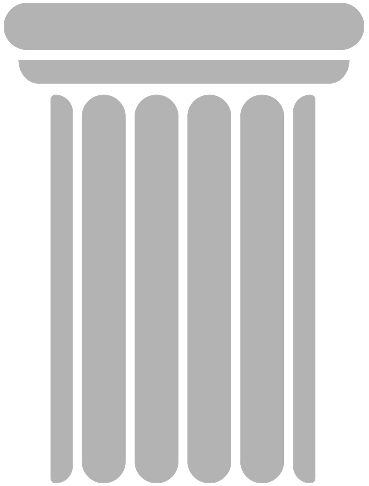


NHLBI BioData Catalyst

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.

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

WHO?

WHAT?

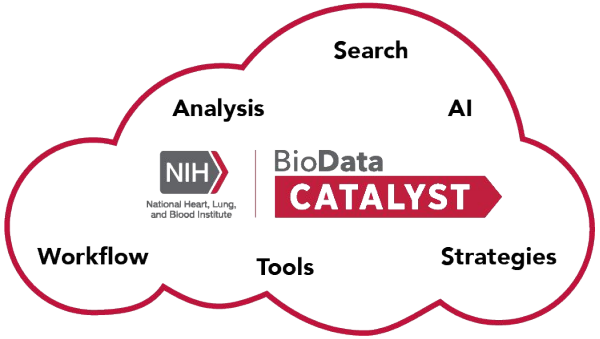
Genomics

Clinical

Imagery

DATA HARMONIZATION

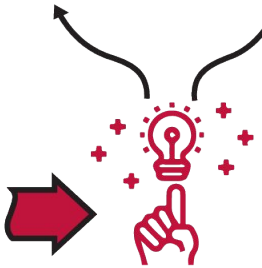
WHERE?



- UNDERSTAND
- OPEN SCIENCE
- CROSS-LINK
- COLLABORATE
- SCALE
- SHARE
- INTEROPERATE

HOW?

SCIENCE!



Prevention Strategies

WHY?

PATIENTS!

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

What BioData Catalyst offers

Managing the Computing Environment

Elastic Computing

Easier Access to many High Value Datasets

Tooling

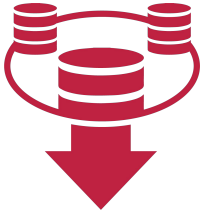
Data Discovery

Statistical Analysis Tools (R, SAS)

Other Specialized Workflows

Community and Peer Interactions

The Computing Environment



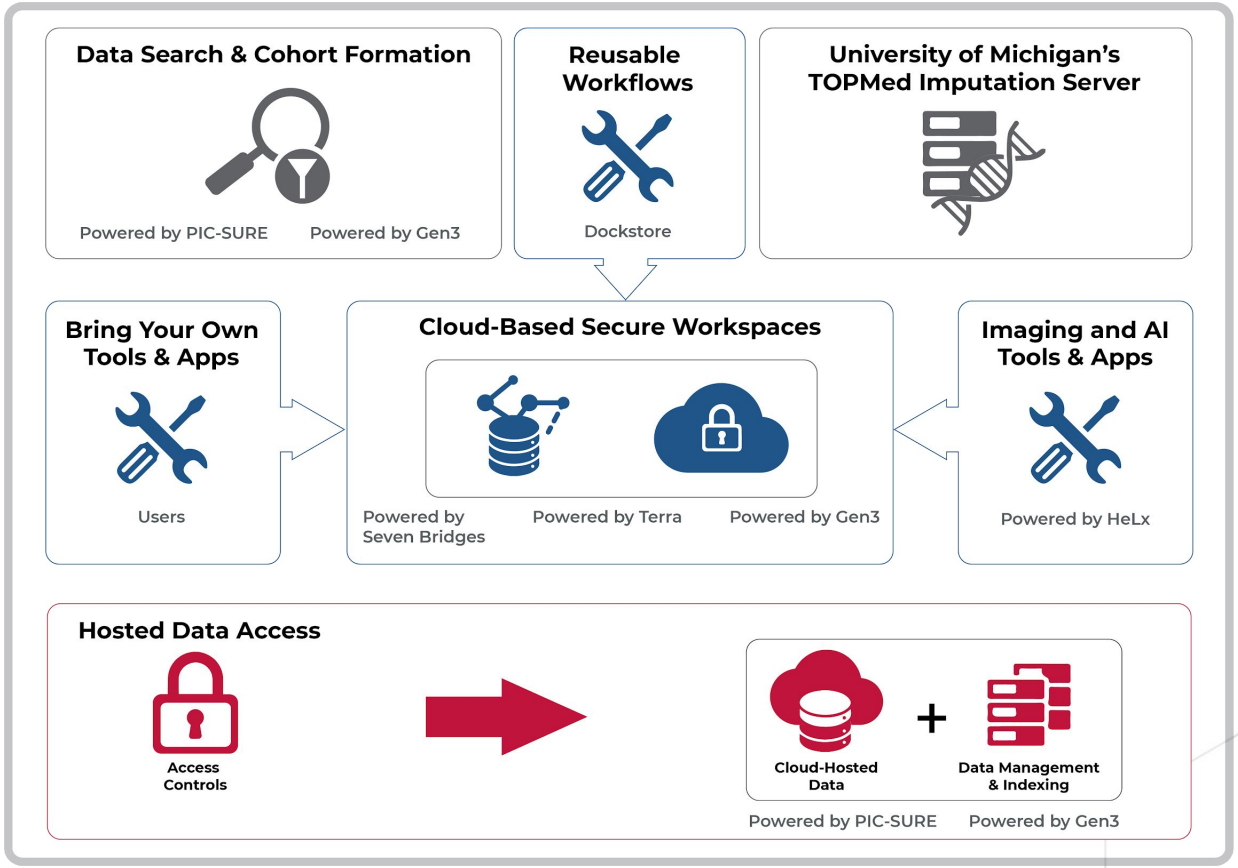
No need to **download** and **manage** (multiple) large datasets

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

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

Help desk and **documentation**

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



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

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

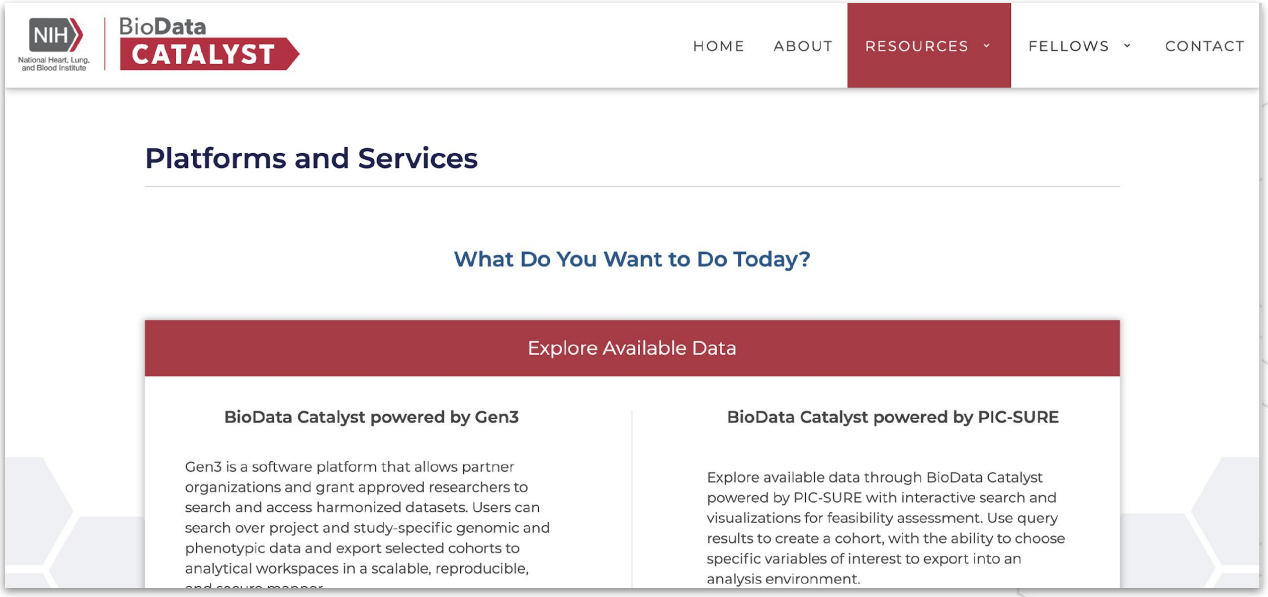
Workshop Forum: https://bit.ly/BDC-Howard-Workshop

Platforms and Services

Explore Data
- PIC-SURE
- Gen3

Analyze Data
- Seven Bridges
- Terra

Community Tools
- Dockstore



Web resource: Services

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

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

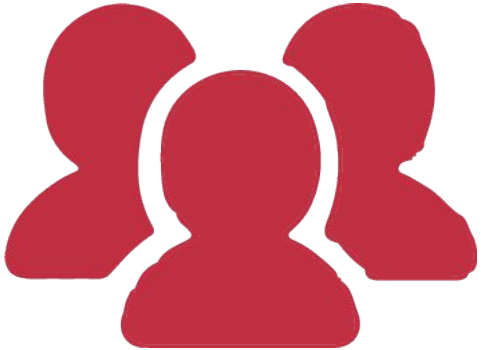
NIH National Heart, Lung, and Blood Institute | BioData CATALYST

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

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

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

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

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

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

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

Presenter: Fayuan Wen, PhD

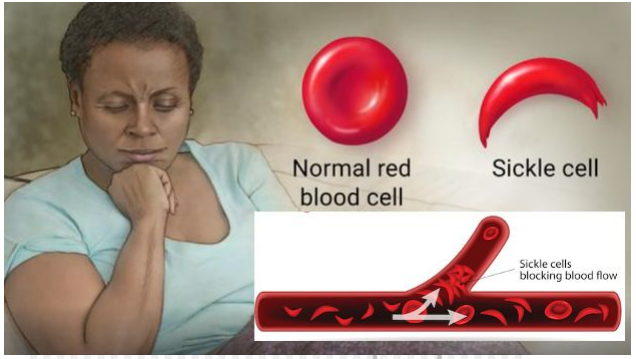
Postdoctoral Associate, Center for Sickle Cell Disease

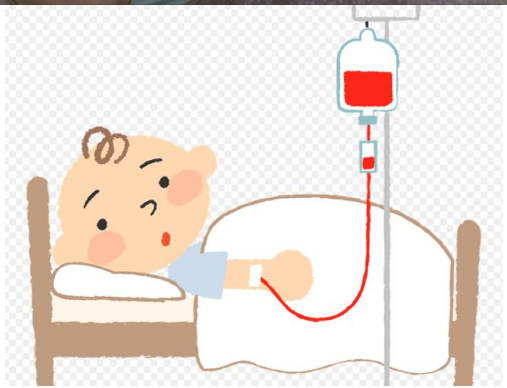
Adjunct Lecturer, Biology Department

Howard University

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

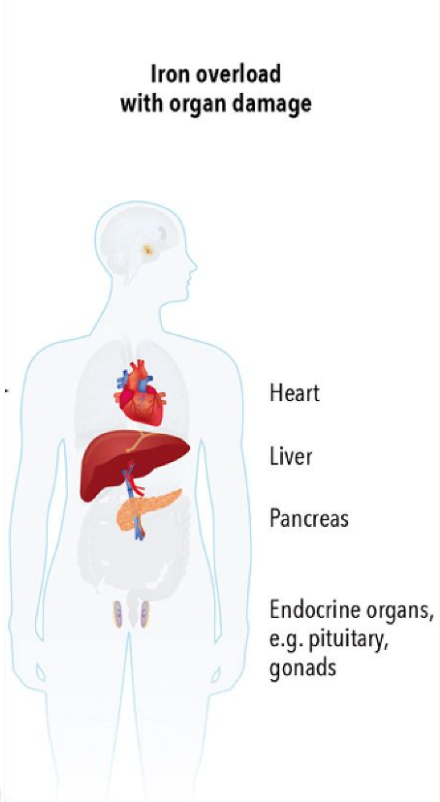
Introduction





- 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

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

→

GENESIS Single Variant Association Testing on Seven Bridges

↓

Result visualization
GENESIS LocusZoom
Annotation
TOPMed Annotation Explorer

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

GENESIS Single Variant Association Test on Seven Bridges

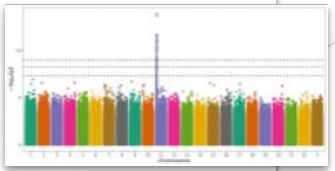
COMPLETED **20201214FourCohortsFerritinLOGGENESISSingleVariantAssociation**

Get support | View stats & logs | Edit and rerun

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[illegible].filtered.gds
  - chr11FourCohortsAc4Qual1[illegible].filtered.gds
  - chr12FourCohortsAc4Qual1[illegible].filtered.gds
  - chr13FourCohortsAc4Qual1[illegible].filtered.gds
  - chr14FourCohortsAc4Qual1[illegible].filtered.gds
  - ...and 18 more items
- ***Known hits file***
  - *No files selected*
- **Null model file**
  - 20200904ferritinLOGNullmode[illegible]_null_model_invnorm.RData
- **Phenotype file**
  - 20200904ferritinLOG[illegible]ullmodel_phenotypes.RData
- ***Variant Include Files***
  - *No files selected*

App Settings — Show non-default

- **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

Interactvie 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	allele.index	n.obs	freq	MAC	Score	Score.SE	Score.Stat	Score.pval	Est	Est.SE	PVE
1	24036289	11	102993	1	3410	0.0082111437	56	-10.0861361	16.419693	-0.612802199	0.54000714	-0.0372304754	0.06075447	1.106442e-04
2	24036477	11	116798	1	3410	0.0014662757	10	-4.6424502	6.851970	-0.677139800	0.49831726	-0.0987664460	0.14585828	1.350967e-04
3	24036510	11	119612	1	3410	0.0008797654	6	7.4686349	5.307531	1.407176853	0.15937495	0.2651283282	0.18841152	5.834257e-04
4	24036558	11	121159	1	3410	0.0180351906	123	31.6194804	24.439475	1.293787214	0.19573893	0.0529184207	0.04091741	4.931896e-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	0.0029325513	20	-5.4852965	9.749215	-0.562639833	0.57368016	-0.0577112977	0.10257236	9.327153e-05
7	24037096	11	135572	1	3410	0.0008797654	6	-6.4108370	5.316980	-1.205728908	0.22792203	-0.2267694833	0.18807667	4.283389e-04
8	24037238	11	144932	1	3410	0.0008797654	6	-2.8786843	4.861025	-0.592196992	0.55371868	-0.1218255425	0.20571794	1.033286e-04
9	24037337	11	148591	1	3410	0.0193548387	132	15.1015935	24.607130	0.613708043	0.53940827	0.0249402529	0.04083863	1.109716e-04
10	24037606	11	161365	1	3410	0.0032258065	22	-12.7818340	10.016592	-1.276066131	0.20193214	-0.1273952370	0.09983435	4.797716e-04
11	24037723	11	168366	1	3410	0.0058651026	40	4.3296083	13.572442	0.318999941	0.74972656	0.0235035029	0.07367871	2.998261e-05
12	24037726	11	168589	1	3410	0.0121700880	83	18.9983867	19.501318	0.974210592	0.32995209	0.0499561308	0.05127859	2.796364e-04
13	24037825	11	172743	1	3410	0.0205278592	140	-0.4604732	25.389460	-0.018136393	0.98553005	-0.0007143276	0.03938842	9.691478e-08
14	24038147	11	176614	1	3410	0.0126099707	86	-32.5308575	19.555075	-1.663550645	0.09620226	-0.0850700215	0.05113762	8.153805e-04

Showing 1 to 16 of 1,519,526 entries, 14 total columns

	variant.id	chr	pos	Score.pval
column 0: rownames	11	4932190	4.449242e-08
2	24149360	11	4991804	7.489210e-10
3	24149363	11	4991953	1.663058e-08
4	24149370	11	4992597	2.772388e-08
5	24149513	11	5000177	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	topmed
24149360	11	topmed
24149363	11	topmed
24149370	11	topmed
24149513	11	topmed
24149522	11	topmed
24149524	11	topmed
24149553	11	topmed

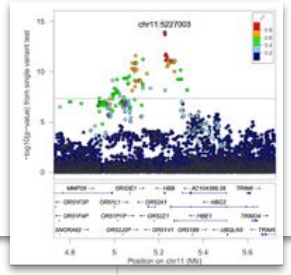
App Settings Show non-default

Genome build	hg38
Locus type	variant
Number of CPUs	4
Output prefix	20210217Chr11LocusZoom
Significance line	5e-8

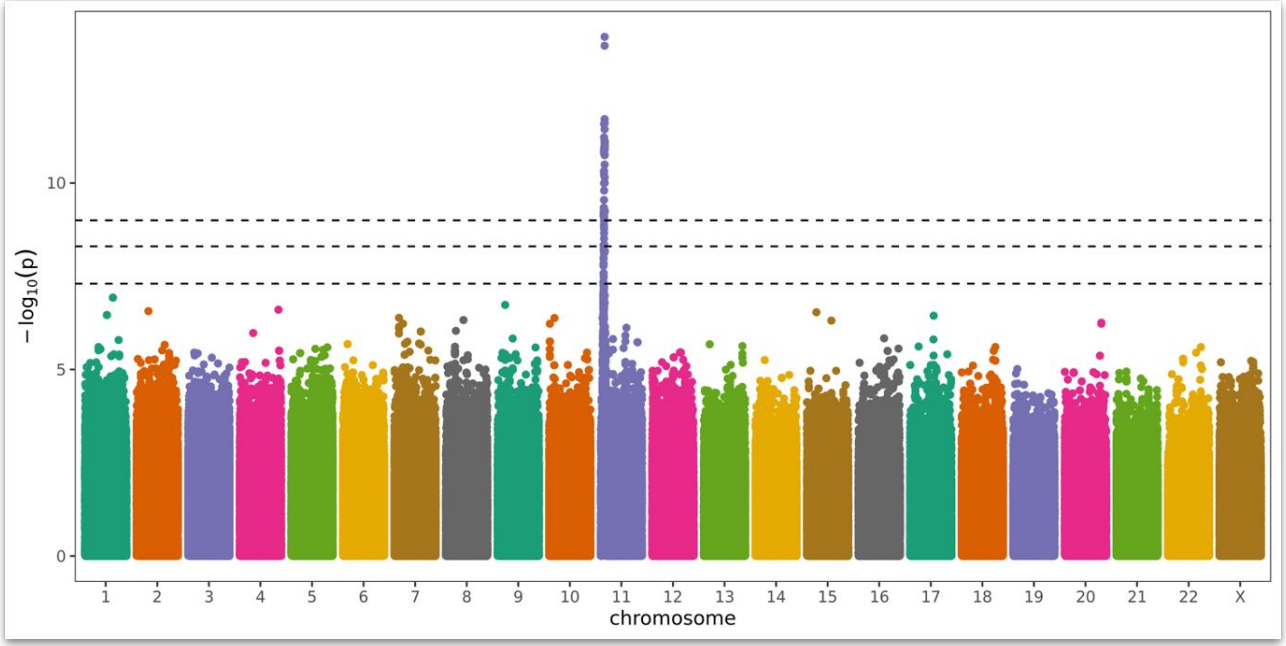
Outputs

- **Locuszoom plots**
  - _1_20210217Chr11LocusZoom_var24148506_ld_TOPMED_chr11_49...
  - _1_20210217Chr11LocusZoom_var24149360_ld_TOPMED_chr11_49...
  - 20210217Chr11LocusZoom_var24149363_ld_TOPMED_chr11_4991...
  - 20210217Chr11LocusZoom_var24149370_ld_TOPMED_chr11_4992...
  - 20210217Chr11LocusZoom_var24149513_ld_TOPMED_chr11_5000...

...and 75 more items

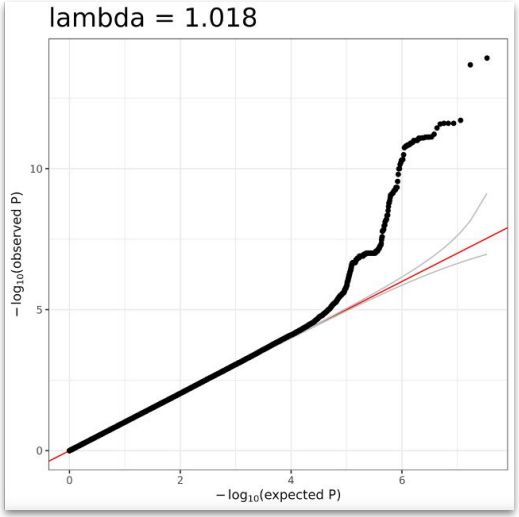


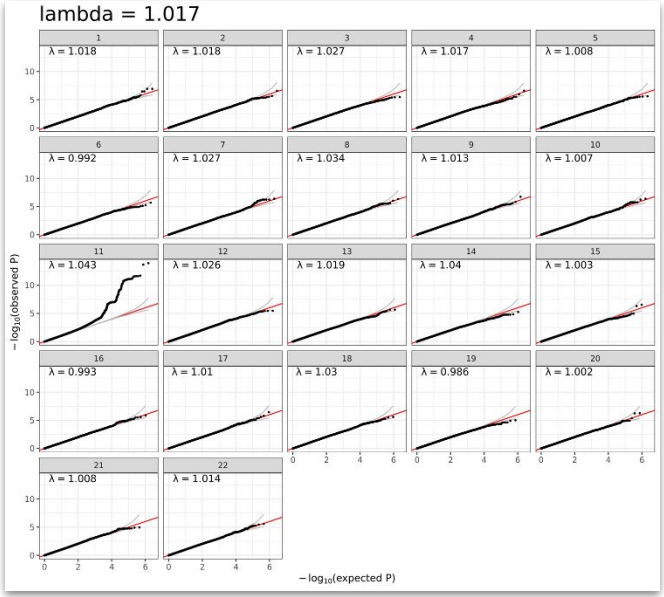
GWAS: Manhattan plot of Ferritin



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

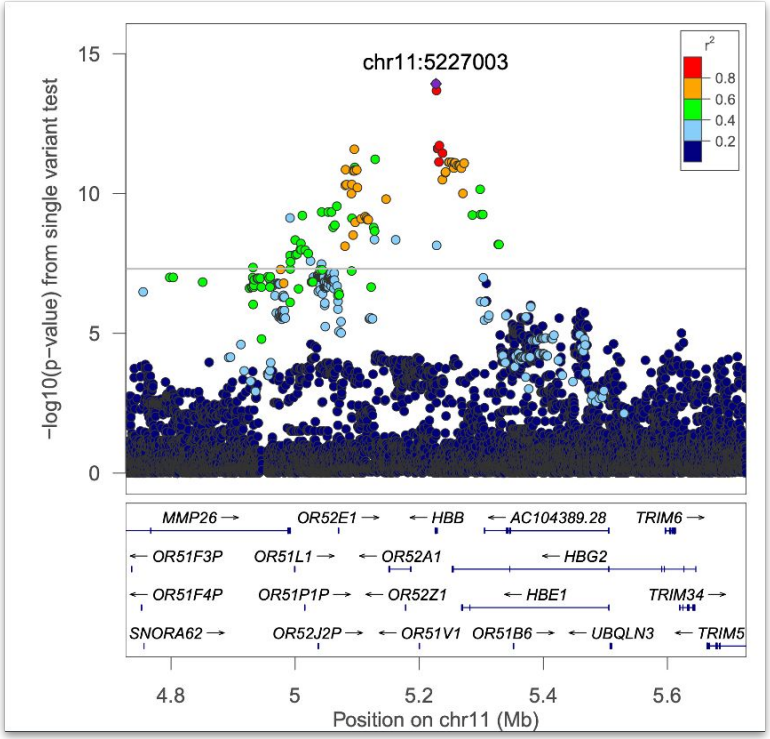




Generated by GENESIS pipeline.

The black dots represented the distribution of observed ordered –log10 (P values) against the theoretical model distribution of expected ordered –log10 P values. The grey line represents the theoretical model distribution of expected –log10 P values under the null distribution.

Regional Association Plot



Variants are plotted with their –log10 (p values) 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

NIH 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)

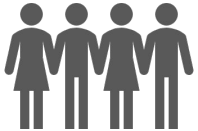
- Coming soon:
  - Additional BioLINCC Teaching and Clinical Trials Datasets
  - Additional studies curated by the Cure Sickle Cell Initiative (clinical trials and cohorts)
  - Additional TOPMed data (rolling basis)
  - COVID-19 data (PETALNet Trials, MIS-C, C3PO, ACTIVE4a, etc.)
  - Pediatric Cardiac Genomics Consortium (PCGC) data

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

Data available in BioData Catalyst

3.42 Petabytes of data

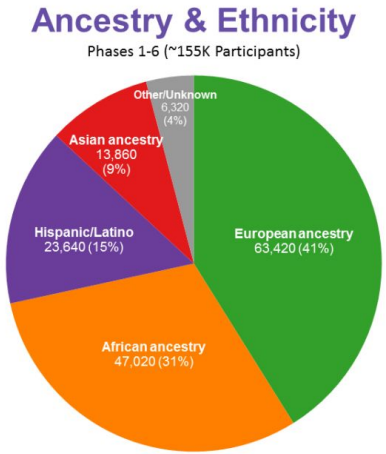
280,000+ Participants

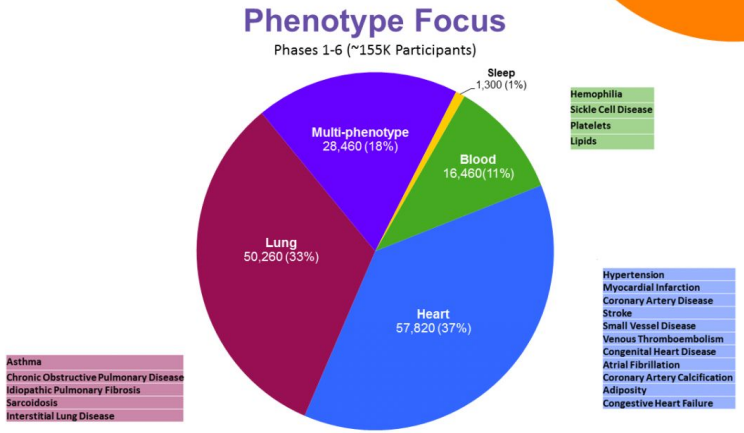


490,000+ Data files

150,000+ Whole genomes

TOPMed Data





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

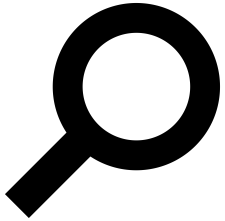
Check Access to Data

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

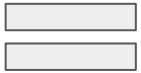
Empowering researchers to access data

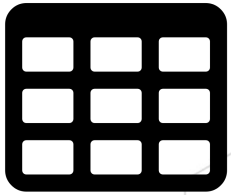
BioData Catalyst Powered by PIC-SURE facilitates approachable research for all skill levels.



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

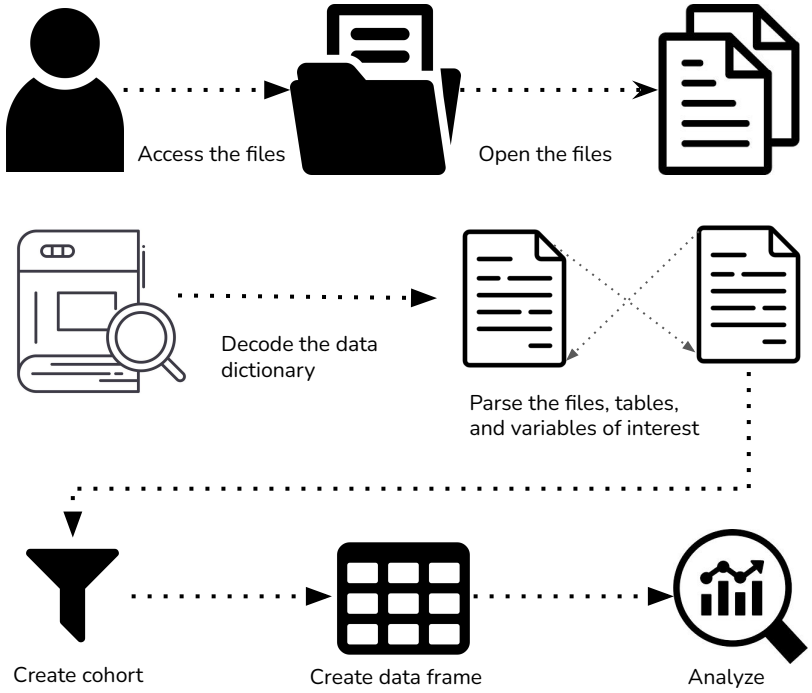
Patient
Information
Commons
-
Standard
Unification of
Research
Elements

- 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

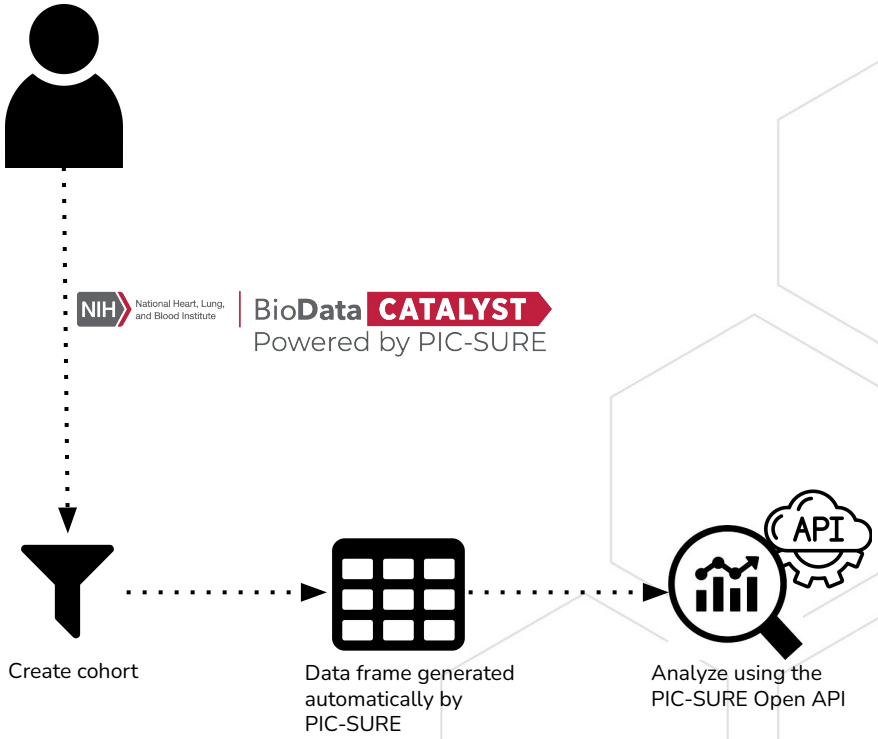
https://picsure.biodatacatalyst.nhlbi.nih.gov/

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

Traditional Workflow

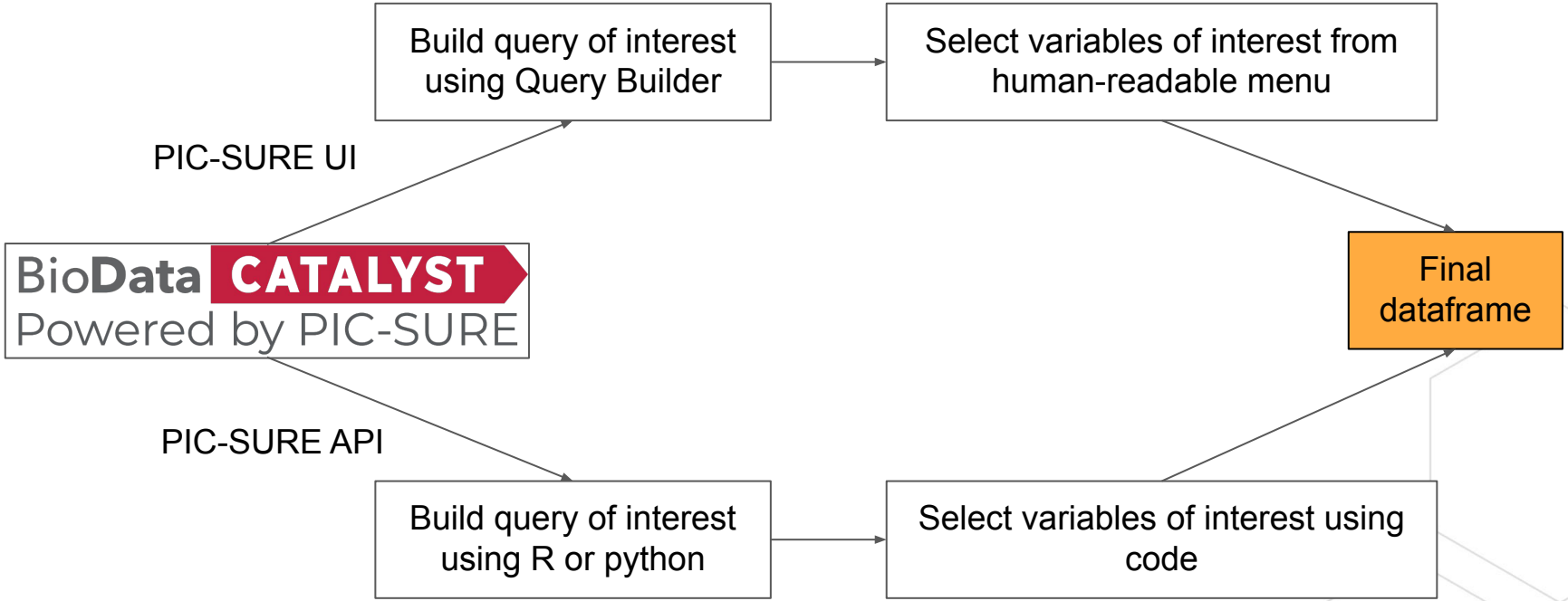


PIC-SURE Workflow



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

PIC-SURE workflow - 2 options



Open vs Authorized Access

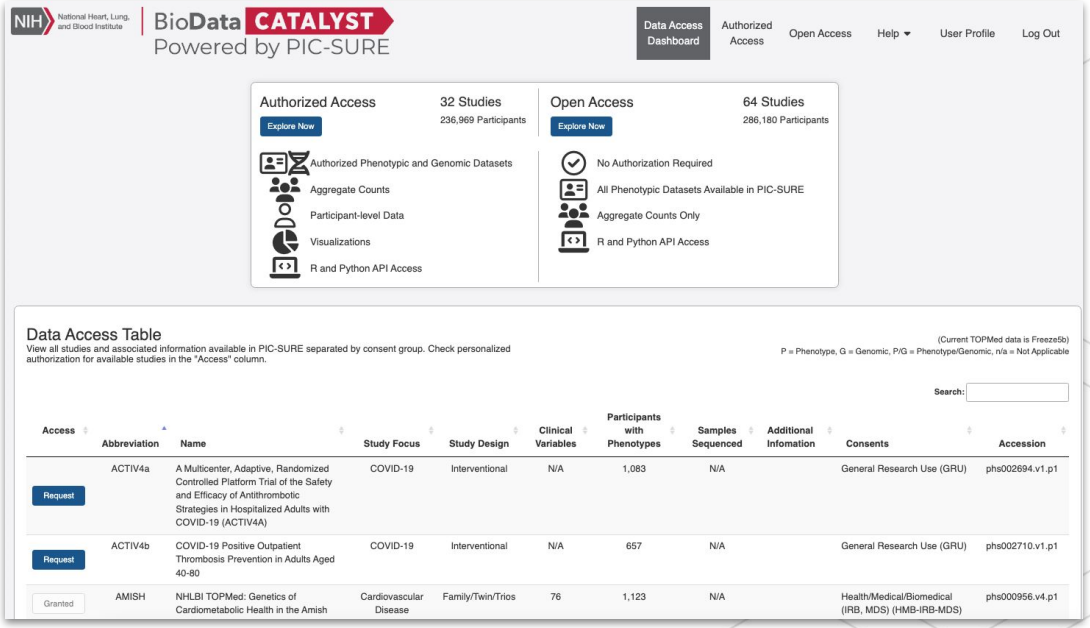
	PIC-SURE Open Access	PIC-SURE Authorized Access
Overview	Allows any user with eRA Commons ID to search any clinical variable in PIC-SURE	Allows users with dbGaP authorization to access data and export to analysis platforms
Access authorization	No approval required, just eRA Commons ID	dbGaP authorization required
Data types	Destigmatized clinical variables	All phenotypic and genomic data
Results	Aggregate counts based on queries	Participant-level data
Use case	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

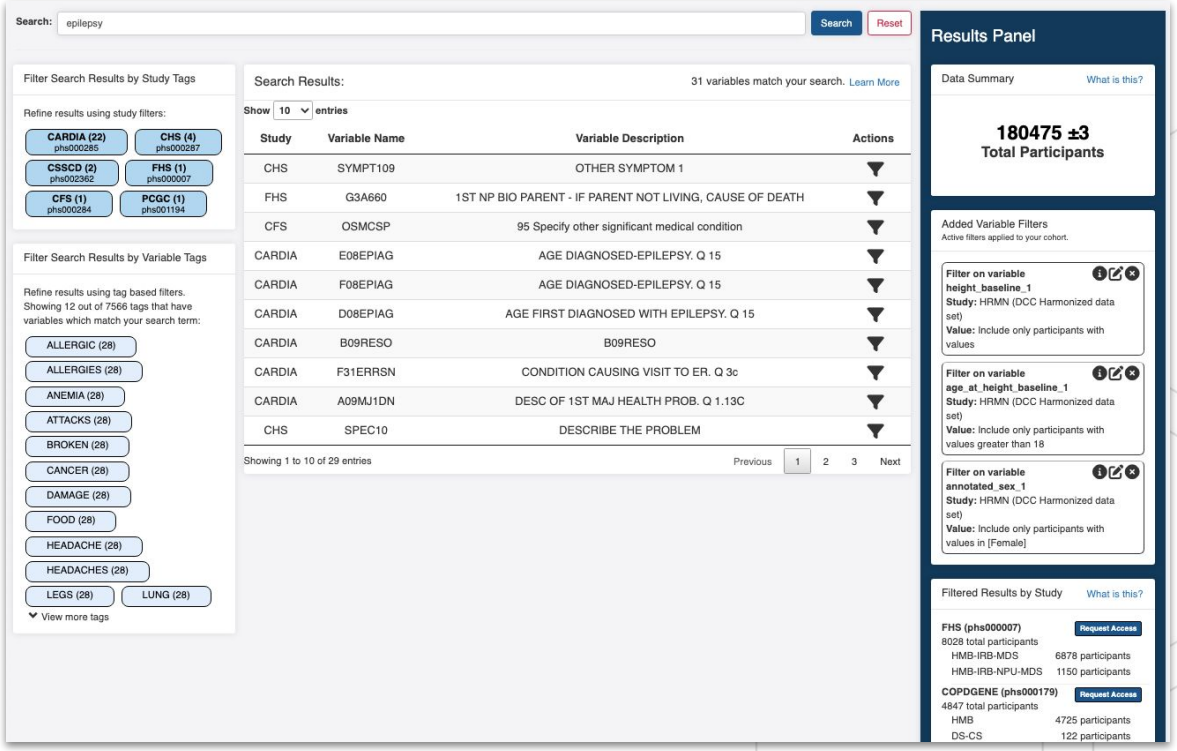


PIC-SURE Open Access

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

Demo

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

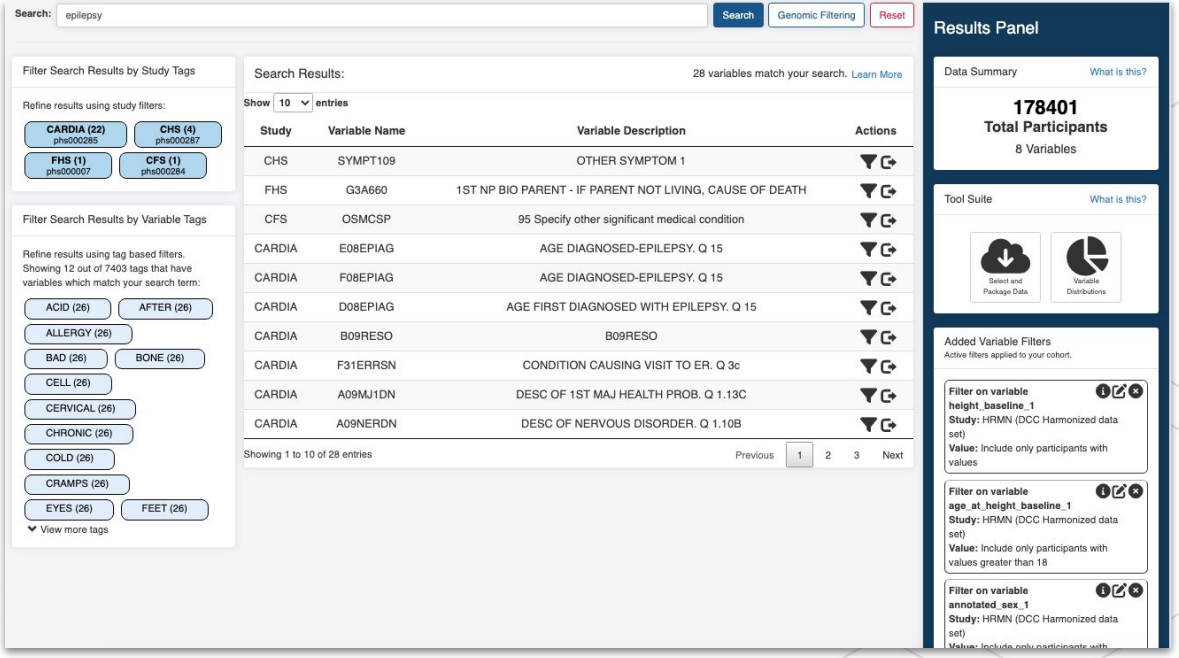


PIC-SURE Authorized Access

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

Demo

https://picsure.biodatacatalyst.nhlbi.nih.gov/picsureui/queryBuilder#



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 esposes 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`.

```
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.

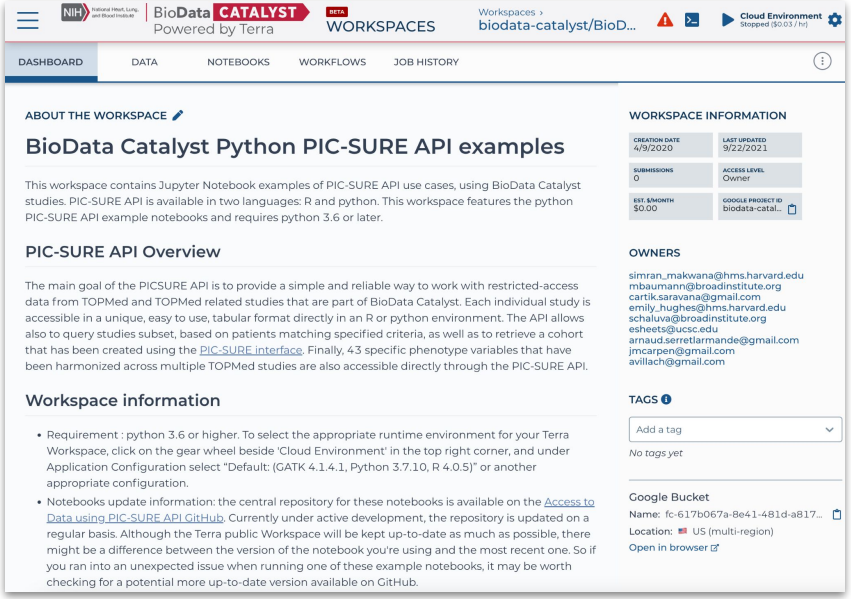
```
i = 0
categories = []
while len(categories) == 0 or len(categories) > 8:
    categorical_var_info = my_variables_df[my_variables_df.is_categorical == True]
```

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

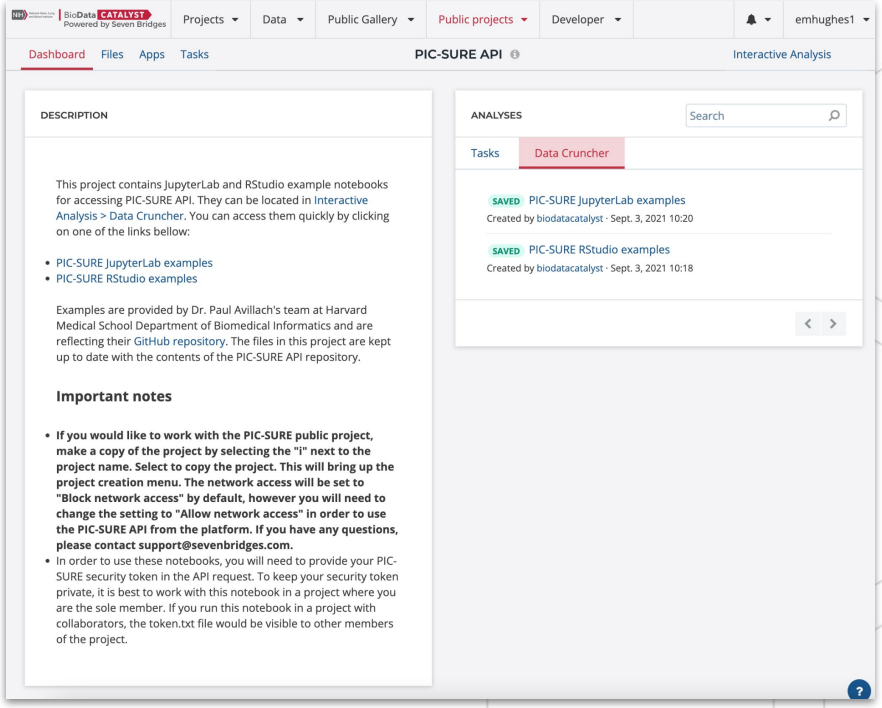
Workshop Forum: https://bit.ly/BDC-Howard-Workshop

PIC-SURE API

BioData Catalyst Powered by Terra



BioData Catalyst Powered by Seven Bridges



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

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

NIH National Heart, Lung, and Blood Institute
BioData CATALYST

Bring-Your-Own Data

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

Web resource: Bring Your Own Data

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

Seven Bridges workspace environment

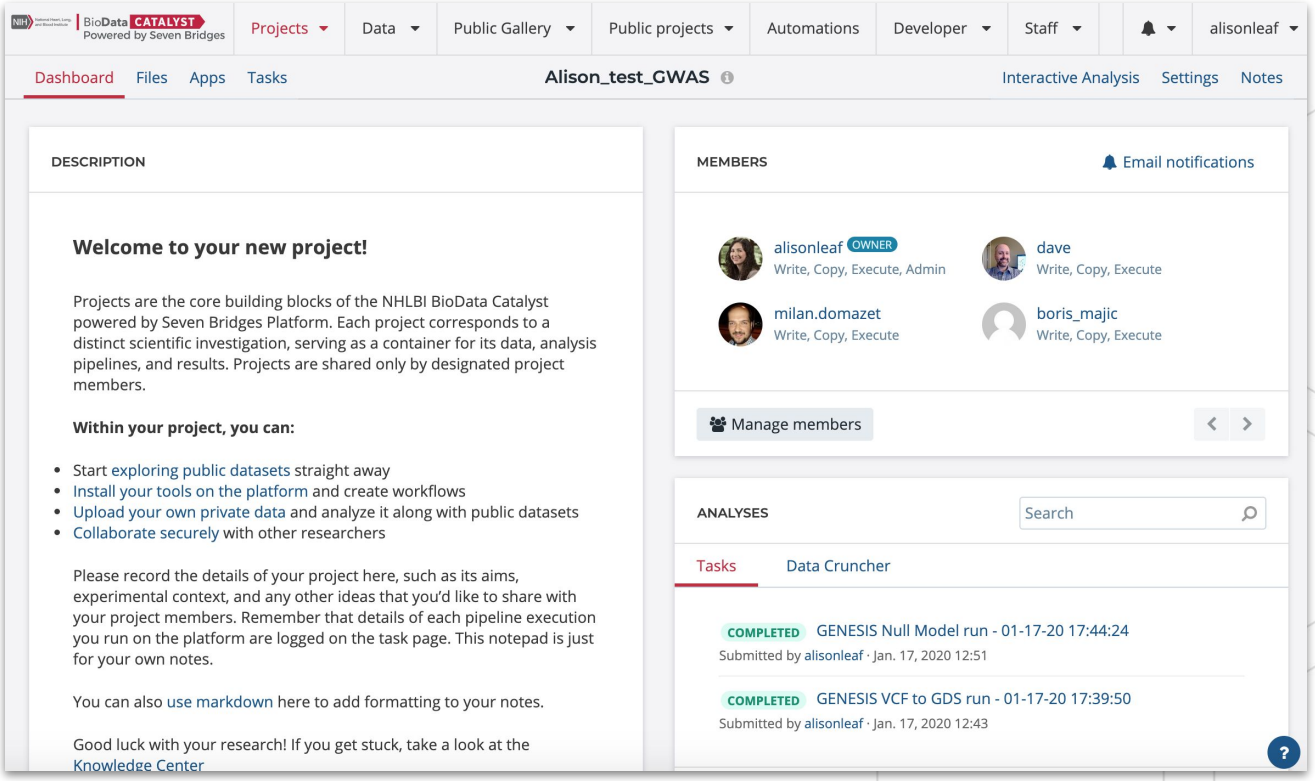
Private, secure workspaces with the option to collaborate

Set up analyses with visual user interface or API

Jupyterlab Notebooks and RStudio

Compute on AWS or Google

Hundreds of hosted CWL pipelines

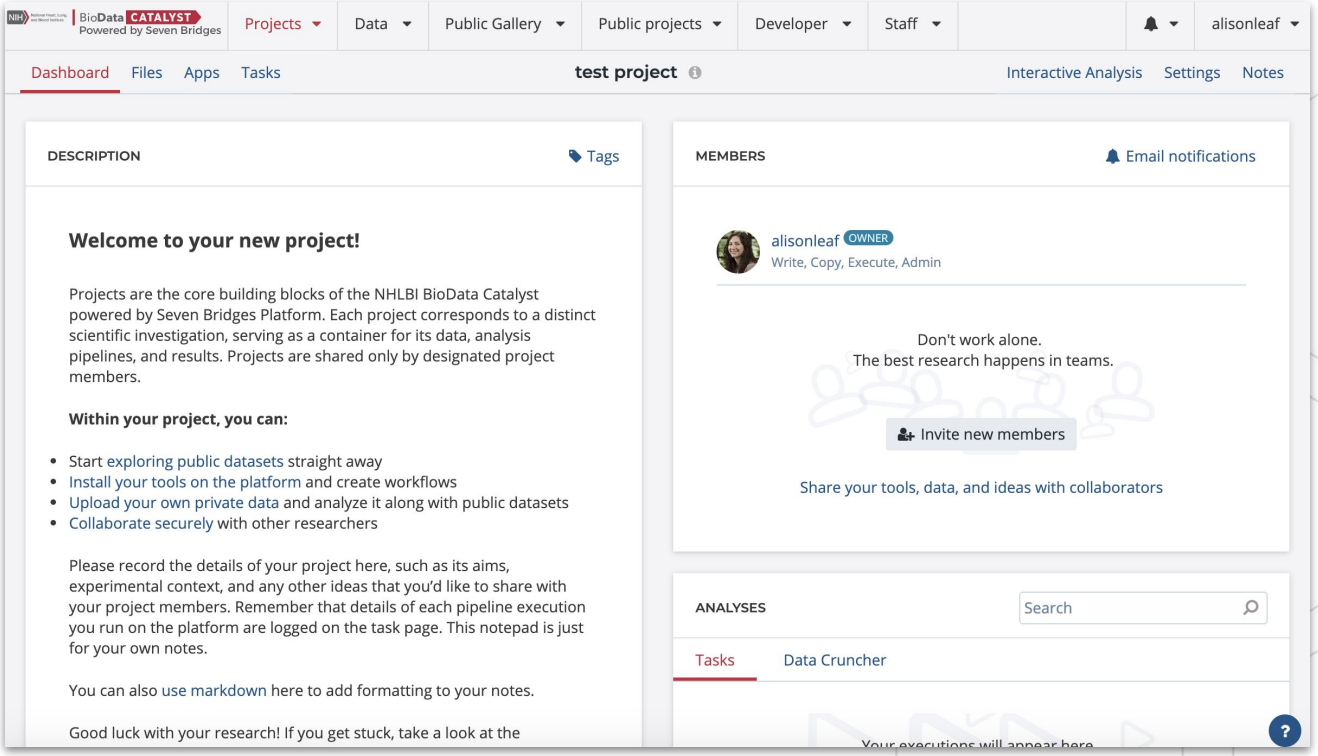


NIH National Heart, Lung, and Blood Institute | BioData CATALYST

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.



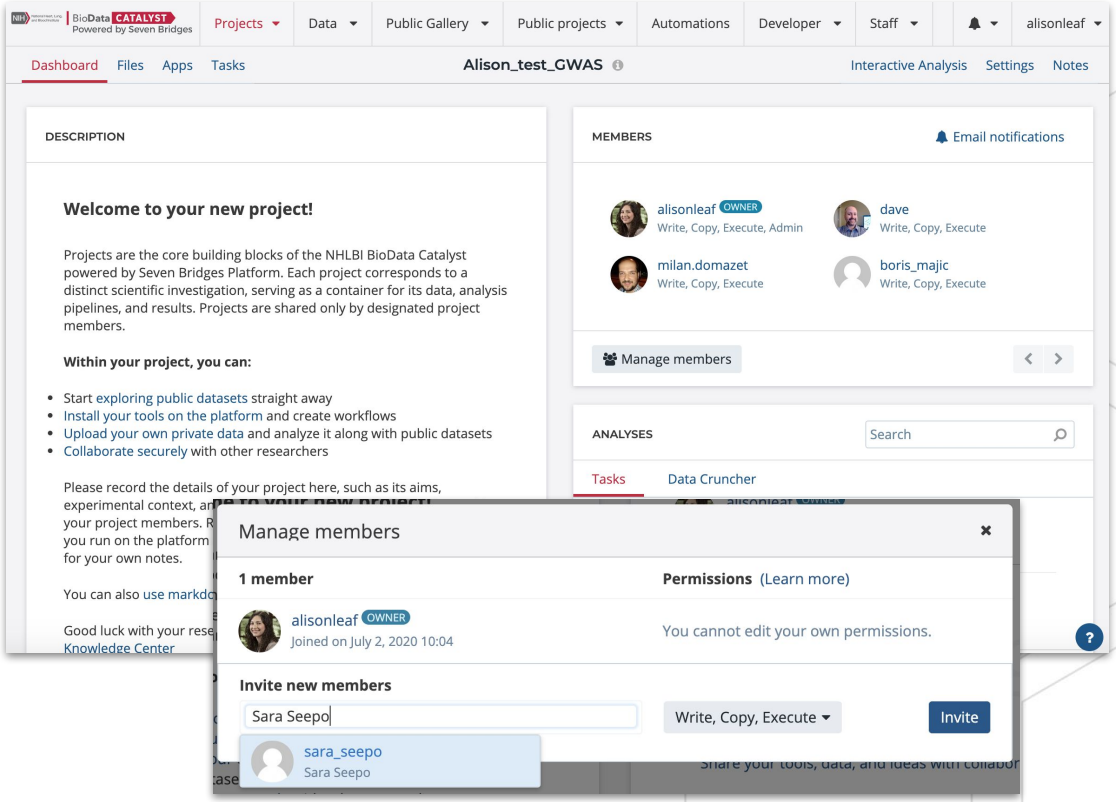
NIH National Heart, Lung, and Blood Institute | BioData CATALYST

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

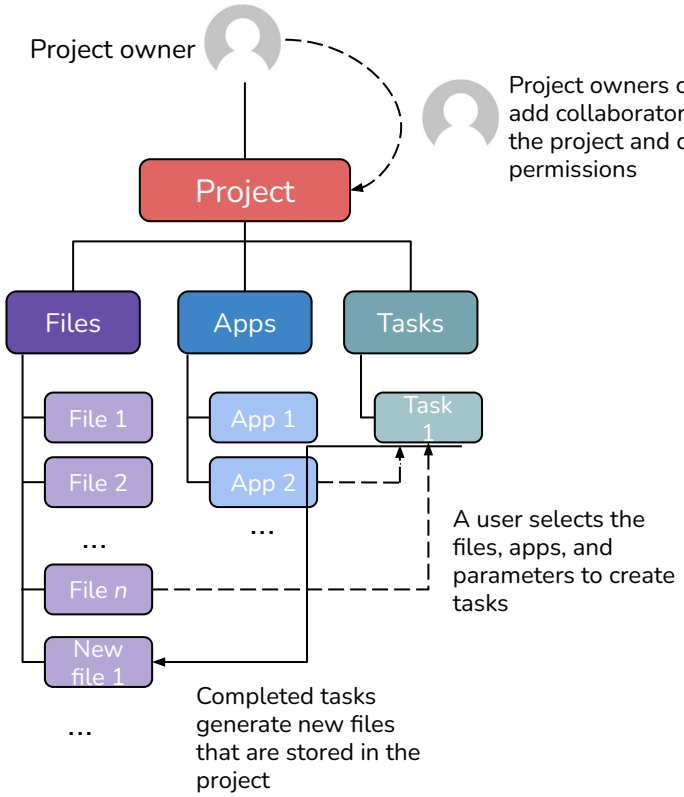
Set granular permissions to limit what project members can see/do



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

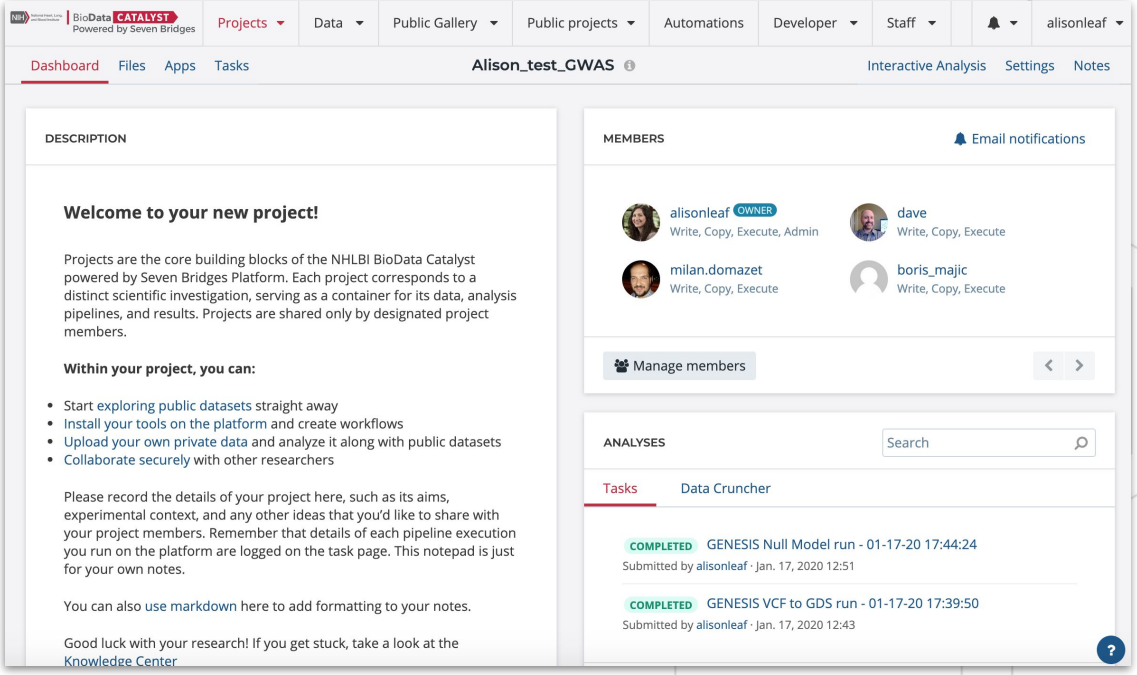
Workshop Forum: https://bit.ly/BDC-Howard-Workshop

Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions



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

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/do
wnloads/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

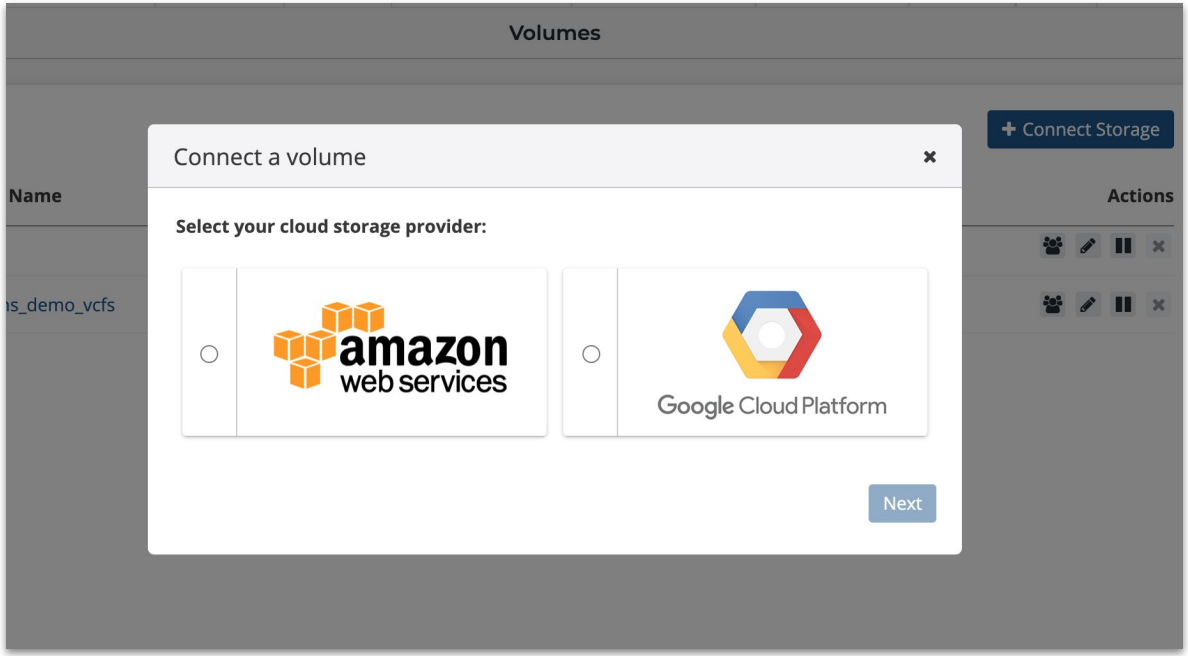
Paste the link of the file(s) you want to import

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

or Browse file on your computer containing the links

Import

Connect private cloud storage directly to platform

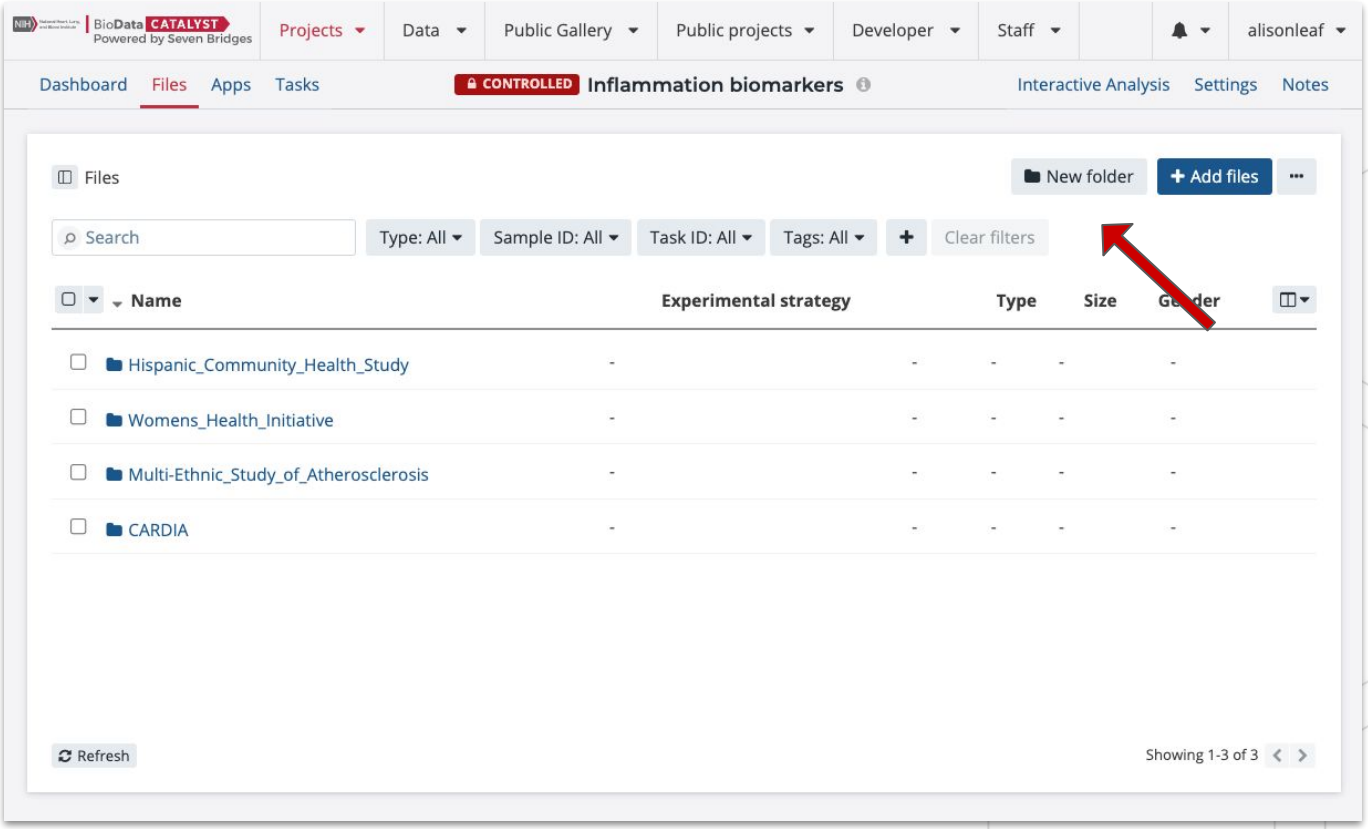


Users retain full control over cloud storage access, management, and integrations.

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

Organize and manage files within projects

Nested folder structure for organizing files



Questions?

Next up: Tools, Workflows, and Interactive Analysis

Overview of the BioData Catalyst Ecosystem

Kat Thayer

NIH 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.

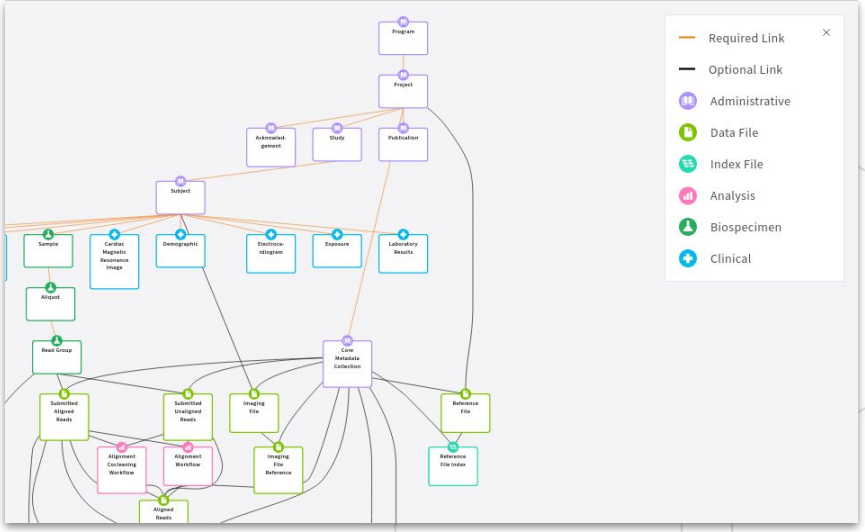
- creates pointers to data files and links them to metadata (***file information***) .

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

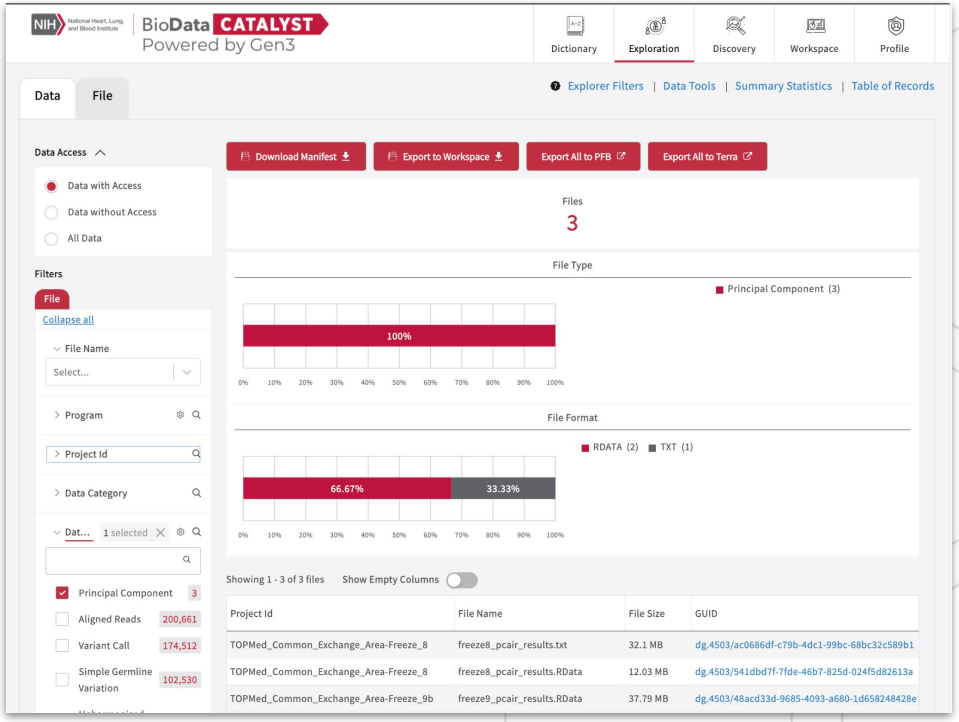


Workshop Forum: https://bit.ly/BDC-Howard-Workshop

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



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

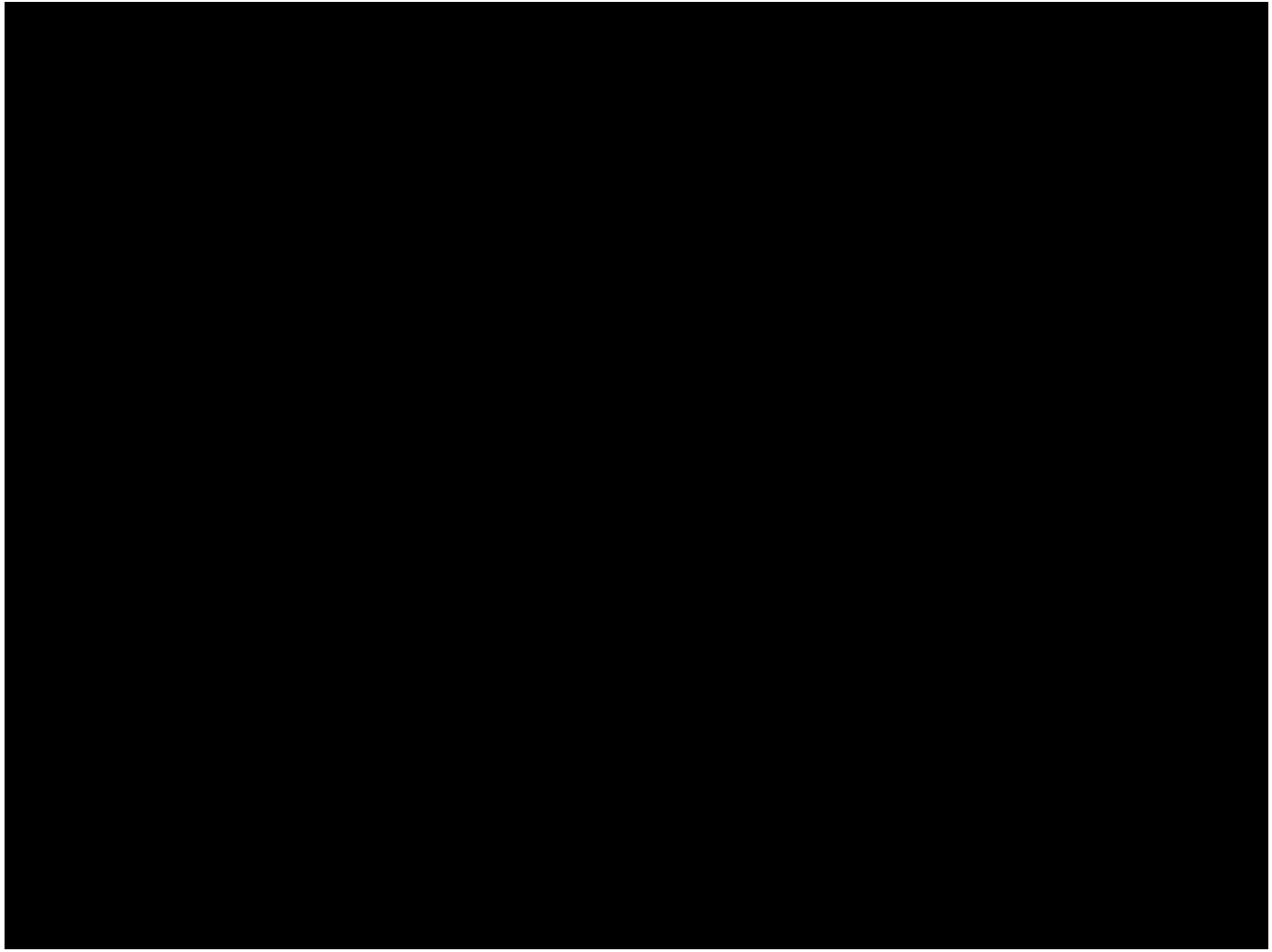
About *BioData Catalyst Powered by Terra*

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 (*coming*)
Applications	• Preloaded applications and options to bring-your-own through a user-friendly interface. • Galaxy, IGV, Seqr
Interactive Analysis Features	• Highly customizable machines with persistent disks set up to save your work • Bioconductor, Hail, GATK and other popular bioinformatics tools preloaded. • "Best practices" workspaces from the tool developers



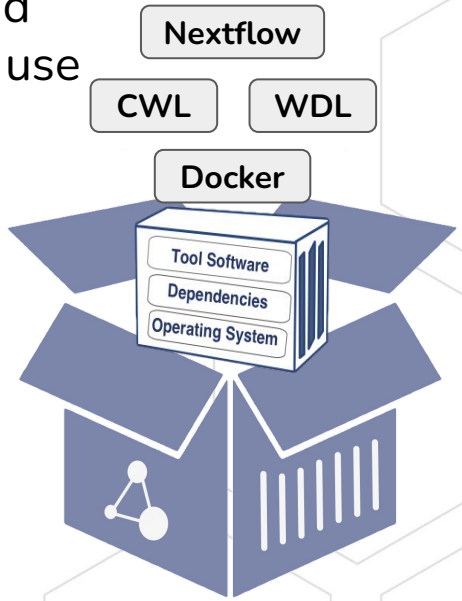
Introduction to Dockstore

“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

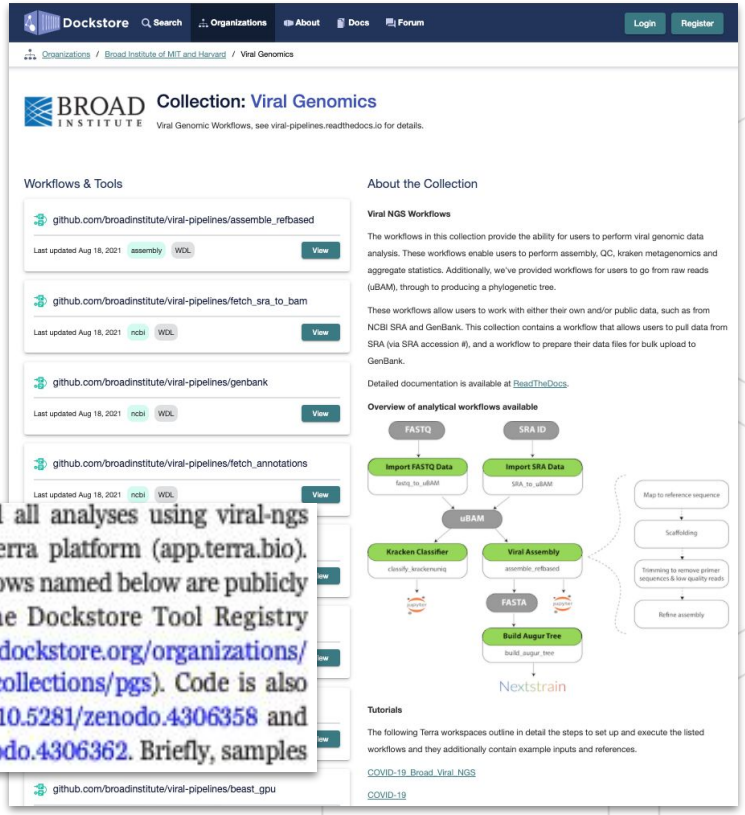
- Increases reproducibility of computational analysis using combination of containers and workflow languages
- Increases the transparency of analysis methods
- Allows others to verify results and apply existing methods into their own research



dockstore.org

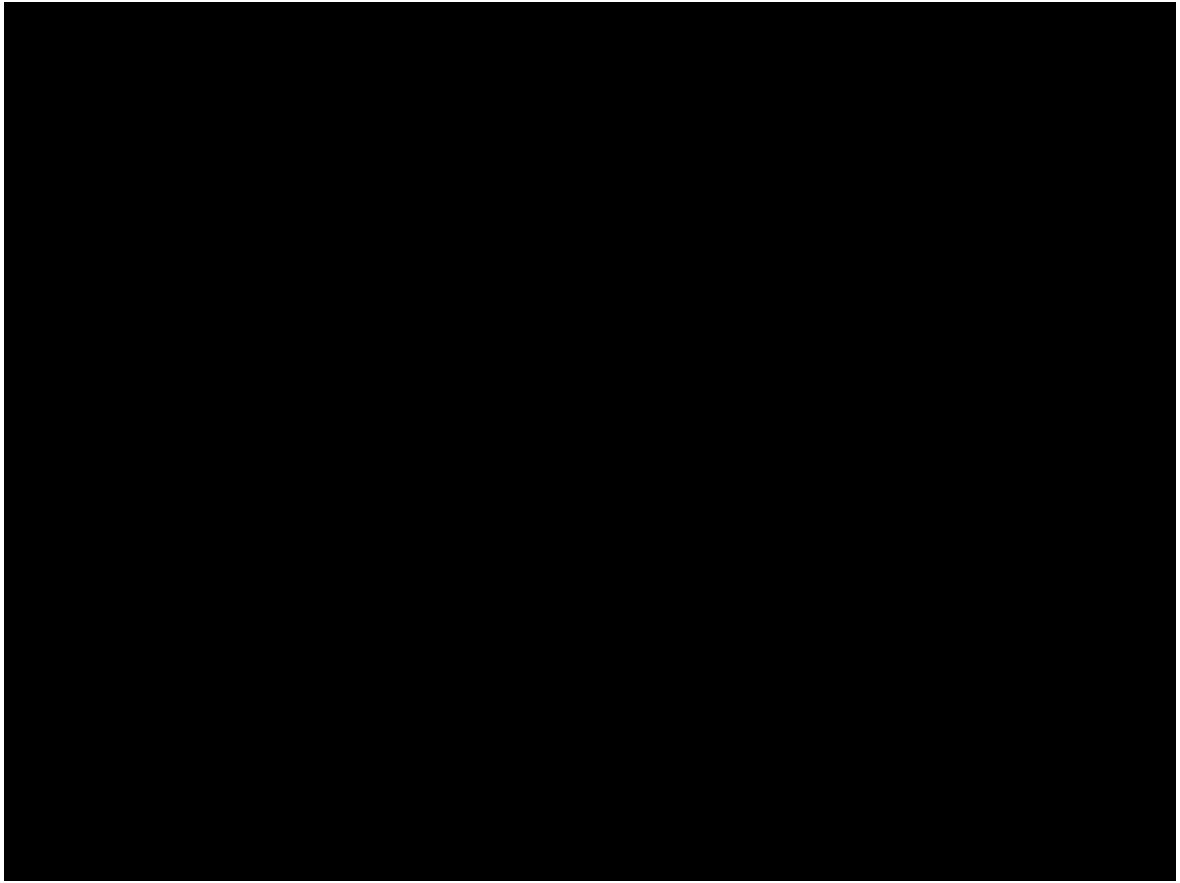
Publish your workflows on Dockstore!

- 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.



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 and doi:10.5281/zenodo.4306362. Briefly, samples

Lemieux et al., Science 371, eabe3261 (2021)



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.

NIH 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.

NIH 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

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

Agenda

Day One: Thursday, November 17th

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

Day Two: Friday, November 18th

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

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

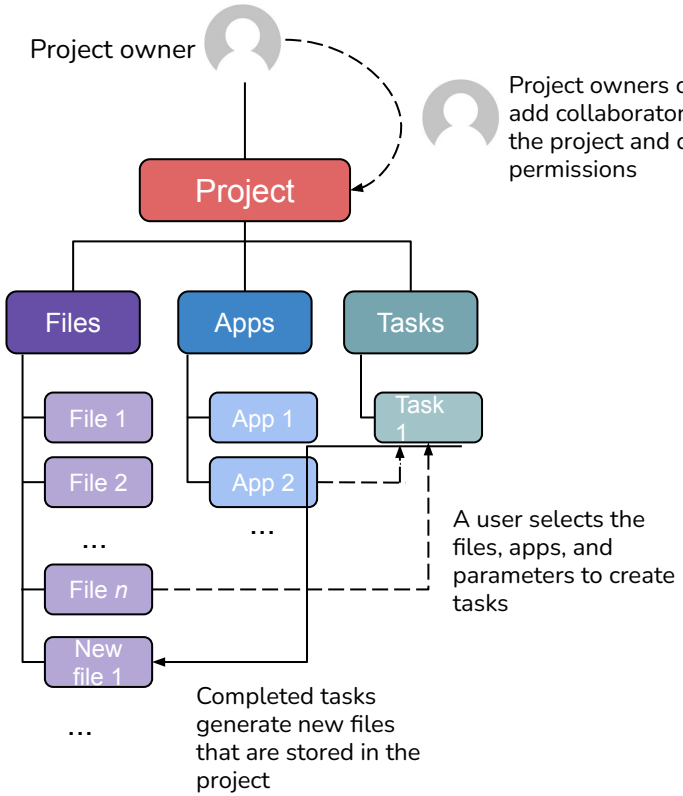
Tools, Workflows, and Interactive Analysis

Dave Roberson, Community Engagement Specialist at Seven Bridges

NIH National Heart, Lung, and Blood Institute
BioData CATALYST

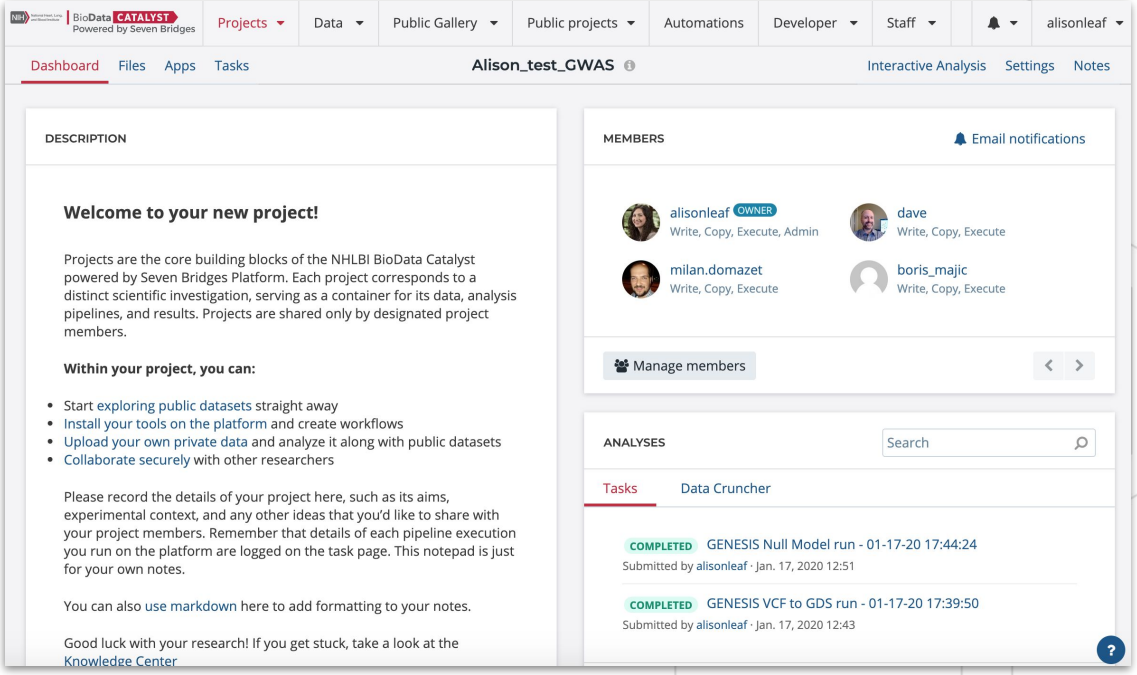
Workshop Forum: https://bit.ly/BDC-Howard-Workshop

Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*

Easily manage collaborators and permissions

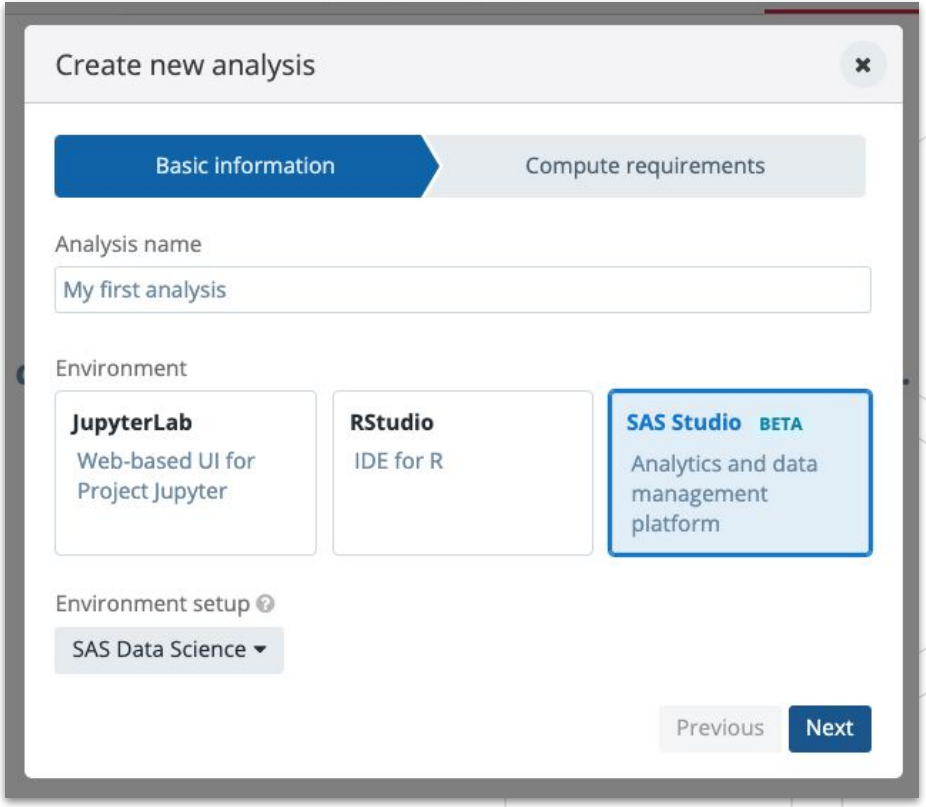


NIH National Heart, Lung, and Blood Institute | BioData CATALYST

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.



User friendly workflow editor enables reproducibility by default

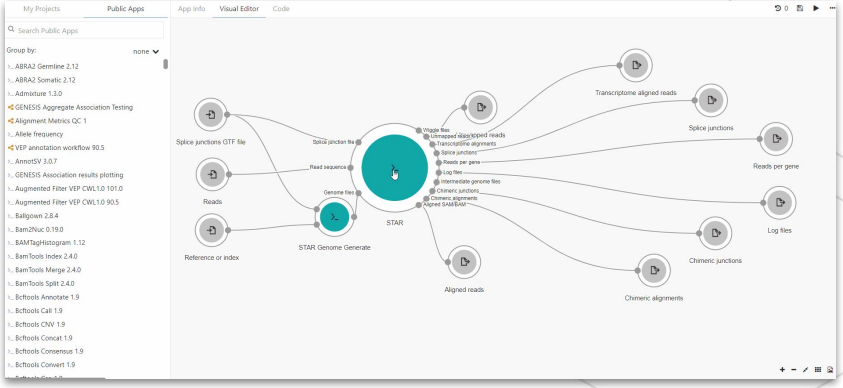
Common Workflow Language enables **portability**, **reproducibility**, and **scalability**

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

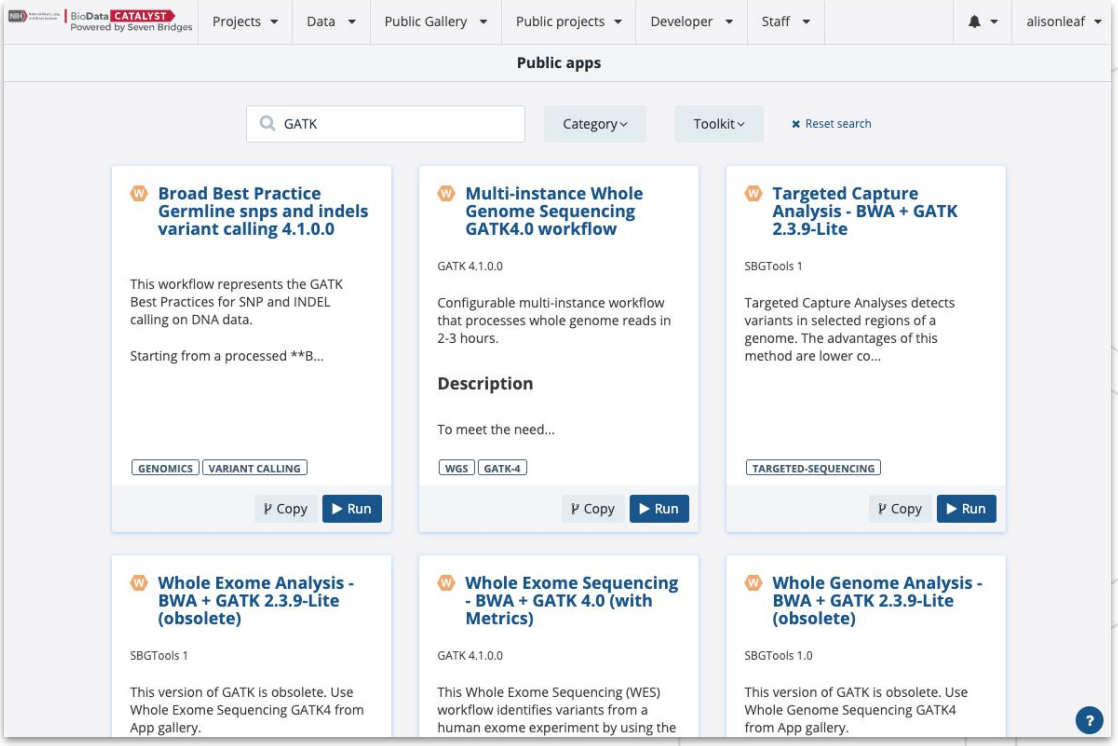


Workshop Forum: https://bit.ly/BDC-Howard-Workshop

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

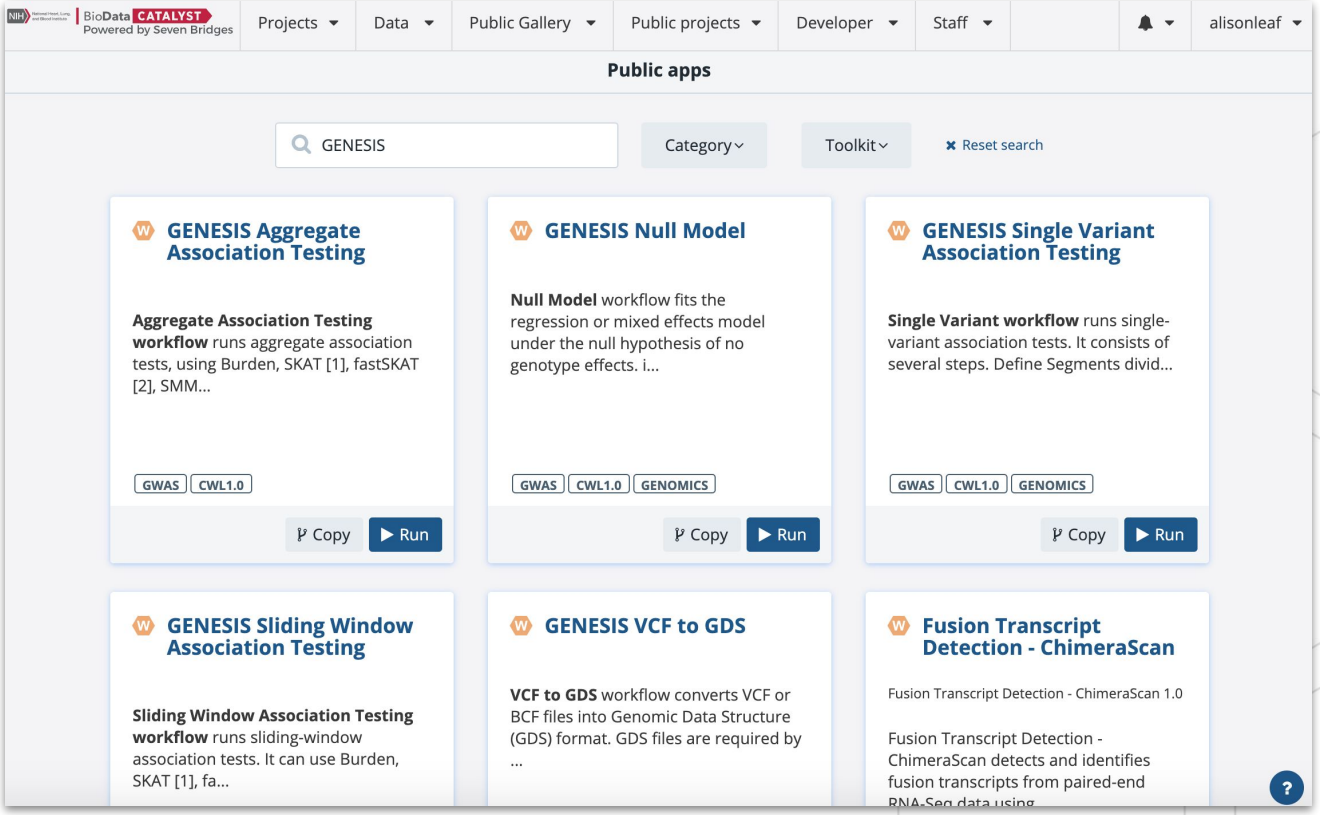


NIH National Heart, Lung, and Blood Institute | BioData CATALYST

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

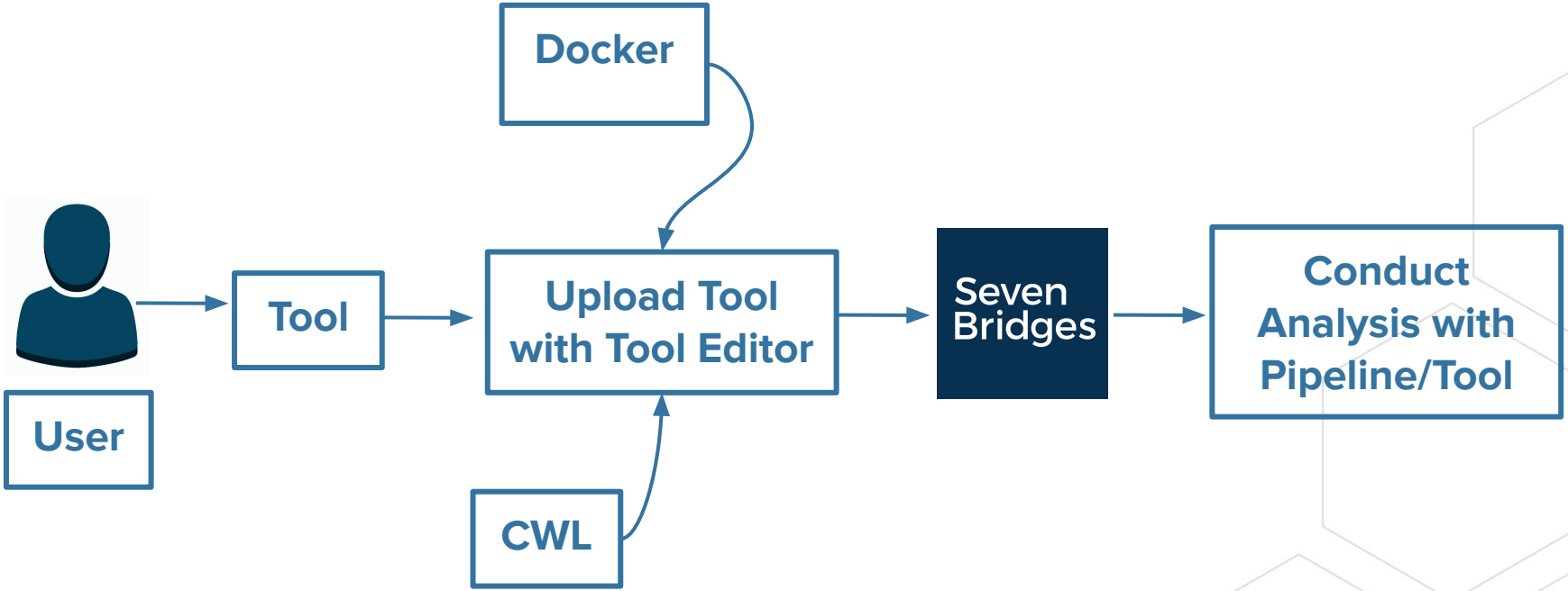
Run association pipelines out of the box

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



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

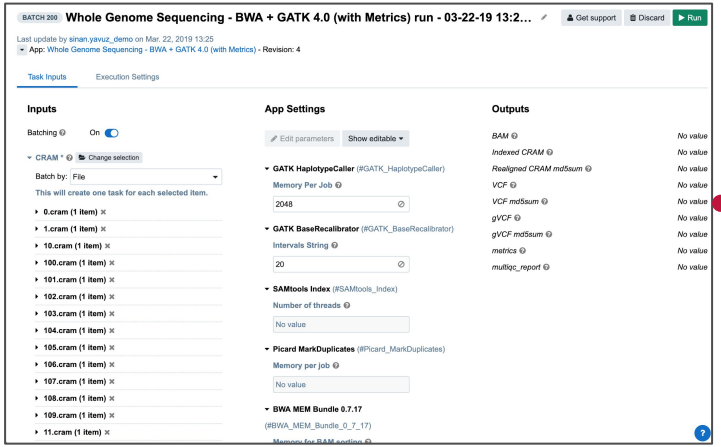
Bringing custom tools to the platform

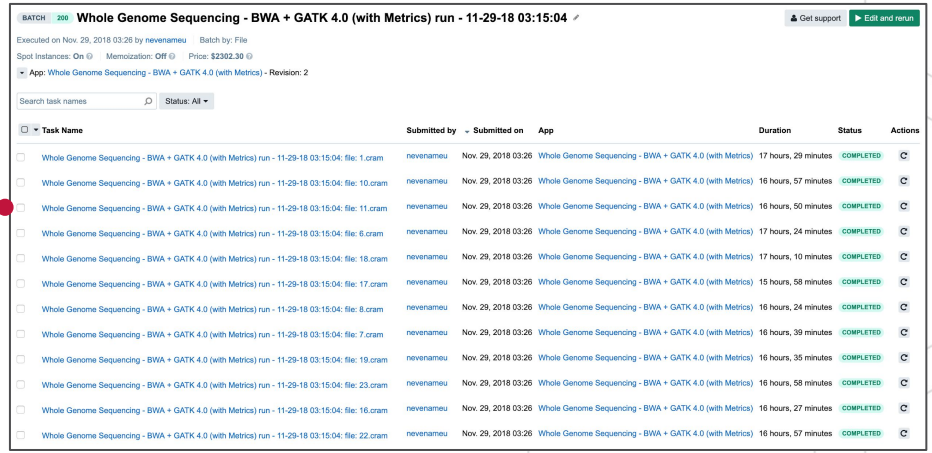


Scale to 100's and 1000's of tasks in parallel using batching

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.





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

Detailed documentation and tutorials

Comprehensive tips for reliable and efficient analysis set-up

BIODATA CATALYST POWERED BY SEVEN BRIDGES

OBJECTIVE

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

HELPFUL TERMS TO KNOW

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

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

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

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

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

Troubleshooting Failed Tasks

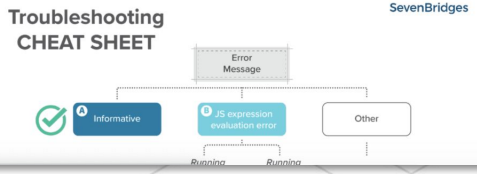
BIODATA CATALYST POWERED BY SEVEN BRIDGES

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

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

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

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



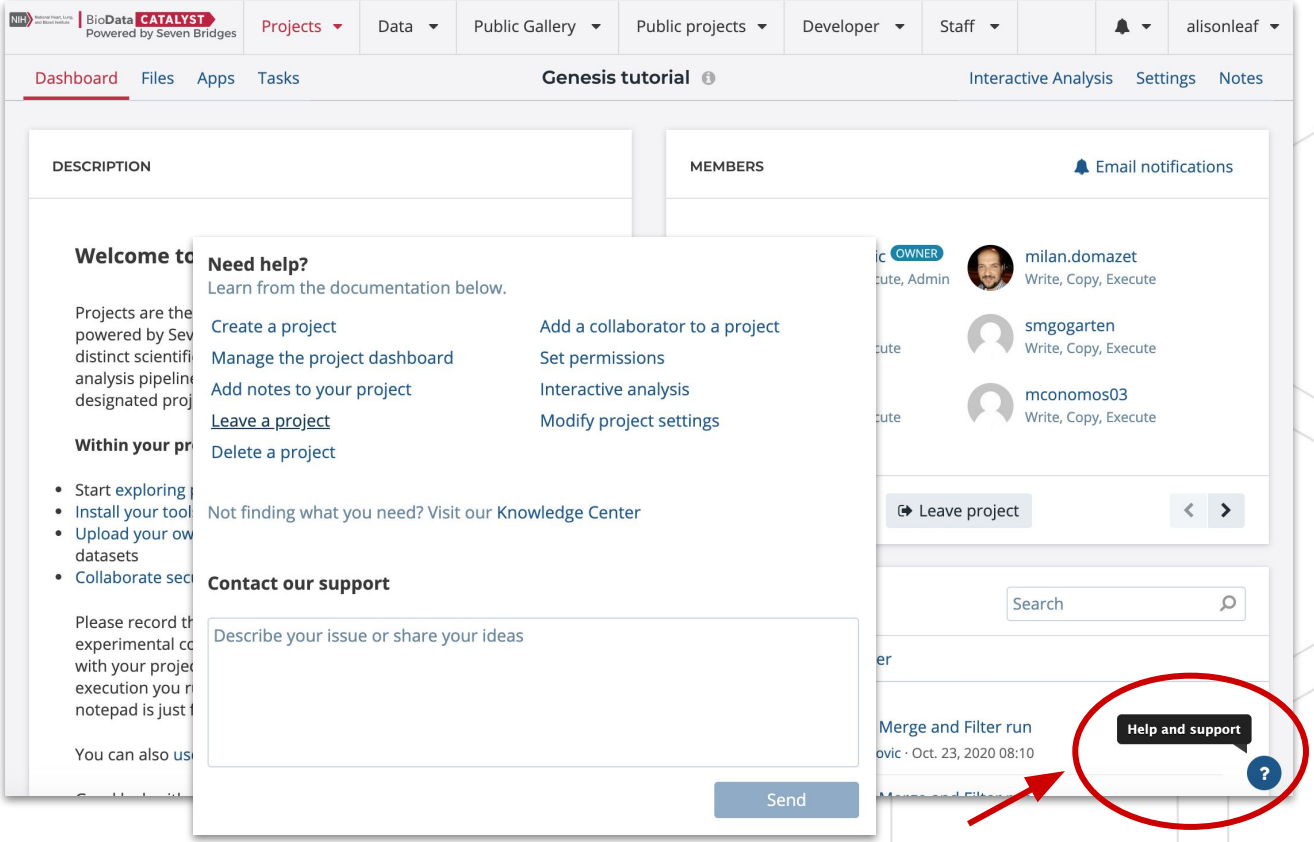
Visit the Knowledge Center

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

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

Getting Help - Contacting Support from the platform

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



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

Questions?

Next up: Understanding, Estimating, and Managing Cloud Costs

Understanding, Estimating, and Managing Cloud Costs

Dave Roberson, Community Engagement Specialist at Seven Bridges

NIH 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

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

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

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

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.

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

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, cost
found in the sections below. In the
benchmarking results and some ti
Lastly, we included a Methods sect
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$*

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

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** | ...n: **11 minutes**

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

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

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 — Show non-default

- **GENESIS Null Model** (#null_model)
  - **Covariates**
    - sex
    - age
    - PC1
    - PC2
    - PC3
    - PC4
  - Family: gaussian
  - Outcome: height
  - Output prefix (Output prefix): demo_height
  - Two stage model: FALSE

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**

BioData CATALYST Powered by Seven Bridges | Projects | Data | Public Gallery | Public projects | Developer | Staff | alisonleaf

Current Usage
Spending Details
Group Members

Billing Group settings: BDC Fellow - Einat Granot-Hershkov

Alison Leaf
Account settings
Payments
Email notifications
Sign out

Info

Organization

Creator	alisonleaf
Primary contact	Seven Bridges
Address	One Broadway, 14th Fl. , Massachusetts, United States
Payment Method	CC - Last 4 Digits: **** **** **** Expiration Date: Change Payment Method

Current usage $ 309.42

Analysis usage

Analysis charges	$ 303.64
Tasks	$ 285.24
Data Cruncher analyses	$ 18.41
Additional charges	$ 0.00

Storage usage

Storage charges	$ 5.78
Active	$ 5.45
Downloaded	$ 0.33
Storage deduction	$ 0.00
Credits	$ 0.00

Instance limits
Total number of instances that can be run in parallel
Current usage: 0 of 60

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

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

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

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

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

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

Questions?

Take a moment to request pilot funds

Next up: Running a GWAS on BioData Catalyst

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

Running a GWAS on BioData Catalyst

Dave Roberson, Seven Bridges

NIH 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

Analysis Steps:

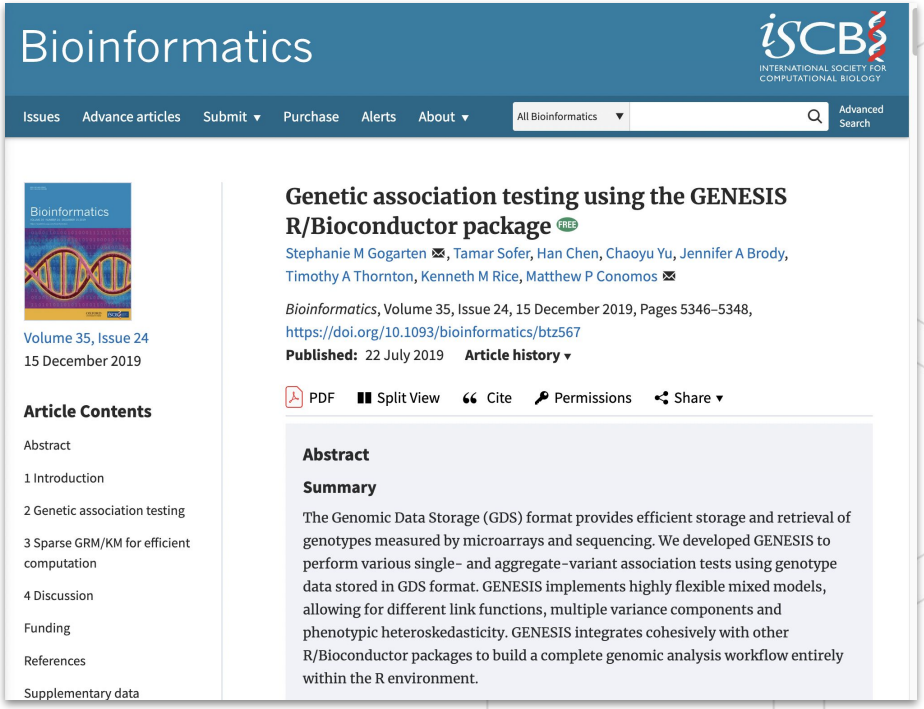
- Conversion to GDS
- Relatedness and Population structure
- Genetic Relationship Matrix
- Association testing
  - Null model
  - Single Variant
  - Rare variant

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



Bioinformatics

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Issues | Advance articles | Submit | Purchase | Alerts | About | All Bioinformatics | Advanced Search

Volume 35, Issue 24
15 December 2019

Article Contents

Abstract
1 Introduction
2 Genetic association testing
3 Sparse GRM/KM for efficient computation
4 Discussion
Funding
References
Supplementary data

Genetic association testing using the GENESIS R/Bioconductor package FREE

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

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

Published: 22 July 2019 **Article history**

PDF | Split View | Cite | Permissions | Share

Abstract

Summary

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

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.

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

Questions?

Feedback on Workshop

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

Thank you for joining us

Join the Community

Interact with the forum

Subscribe to our YouTube channel

Register for Nov 30 Community Hours

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