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

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

BDC for NDSA Workshop

Day 2: July 14 | 11 AM ET

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

Statement of Conduct

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

Web Resource: Statement of Conduct

Recap of Day 1

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

Agenda: Day 2

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

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

Meet Your Instructors

Emily Hughes
BDC Powered by PIC-SURE
Harvard Medical School

Kat Thayer
BDC Powered by Terra
Broad Institute

David Roberson
BDC Powered by Seven Bridges
Seven Bridges/Velsera

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

TAs and Live Support

Cera Fisher
BDC Powered by Seven Bridges
Seven Bridges/Velsera

Aarthi Krishnan
BDC Powered by Seven Bridges
Seven Bridges/Velsera

Michael Corace
BDC Powered by PIC-SURE
Harvard Medical School

Amber Voght
User Engagement Specialist,
BDC Coordinating Center

Kaleena Narwani
User Engagement Specialist,
BDC Coordinating Center

Check-In

Icebreaker

- Did you practice what we covered yesterday?
  - How did it go?
  - What are you interested in or excited about?
  - What are you still unsure about?
- Questions?

Reproducible Research

Curricula of:

Reproducible Research

In this section, you will:

- Gain understanding about the FAIR Principles are and how they can be applied to research
- Discuss the importance of reproducible research
- Learn about some real use-cases of reproducible research on BDC

What are the FAIR principles?

What are the FAIR Principles?

Guiding principles for data management and stewardship - that can be applied to data analysis as well

Findable
Accessible
Interoperable
Reusable

For more information about the FAIR Principles, you can visit the "Go FAIR" website

Findable

Data

- Data and metadata should be easily discoverable by both humans and computers
- Metadata should accurately describe the data
- Data and metadata should be searchable

Analyses and workflows

- Analyses and workflows should be discoverable and searchable by humans
- Documentation should accurately describe the analyses and workflows

Accessible

Data

- Once researchers find data, they should have a clear understanding of how to proceed with accessing the data
- Could include authorization to access the data

Analyses and workflows

- Once researchers have found a workflow of interest, they should have a clear understanding of how to proceed with accessing the code
- Could include citing open source workflows

Interoperable

Data

- Data should be easily integratable with other data
- Data should be available to apply to a variety of workflows

Analyses and workflows

- Analyses and workflows should have the capability to be used on a variety of systems and platforms, where applicable
- Analyses and workflows should be generalized and configurable

Reusable (or Reproducible)

Data

- Other researchers should be able to reuse data
- Metadata should describe the data clearly to allow for reuse
- Other researchers should be able to understand where the data came from, how it was processed, and who has transformed the data (data provenance)

Analyses and workflows

- Analyses and workflows should be described in enough detail for other researchers to reproduce the analysis
- When possible, analyses and workflows should be cited and/or shared

Breakout rooms

In your breakout rooms, discuss the following:

1. What are some methods you apply in your research process to embody FAIR principles?
2. Have you ever had issues trying to recreate an analysis or find/access data?
3. What aspects of BDC embody FAIR principles? How could you apply this in your research on BDC to ensure FAIR principles?

Examples of reproducible science in BDC

BDC & Sickle Cell Disease

Acknowledgements

Sickle Cell Disease reproduction analysis and slides completed by **Andrew St. Martin**

Center for International Blood & Marrow Transplant Research (CIBMTR)

Cure Sickle Cell Initiative

- Sickle cell disease (SCD) is a debilitating hematologic disorder
- Hematopoietic cell transplantation (HCT) is potentially curative for SCD
- CIBMTR collects patient outcome data for SCD transplant recipients
- Several SCD studies available on BDC per the Cure Sickle Cell Initiative

Project Objectives

1. Reproduce the results from a recent NHLBI publication

Effect of donor type and conditioning regimen intensity on allogeneic transplantation outcomes in patients with sickle cell disease: a retrospective multicentre, cohort study

Mary Eapen Prof, Ruta Brazauskas PhD, Mark C Walters Prof, Françoise Bernaudin Prof, Khalid Bo-Subait MPH, Courtney D Fitzhugh MD, Jane S Hankins MD, Julie Kanter MD, Joerg J Meerpohl Prof, Javier Bolaños-Meade MD, Julie A Panepinto Prof, Damiano Rondelli Prof, Shalini Shenoy Prof, Joi Williamson, Teonna L Woolford, Eliane Gluckman Prof, John E Wagner Prof and John F Tisdale Prof

Lancet Haematology, The, 2019-11-01, Volume 6, Issue 11, Pages e585-e596, Copyright © 2019 Elsevier Ltd

2. Perform novel study on long-term survival for SCD transplant patients

3. Do 1 & 2 using *BDC-PIC-SURE*

Eapen et al. Lancet Haematol. 2019

- N=910 SCD patients in the US who received HCT between 2008 and 2017
- Donor types: Matched sibling (61%), haplo relative (15%), matched unrelated donor (12%), and mismatched unrelated donor (12%)
- Median follow-up by donor group: 36, 25, 37, 47 months respectively
- Recipient age, donor type, and conditioning intensity were found to be significantly associated with event-free survival following HCT

Publication data vs BDC

		Publication Results			Reproducible Results	
Event-free survival	Events/Patients	Hazard Ratio (95% confidence interval)	p-value	Events/Patients	Hazard Ratio (95% confidence interval)	p-value
Age						
≤12 years	72/491	1.00		90/491	1.00	
13 – 49 years	102/418	1.74 (1.24 – 2.45)	0.0014	138/418	1.66 (1.24 – 2.23)	0.0007
Regimen intensity						
Non-myeloablative	36/181	1.00		60/181	1.00	
Myeloablative	75/478	1.57 (0.95 – 2.61)	0.079	95/478	0.91 (0.61 – 1.36)	0.65
Reduced intensity	63/250	1.97 (1.15 – 3.36)	0.013	73/250	1.01 (0.66 – 1.57)	0.95
Donor type						
HLA-matched sibling	52/557	1.00		79/557		
Haploidentical relative	45/137	5.30 (3.17 – 8.86)	<0.0001	56/137	2.91 (1.96 – 4.34)	<0.0001
HLA-matched unrelated	38/111	3.71 (2.39 – 5.75)	<0.0001	45/111	3.11 (2.12 – 4.57)	<0.0001
HLA-mismatched unrelated	39/104	4.34 (2.58 – 7.32)	<0.0001	48/104	4.38 (3.02 – 6.35)	<0.0001

- Results aren't identical due to updated follow-up annually, but consistent

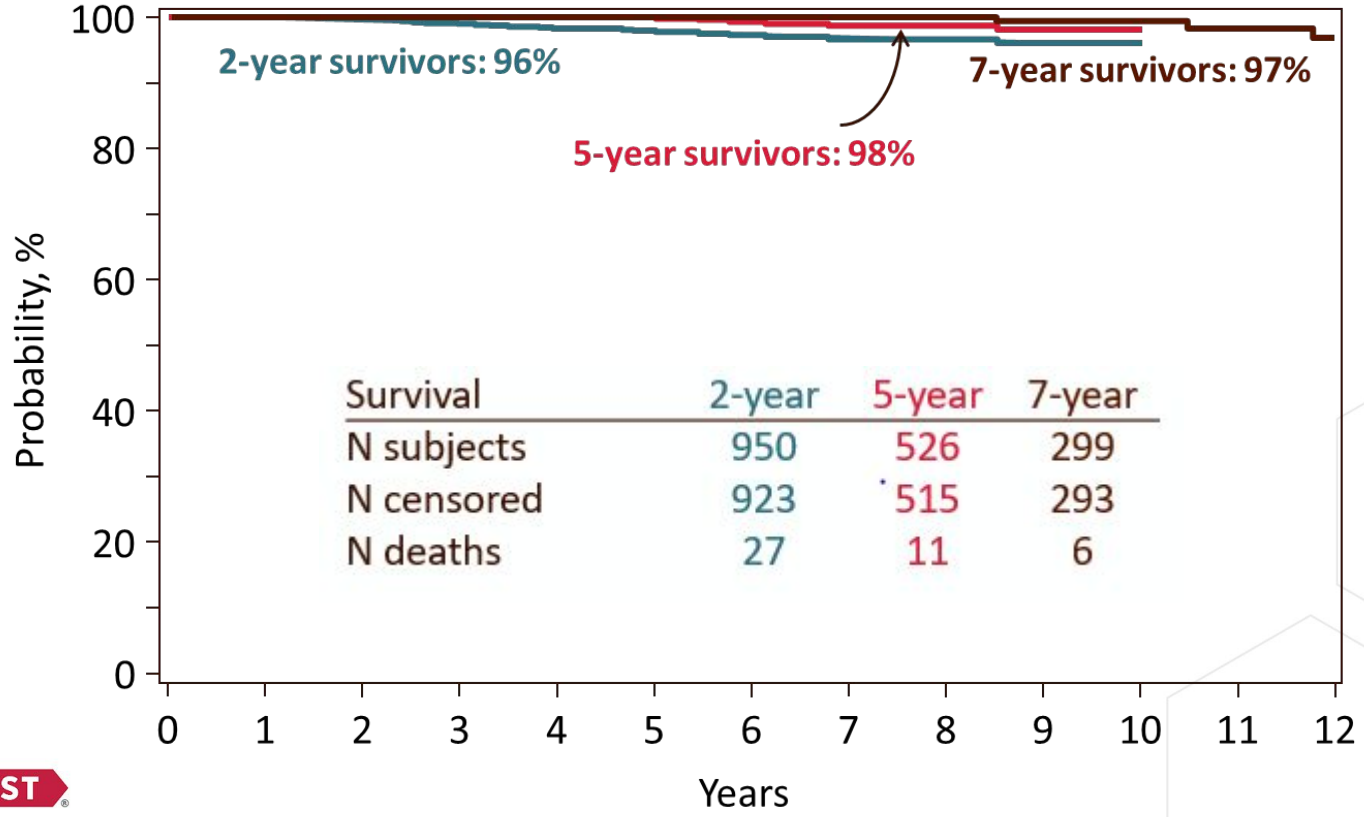
Novel research: long-term survival

- Life expectancy of adults with SCD in the US is shortened by ≥2 decades compared to the general population
- HCT is associated with life threatening complications, most of which occur within 2 years after transplantation
- If patients survive the early mortality risks, how do they fare long-term?
- Expanded population to all SCD patients in BDC who survived at least 2 years after HCT

Study population: first HCT for SCD in US, 2000-2017

CHARACTERISTIC	
Number	950
Age, years, median (range)	11 (<1 – 57)
18 – 29 years	15%
≥ 30 years	8%
Race – African American	92%
Donor type	
HLA-identical sibling	66%
HLA-mismatched relative	13%
HLA-matched unrelated donor	10%
HLA-mismatched unrelated donor	11%
Graft type – bone marrow	71%
Conditioning intensity - myeloablative	53%
Follow-up, years, median (range)	5 (2 - 20)

Overall Survival: 10-year Probability of Survival



Survival	2-year	5-year	7-year
N subjects	950	526	299
N censored	923	515	293
N deaths	27	11	6

Risk Factors – Death beyond 2 years after HCT

RISK FACTOR	HAZARD RATIO	P-VALUE
Age (10-year increment)	1.75	0.0004
Donor type		
HLA-matched sibling	1.00	
Alternative donors	3.49	0.0030

- Every 10-year increment in patient age at HCT was associated with a 75% increased risk for death
- Risk for death is 3.5 times higher after alternative donor compared to HLA-matched sibling HCT
  - Among the alternative donors (mismatched relative, matched unrelated, and mismatched unrelated), mortality risks were not different (p=0.48)

Standardized Mortality Ratio to that of U.S. Population

SURVIVAL TIME AFTER HCT	RISK FOR DEATH (95% CONFIDENCE INTERVAL)	P-VALUE
2-years	**6.9** (4.9 – 9.6)	<0.0001
5-years	**3.8** (2.1 – 6.2)	<0.0001
7-years	**3.2** (1.4 – 6.4)	0.0200

- Risk for death those who survived at least 2 years after HCT was 7-fold higher compared to an age, sex, and race matched general US population
- For who survived 5 and 7 years after HCT, the risk of death was 4- and 3- fold higher compared to an age, sex, and race matched general US population

Project Summary

- Data upload to BDC is seamless and the PIC-SURE application programming interface (API) easily enables analyses reproduction of published data
- Long-term survival is excellent for SCD patients who survive ≥2 years after HCT despite lower life expectancy than that of an age, sex, race matched U.S. population
    - The risk recedes over time but remains higher as late as 7-years after HCT

Long-term Survival after Hematopoietic Cell Transplant for Sickle Cell Disease Compared to the United States Population

ORCHID Clinical Trial: Results reproduction and extension using BDC

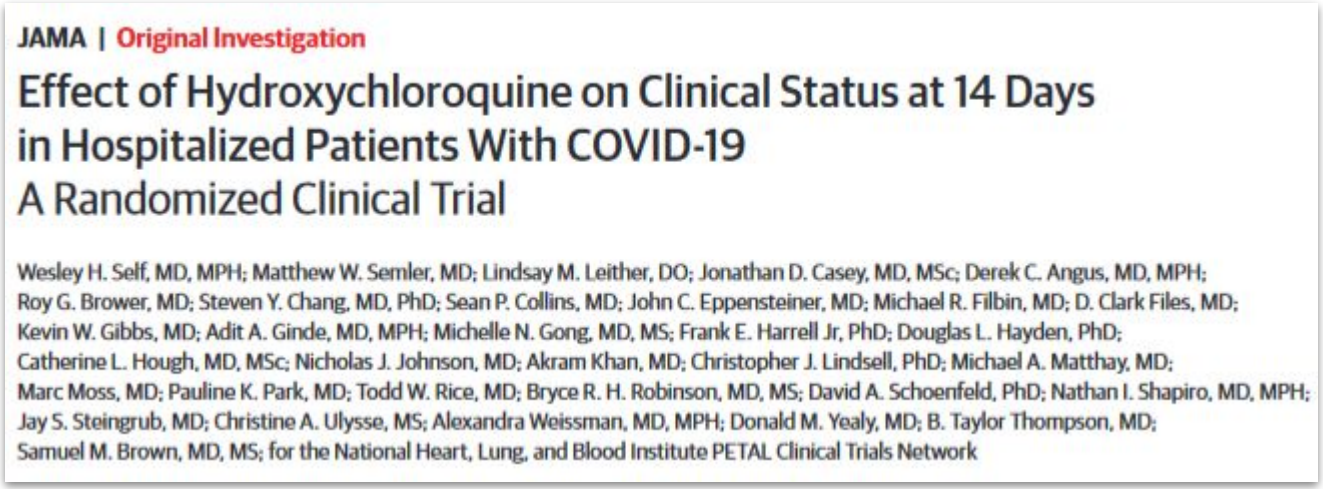
Acknowledgements

ORCHID reproduction and slides created by
Arnaud Serret-Larmande

As part of Paul Avillach's lab with the Department of Biomedical Informatics at Harvard Medical School

Using BDC and PIC-SURE to reproduce and expand on published results

On November 9th, 2020, publication of the results of the ORCHID clinical trial:



JAMA | Original Investigation

Effect of Hydroxychloroquine on Clinical Status at 14 Days in Hospitalized Patients With COVID-19
A Randomized Clinical Trial

Wesley H. Self, MD, MPH; Matthew W. Semler, MD; Lindsay M. Leither, DO; Jonathan D. Casey, MD, MSc; Derek C. Angus, MD, MPH; Roy G. Brower, MD; Steven Y. Chang, MD, PhD; Sean P. Collins, MD; John C. Eppensteiner, MD; Michael R. Filbin, MD; D. Clark Files, MD; Kevin W. Gibbs, MD; Adit A. Ginde, MD, MPH; Michelle N. Gong, MD, MS; Frank E. Harrell Jr, PhD; Douglas L. Hayden, PhD; Catherine L. Hough, MD, MSc; Nicholas J. Johnson, MD; Akram Khan, MD; Christopher J. Lindsell, PhD; Michael A. Matthay, MD; Marc Moss, MD; Pauline K. Park, MD; Todd W. Rice, MD; Bryce R. H. Robinson, MD, MS; David A. Schoenfeld, PhD; Nathan I. Shapiro, MD, MPH; Jay S. Steingrub, MD; Christine A. Ulysse, MS; Alexandra Weissman, MD, MPH; Donald M. Yealy, MD; B. Taylor Thompson, MD; Samuel M. Brown, MD, MS; for the National Heart, Lung, and Blood Institute PETAL Clinical Trials Network

Finds no statistically significant benefit of hydroxychloroquine over placebo in the treatment of patients hospitalized for COVID-19

How to infuse confidence in these results?

- Results noticeable, in the context of the polemics surrounding usage of hydroxychloroquine as a COVID-19 treatment
- Around the same time, numerous publications about COVID-19 had been retracted (69 referenced by retractionwatch.com as of February 7th 2021)

⇒ Trust in these results can be reinforced through transparency and reproducibility of the analysis

A framework to streamline reproducibility

BDC:

- Ease process for investigators to upload their data and to manage access requests
- Ease process for investigators willing to find and be granted access to the data

PIC-SURE Application Programming Interface (API):

- Provide data access under a simple and accessible format (following FAIR principles)
- Allows the production of analyses in a self-contained format (using Jupyter Notebook)
    - Can be run in one-click
    - Provide a quick-start for investigators willing to work on the data

Results of the reproduction process published in JAMIA Open

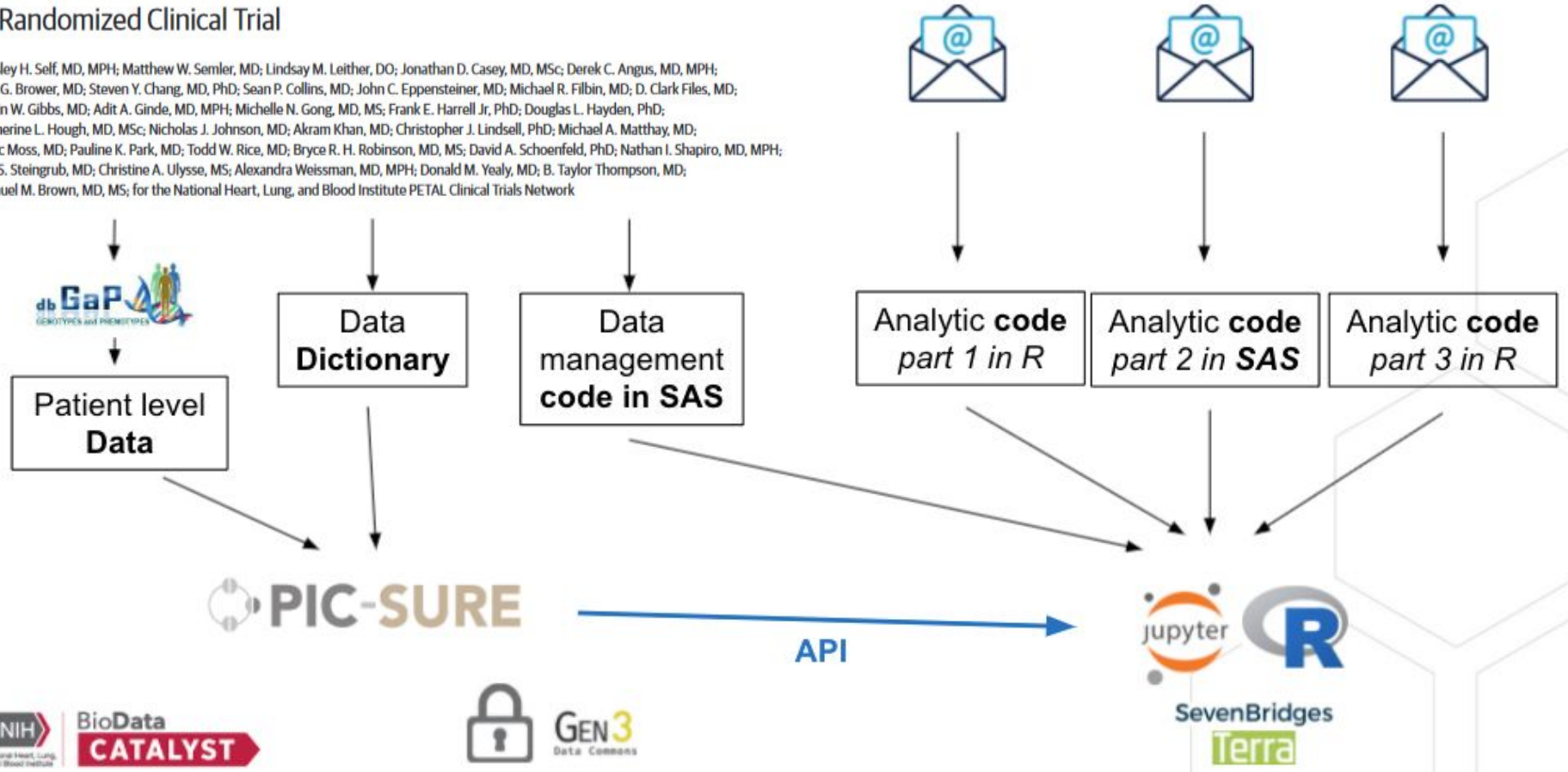
[Streamlining statistical reproducibility: NHLBI ORCHID clinical trial results reproduction]

JAMA | Original Investigation

Effect of Hydroxychloroquine on Clinical Status at 14 Days in Hospitalized Patients With COVID-19
A Randomized Clinical Trial

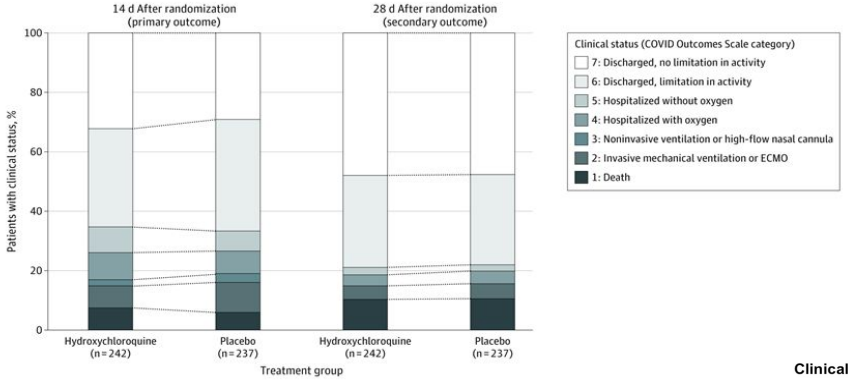
Wesley H. Self, MD, MPH; Matthew W. Semler, MD; Lindsay M. Leither, DO; Jonathan D. Casey, MD, MSc; Derek C. Angus, MD, MPH; Roy G. Brower, MD; Steven Y. Chang, MD, PhD; Sean P. Collins, MD; John C. Eppensteiner, MD; Michael R. Filbin, MD; D. Clark Files, MD; Kevin W. Gibbs, MD; Adit A. Ginde, MD, MPH; Michelle N. Gong, MD, MS; Frank E. Harrell Jr, PhD; Douglas L. Hayden, PhD; Catherine L. Hough, MD, MSc; Nicholas J. Johnson, MD; Akram Khan, MD; Christopher J. Lindsell, PhD; Michael A. Matthay, MD; Marc Moss, MD; Pauline K. Park, MD; Todd W. Rice, MD; Bryce R. H. Robinson, MD, MS; David A. Schoenfeld, PhD; Nathan I. Shapiro, MD, MPH; Jay S. Steingrub, MD; Christine A. Ulysse, MS; Alexandra Weissman, MD, MPH; Donald M. Yealy, MD; B. Taylor Thompson, MD; Samuel M. Brown, MD, MS; for the National Heart, Lung, and Blood Institute PETAL Clinical Trials Network

Original analysis



Main results reproduced identically

Figure 2. Clinical Status on the Coronavirus Disease (COVID) Outcomes Scale 14 Days and 28 Days After Randomization

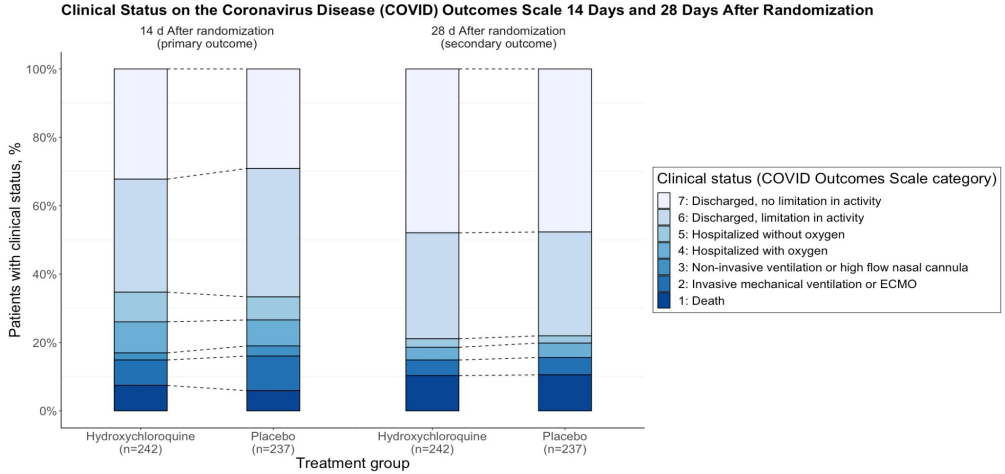


JAMA The Journal of the American Medical Association

Published online on November 9th, 2020

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

Independent main result reproduction uploaded on GitHub, December 10th, 2020

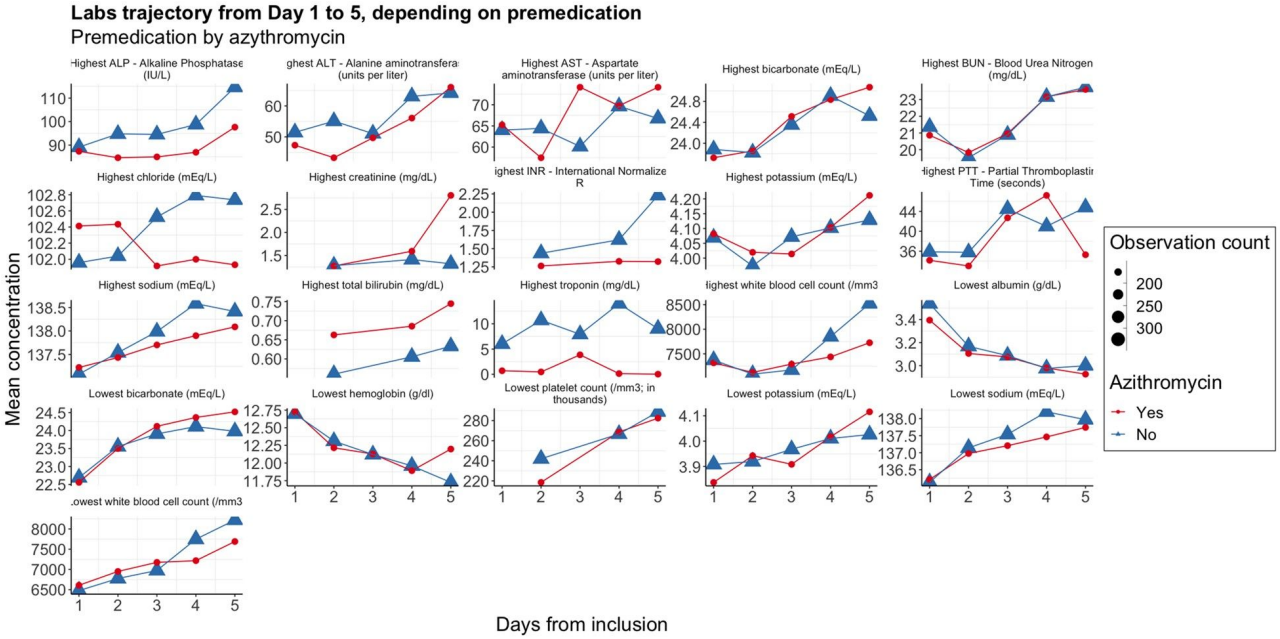


NIH National Heart, Lung, and Blood Institute | BioData CATALYST®

And expansion of the results published in the original article

Study of the labs trajectory according to premedications

More extreme values if premedicated by azithromycin, highlighting the baseline characteristic differences within treatment arms



Take-home messages from this effort

- Delivers a simple and easy-to-follow workflow, leveraging open-source tools to publish fully reproducible and transparent results
- Lay out the path for other research efforts **willing to use BDC to improve transparency of their results**, strengthening their findings

Link to publication:

Streamlining statistical reproducibility: NHLBI ORCHID clinical trial results reproduction

Workflows & Cost Estimation

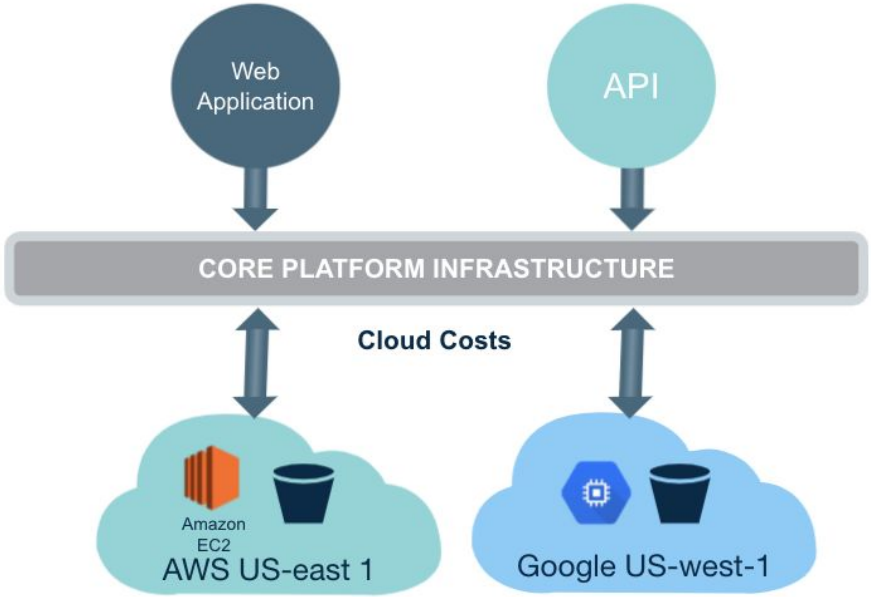
Curricula of:

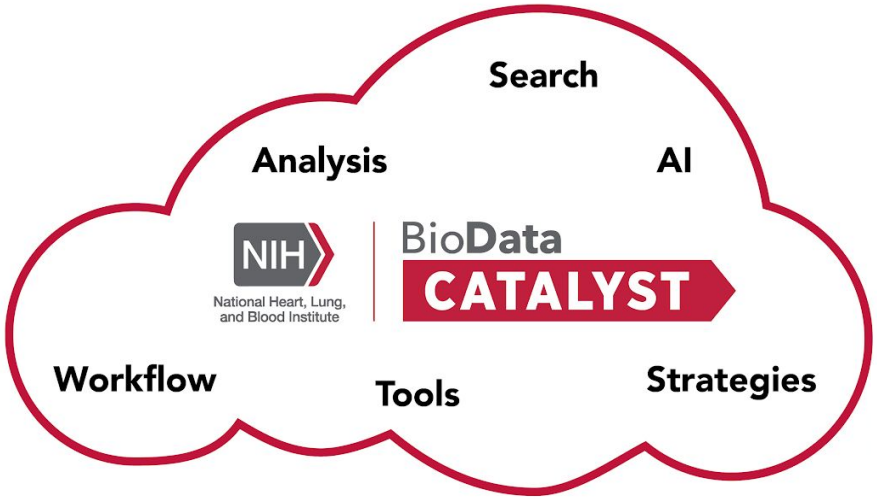
Workflows & Cost Estimation

Learning goals

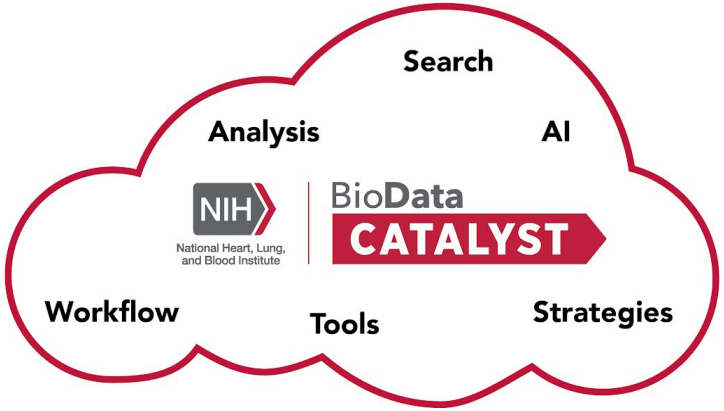
- Compare and estimate costs of using the cloud
- Request pilot credits from BDC
- Incorporate costs into a grant proposal

What is Cloud Computing?





Cloud or High Performance Computing?



- Faster and scalable
- Easier collaboration
- Access
- Use what you need

- Up-front costs
- Maintenance costs
- Institutional fee
- Queue

Where do cloud costs come from?

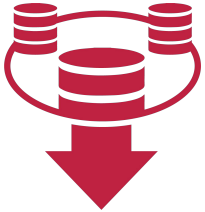
Computation

- CPU hours
- Instance type

Data storage

- Hosted data
- Derived data
- Egress

Value of cloud computing



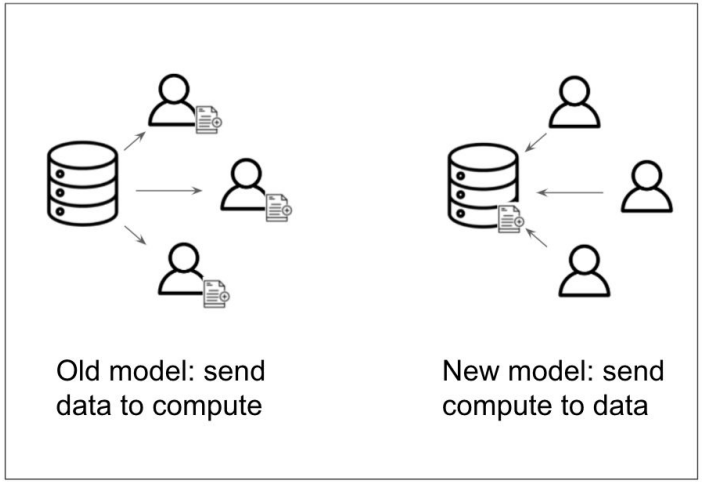
No need to **download** + **manage** (multiple) large datasets

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

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

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.



Request Pilot Credits

What are Cloud Credits?

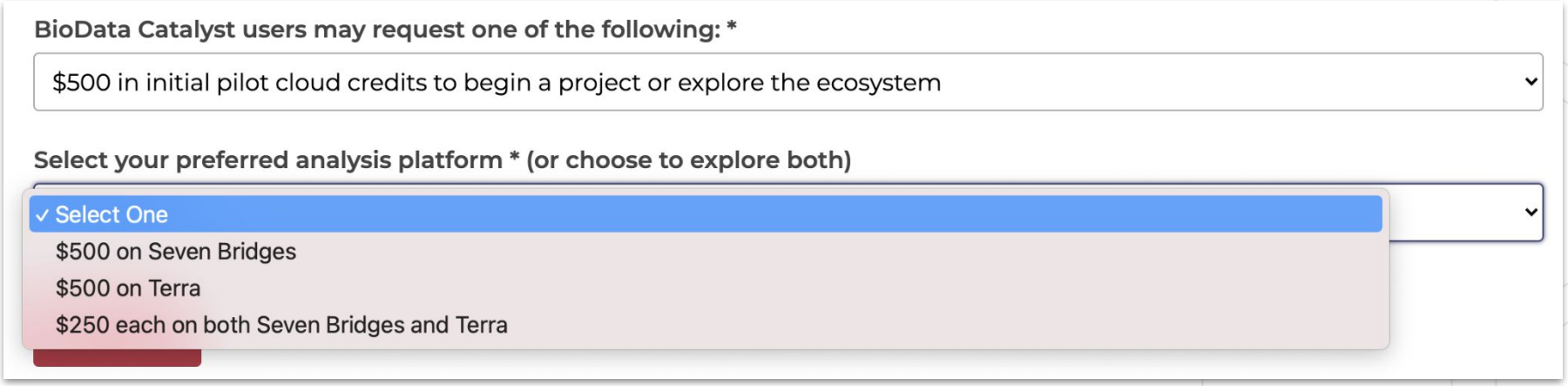
Users are not charged for the storage of hosted datasets; however, if hosted data is used in analyses, users incur costs for computation and storage of derived results.

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

Web resource: Cloud Costs and Credits

Try out BDC with Pilot Credits

- New users of BDC may apply for an initial $500 in cloud credits, known as **pilot credits**; many analyses can be completed for that amount or less.
- For larger tasks, you can use the credits to test and evaluate the ecosystem for things such as piloting pipelines on smaller samples and estimating how much a full analysis will cost.



Cloud Credits Workflow

1

Sign up for the community

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

2

Sign up for Seven Bridges

Seven Bridges and/or Terra

3

Apply for Pilot Credits

Fill out the Cloud Credits Request form.

Use all credits on a single platform, or split.

4

Apply for additional credits or pay yourself

Cover costs after pilot funding has been exceeded.

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

Web Form

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

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

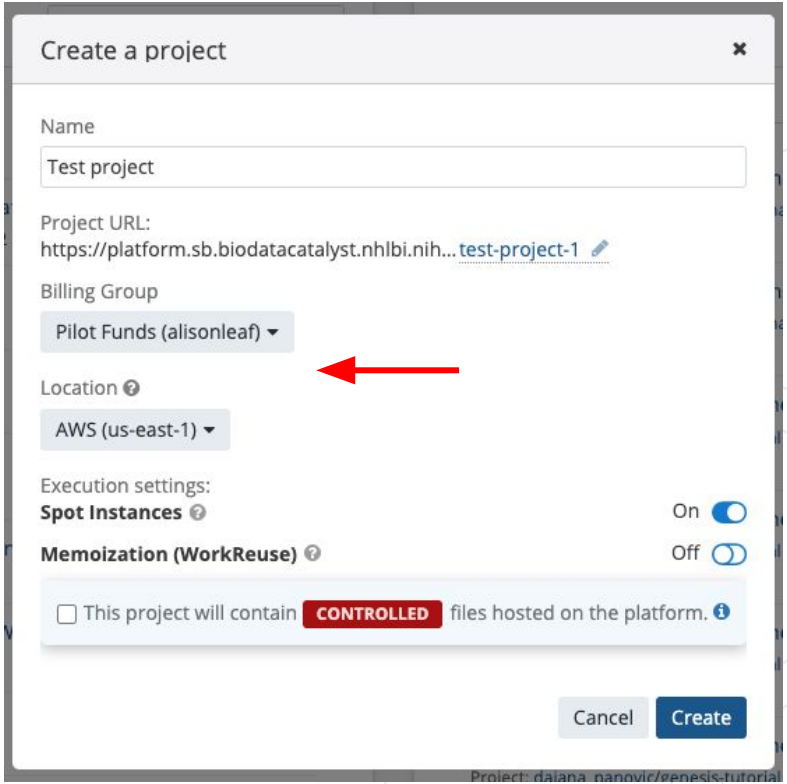
Breakout: Request Pilot Credits

- Follow the Cloud Credits workflow to request $500 in pilot funding
  - Join the Community
  - Sign up for Seven Bridges
  - Apply for Pilot Credits on the BDC website

Benchmarking Pilot Analyses: Best Practices

A billing group is auto created

- A **Pilot Funds** billing group is automatically created for every new user
- This can be populated with up to **$500 in credits** by a request to NHLBI
- **Billing groups** are easy to select during project creation or later in the project settings

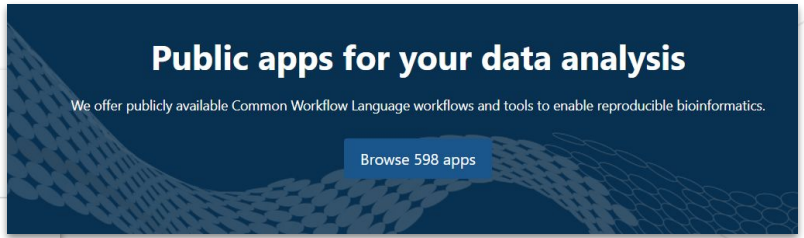


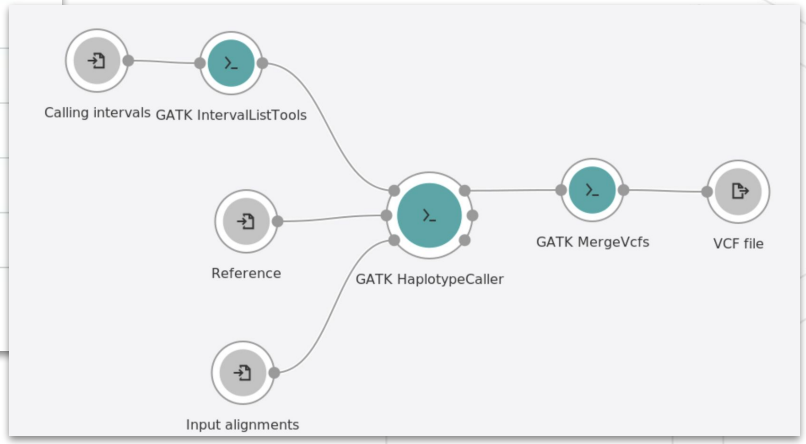
Using Seven Bridges benchmarking information

Broad Best Practice Germline snps and indels variant calling 4.1.0.0

Performance Benchmarking

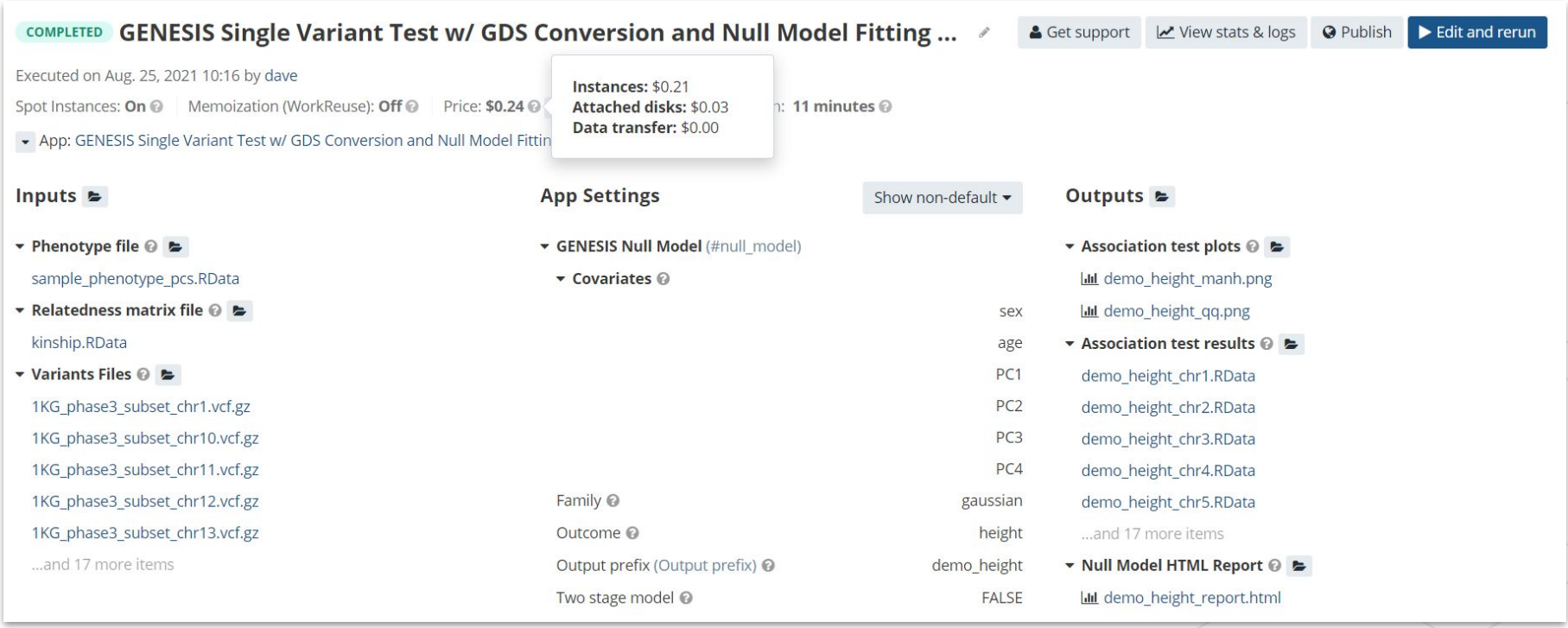
BAM Input size	Experiment type	Coverage	Duration	Cost	Instance
55.8GiB	WGS (scatter count = 20)	~50x	17h 35min	$9.42	c4.2xlarge
55.8GiB	WGS (scatter count = 80)	~50x	10h 32min	$5.64	c4.2xlarge
24.6GiB	WGS (scatter count = 80)	~10x	4h 12min	$2.25	c4.2xlarge
3.5GiB	WES (scatter count = 1)	~70x	17min	$0.16	c4.2xlarge
1.9GiB	WES (scatter count = 1)	~40x	11min	$0.11	c4.2xlarge
1.1GiB	WES (scatter count = 1)	~20x	9min	$0.08	c4.2xlarge
434MiB	WES (scatter count = 1)	~10x	6min	$0.06	c4.2xlarge





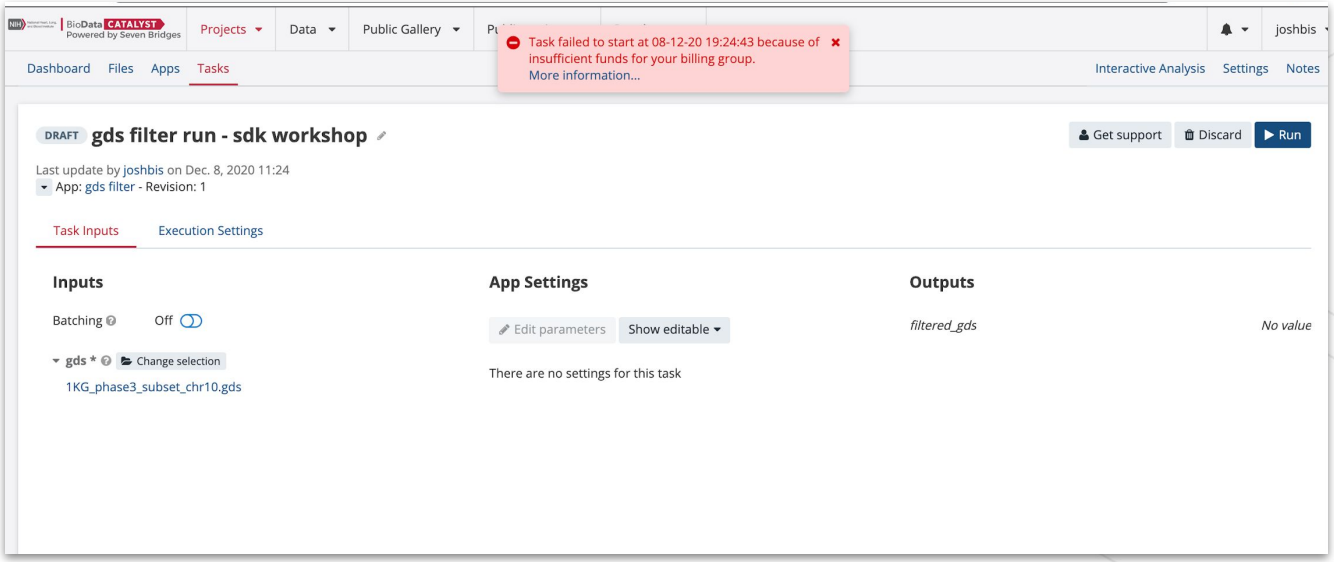
NIH National Heart, Lung, and Blood Institute | BioData CATALYST

Tasks have detailed credit usage information



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

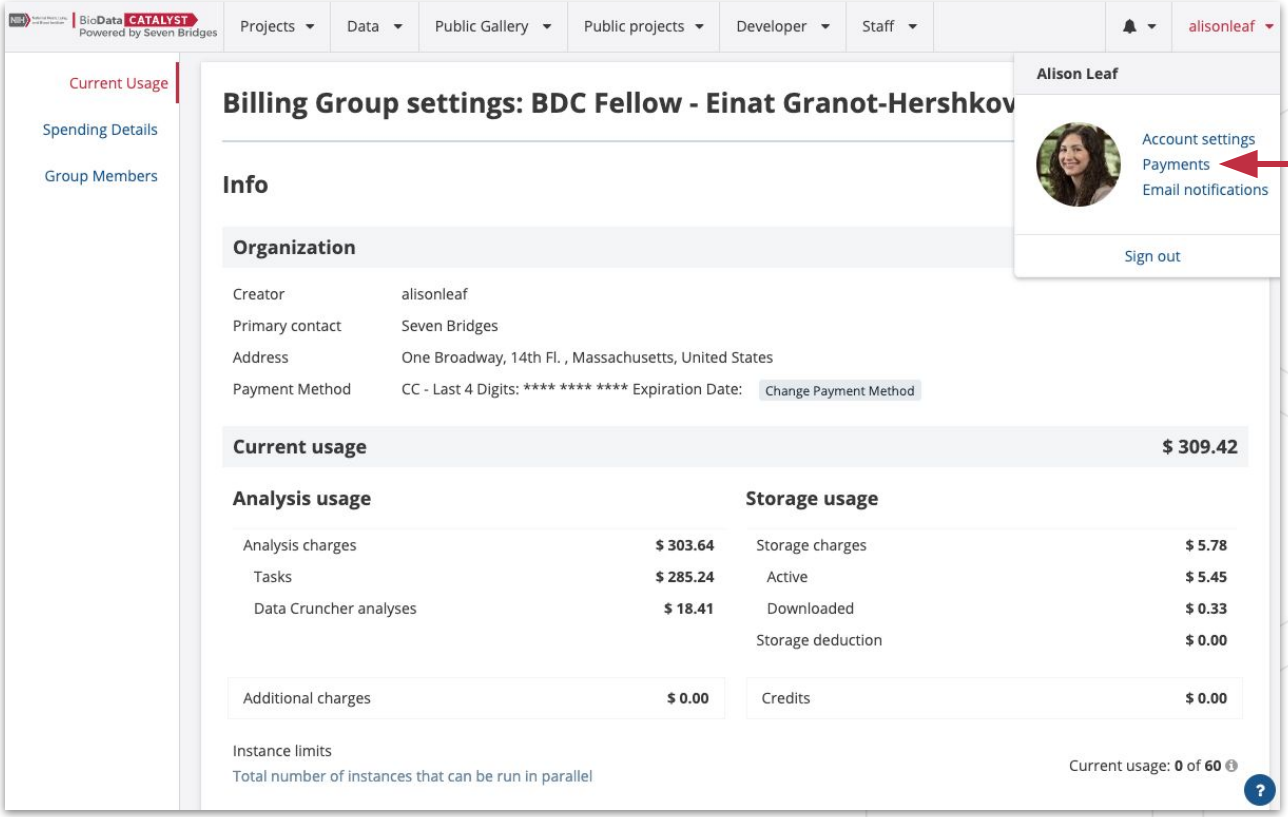
If you try to execute an analysis after you run out of funding in billing group, the platform will not allow the analysis to start and you will see an error message “insufficient funds in billing group.”



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

Track costs on platform payments page

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



Budgeting for Scale

Budgeting for Scale: Computation

Performance Benchmarking

Below is a table describing the runtimes and task costs for a couple of samples with different file sizes, with the following workflow options in mind - indexing is not performed, unmapped reads are sorted by read id, output BAM is sorted by coordinate and basic two pass mode is turned on:

Experiment type	Input size	Paired-end	# of reads	Read length	Duration	Cost	Instance (AWS)
RNA-Seq	2 x 230 MB	Yes	1M	101	18min	$0.40	c4.8xlarge
RNA-Seq	2 x 4.5 GB	Yes	20M	101	30min	$0.60	c4.8xlarge
RNA-Seq	2 x 17.4 GB	Yes	76M	101	64min	$1.20	c4.8xlarge

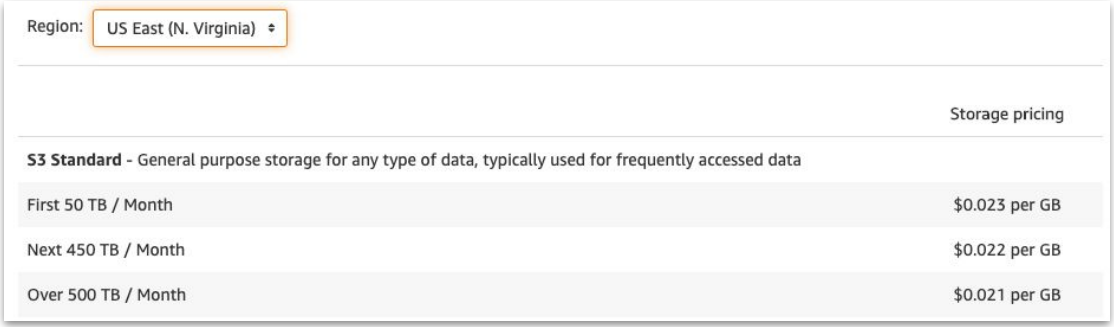
*Cost can be significantly reduced by using **spot instances**. Visit the Knowledge Center for more details.*

User brings 100 RNA-seq samples

~ 4 GB each; $0.60 per STAR run.

$0.60 per run * 100 samples = **$60**

Budgeting for Scale: Storage



Region: US East (N. Virginia)

	Storage pricing
S3 Standard - General purpose storage for any type of data, typically used for frequently accessed data	
First 50 TB / Month	$0.023 per GB
Next 450 TB / Month	$0.022 per GB
Over 500 TB / Month	$0.021 per GB

User brings 100 RNA-seq samples to the platform

~ 4 GB each x2; User stores files on platform for 12 months. Total size ~800 GB.

$0.023 per GB * 800 GB = $18.40 per month.

$220.80 to store files on platform for 12 months

Writing Cost into a Grant Proposal

Estimating Cloud Costs

Researchers incur fees for:

- Data Storage
- Computing / Analysis
- Egress charges
- Help Desk / Platform Support *

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.

Estimating Cloud Costs

Help Desk / Platform Support

General support is included on all ecosystem platforms free of charge, but some projects may require more support than others.

If you anticipate needing **a large amount of support** for activities, reach out to the platform liaisons to research what charges you may accrue.

Writing a Budget Justification

- **BioData Catalyst Data Storage**
- **BioData Catalyst Analysis Costs**
- **BioData Catalyst Egress Fees**
- **Liaison support**

Resource: We are working on **sample language** for writing a budget justification for including BioData Catalyst in a proposal. Contact us, or keep watch on our documentation & forum for its publication.

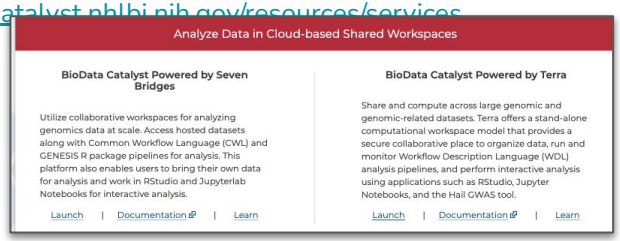
Resources for estimating your cloud costs

- BioData Catalyst Powered by Terra: Understanding and controlling Cloud costs
- BioData Catalyst Powered by Seven Bridges: Estimate and Manage Your Cloud Costs
- NHLBI BioData Catalyst help desk

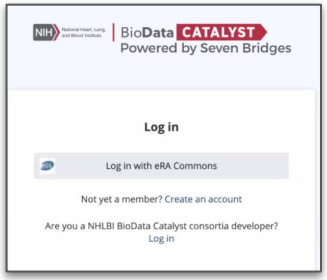
While You Are Waiting for Funding

BioData Catalyst Powered by Seven Bridges

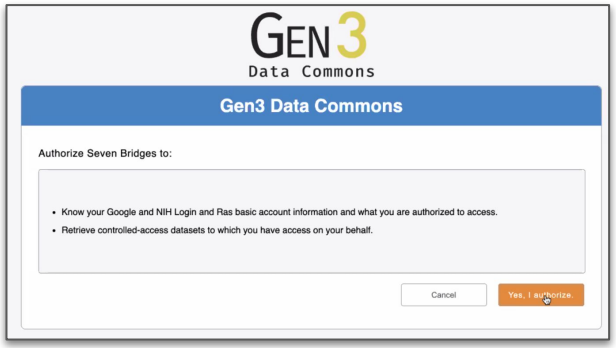
1. Launch: https://biodatacatalyst.nhlbi.nih.gov/resources/services



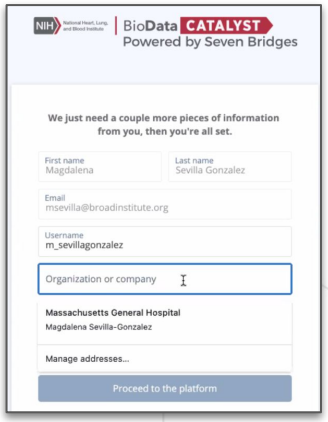
2. Login with **eRA Commons ID**



3. Confirm connection to Gen3 Data Commons

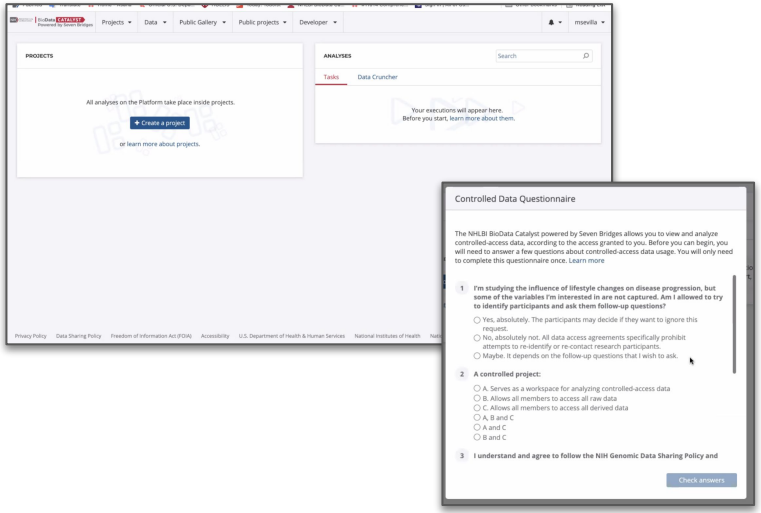


4. Provide your information

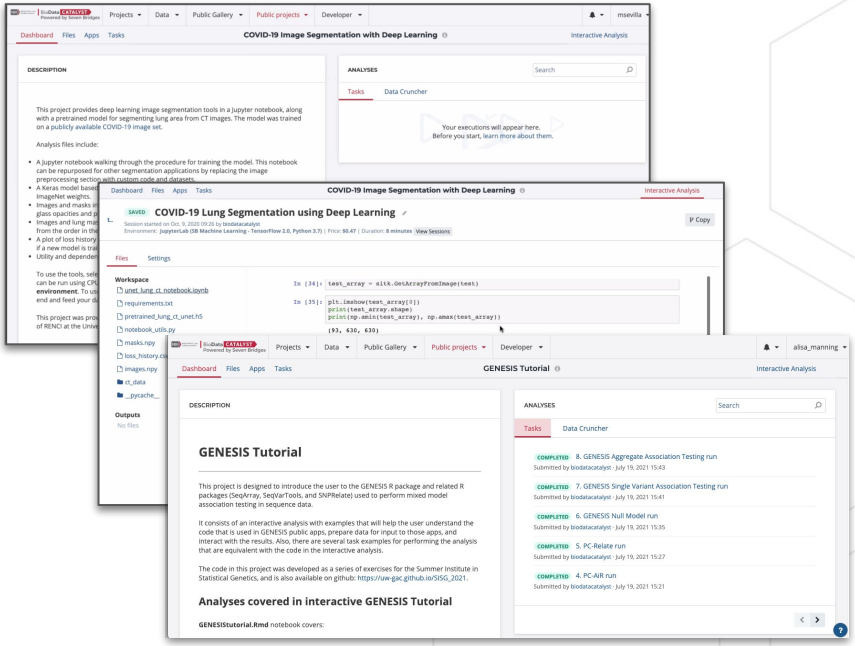


BioData Catalyst Powered by Seven Bridges

5. Complete **Controlled Data Questionnaire** and gain access to project and analysis dashboards



6. Browse public projects; view interactive analysis in **Preview** and see details from analysis pipelines in **Tasks**

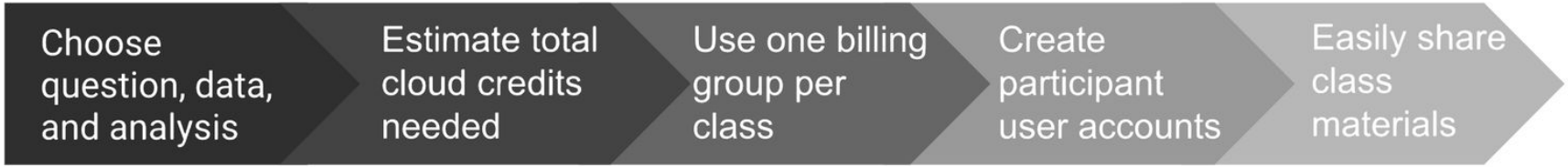


Using BDC as a Teaching Platform

Agenda

- The BDC approach to teaching
- Recent examples of using BDC in a course or workshop
- Guest lecturer and Q+A
    - Use Case: University of Washington Summer Institute in Statistical Genetics
    - Matt Conomos, UW
- Breakouts with Jamboard

Systematic Approach to Teaching



Flexible: UI and coding/API options available. Different options for different student requirements

- Private JupyterLab, RStudio and SAS Studio environments available
- Simple visual representation of the steps in the workflow
- Easy for participants to grasp basic concepts, before using command line
- Easy to run pipelines, mostly point and click
- Easy for instructor to monitor in class work and homework
- One billing group allows for easy monitoring of resources
- Collaborative environment with BDC providing support to the class
- Participants are encouraged to attend BDC/Seven Bridges Office Hours and Community Hours

Useful Guide

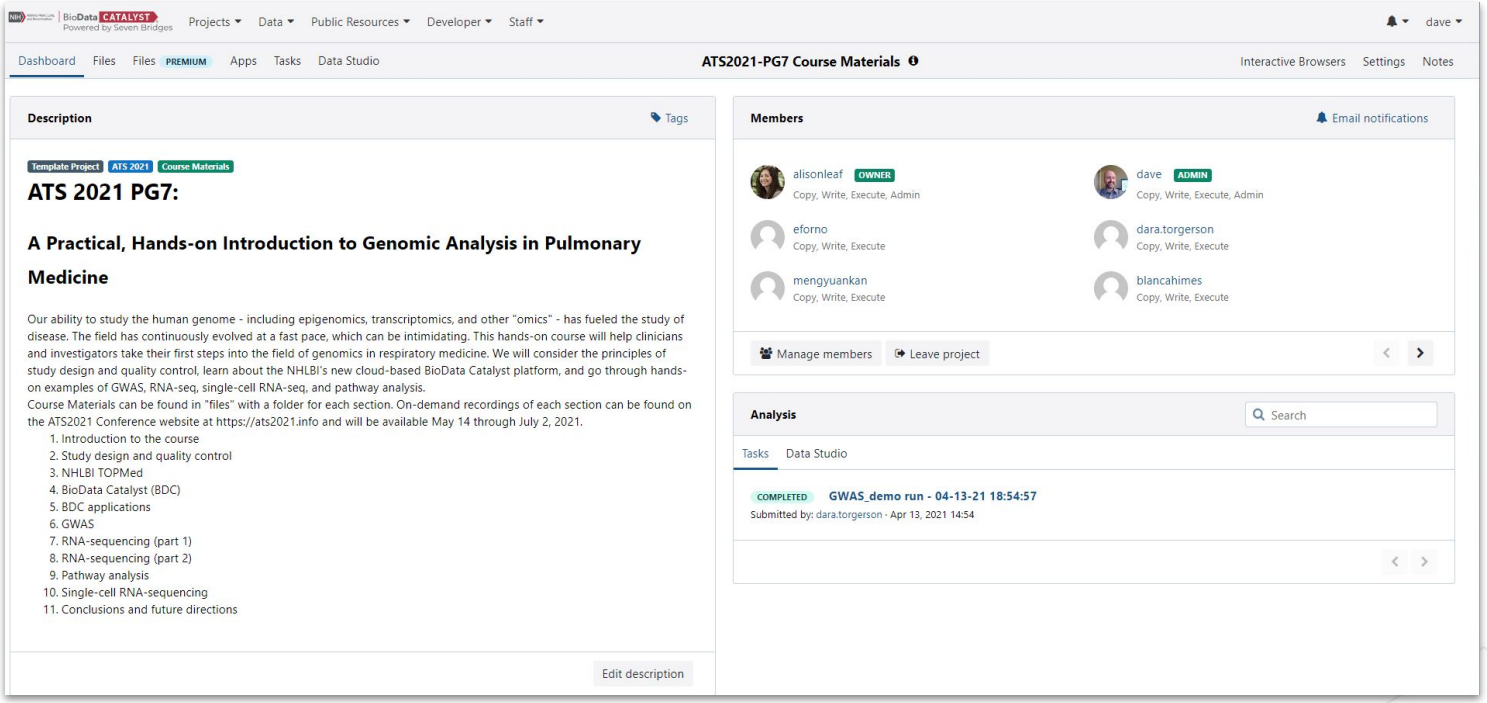
Using BDC for Workshops and Courses

Case studies

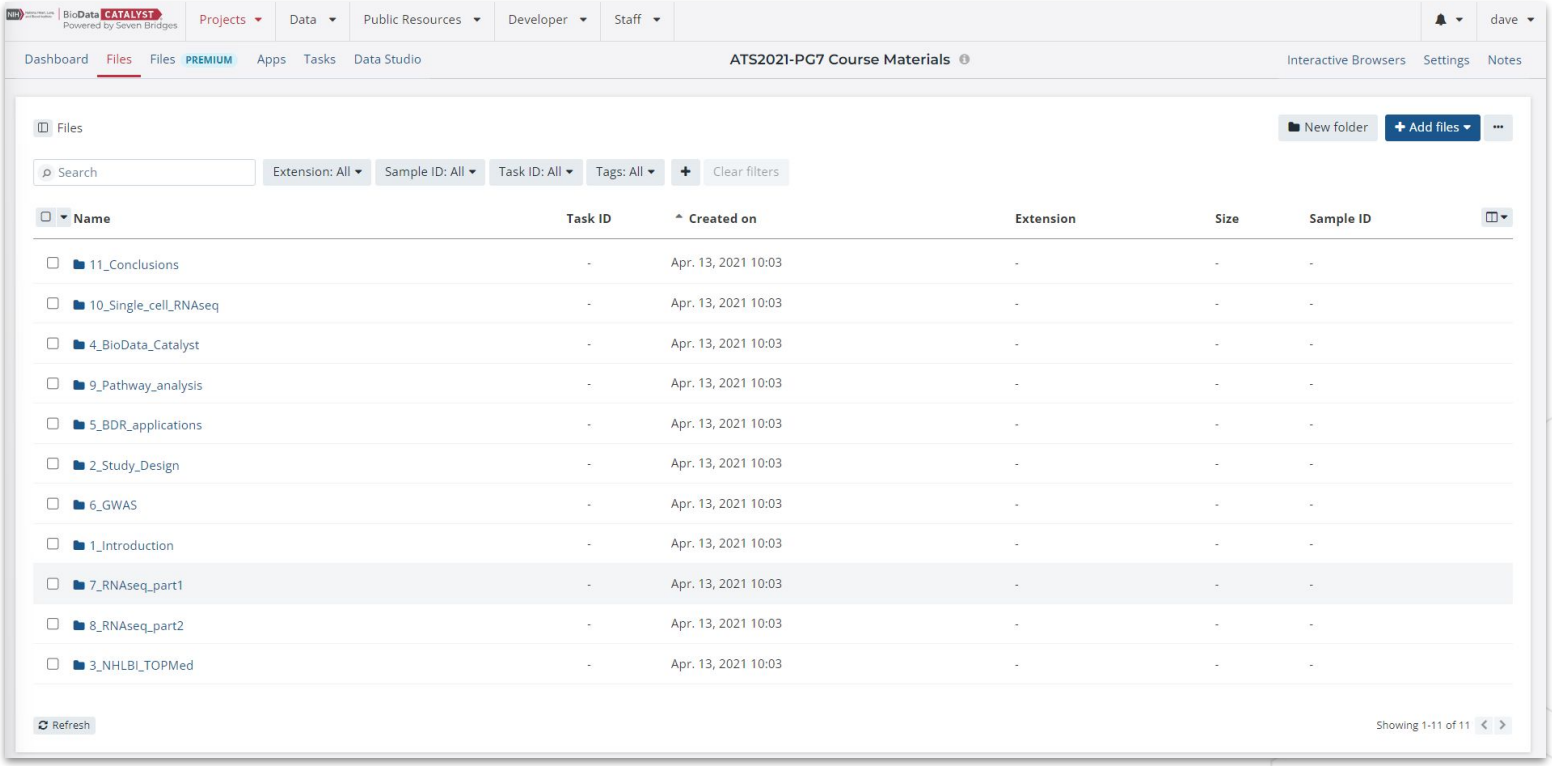
- UW Summer Institutes in Statistical Genetics
- Genomics course for American Thoracic Society Annual Meeting

https://sb-biodatacatalyst.readme.io/docs/using-nhlbi-biodata-catalyst-for-workshops-and-courses

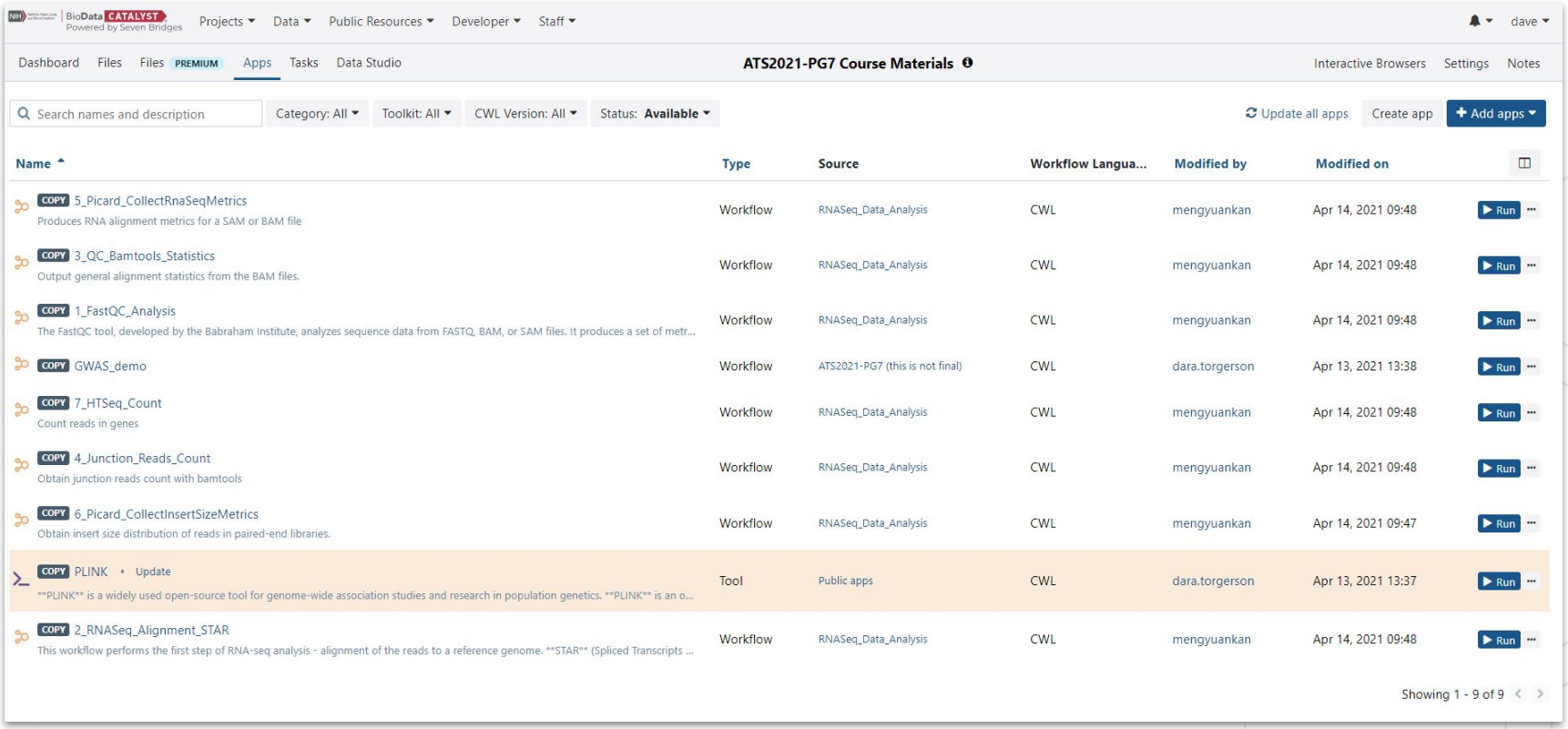
Dashboard provides an overview



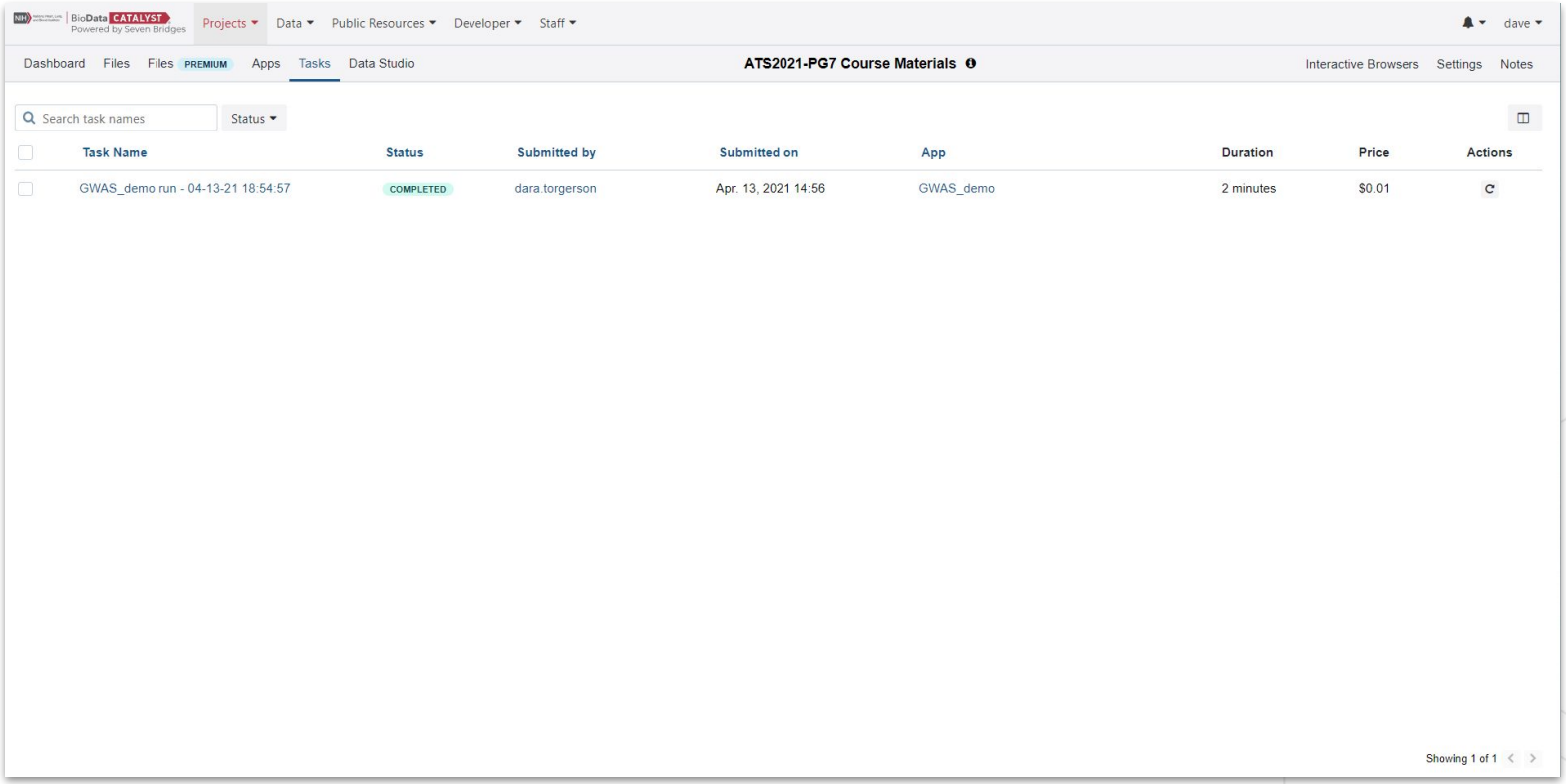
Organize course materials in folders



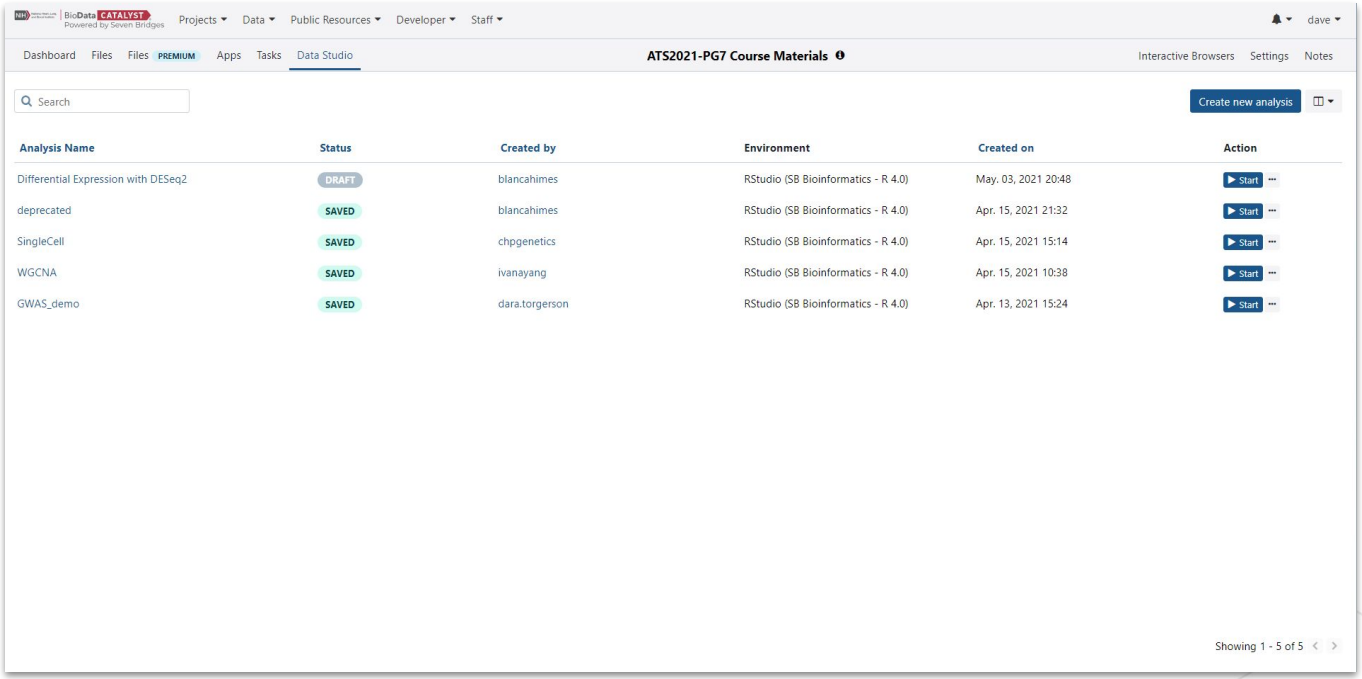
Stage apps for students for quick wins



Pre-run example tasks



Multiple pre-setup scripting environments per class



Success stories

- **Georgetown University**
  - Using a Seven Bridges platform for the past 3 years to train the next generation of data scientists in their **Masters in Health Informatics and Data Science Program**
- **Purdue University**
  - Partnered with Min Zhang at Purdue University to deliver a **four-part series on RNAseq** as part of their STAT-581 course.
- **Summer Institute in Statistical Genetics**
  - University of Washington has used BDC as a GWAS module teaching platform for several years
  - 75 students (each year) use RStudio simultaneously in their private projects

Howard Data Science Workshop

- Two day workshop covering the BioData Catalyst ecosystem and cloud computing
  - Finding data of interest
  - Linking data from other sources
  - Scaling up data analysis
  - Running a Genome Wide Association Study

- Organized by Dr. John Kwagyan

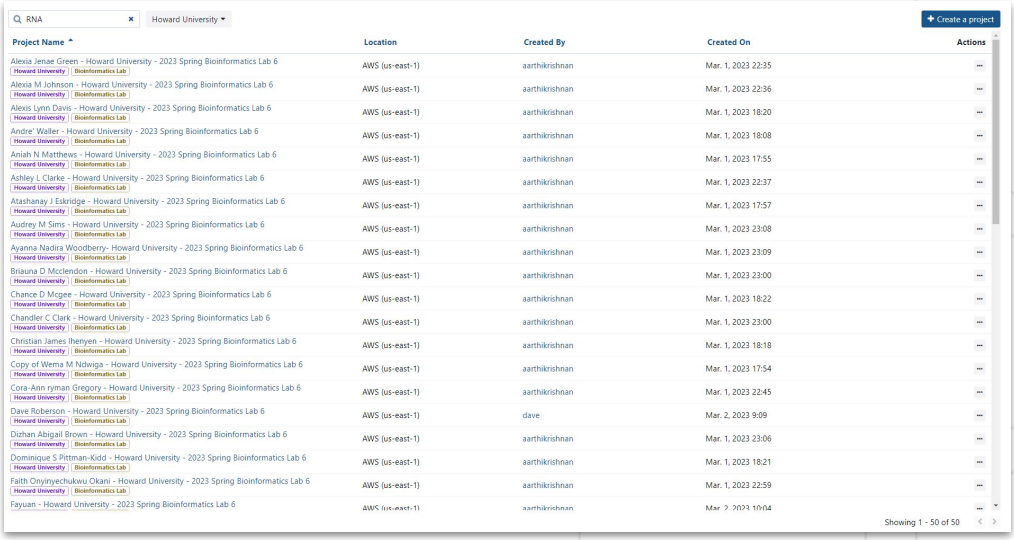
HOWARD UNIVERSITY | Research Centers in Minority Institutions RESEARCH FORWARD

Howard Bioinformatics Courses

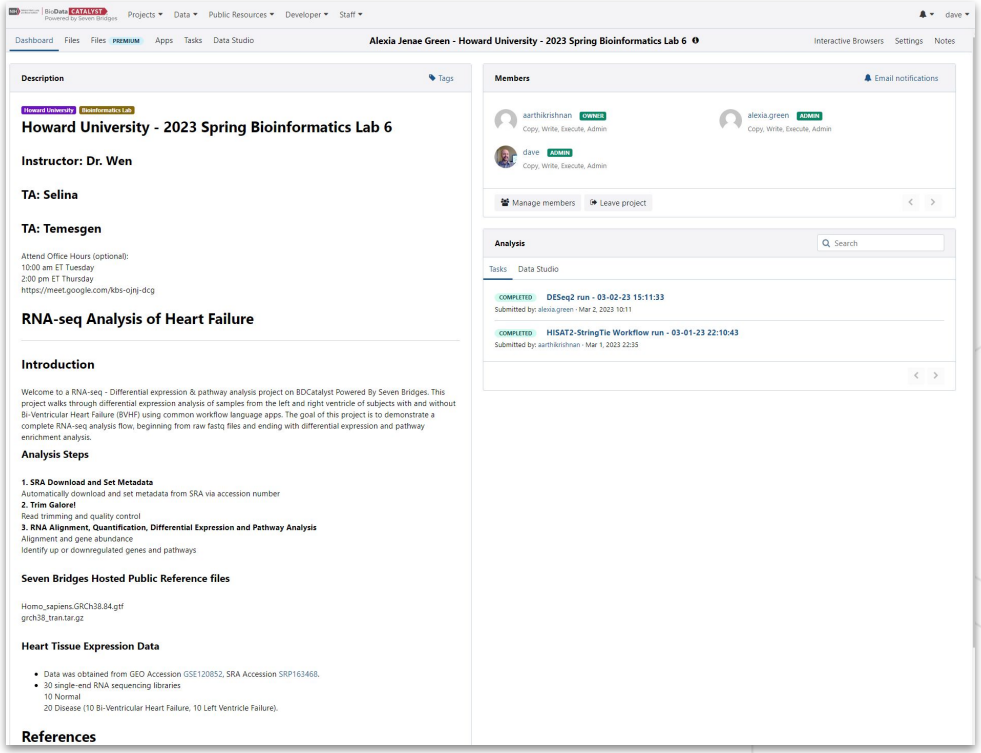
- Two lab sessions catered to advanced undergraduate and early graduate students
  - Genome Wide Association testing
  - RNA Whole Transcriptomic Analysis
- Credits provided by NHLBI
- More courses planned for Fall semester

Dr. Teng

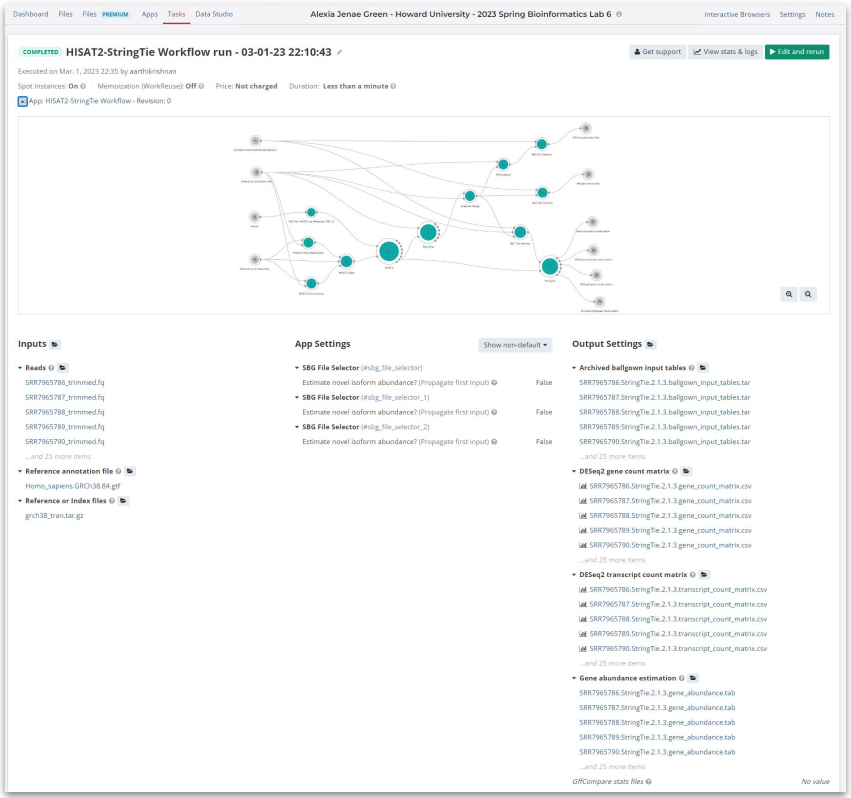
Organized by Dr. Teng and Dr. Fayuan Wen

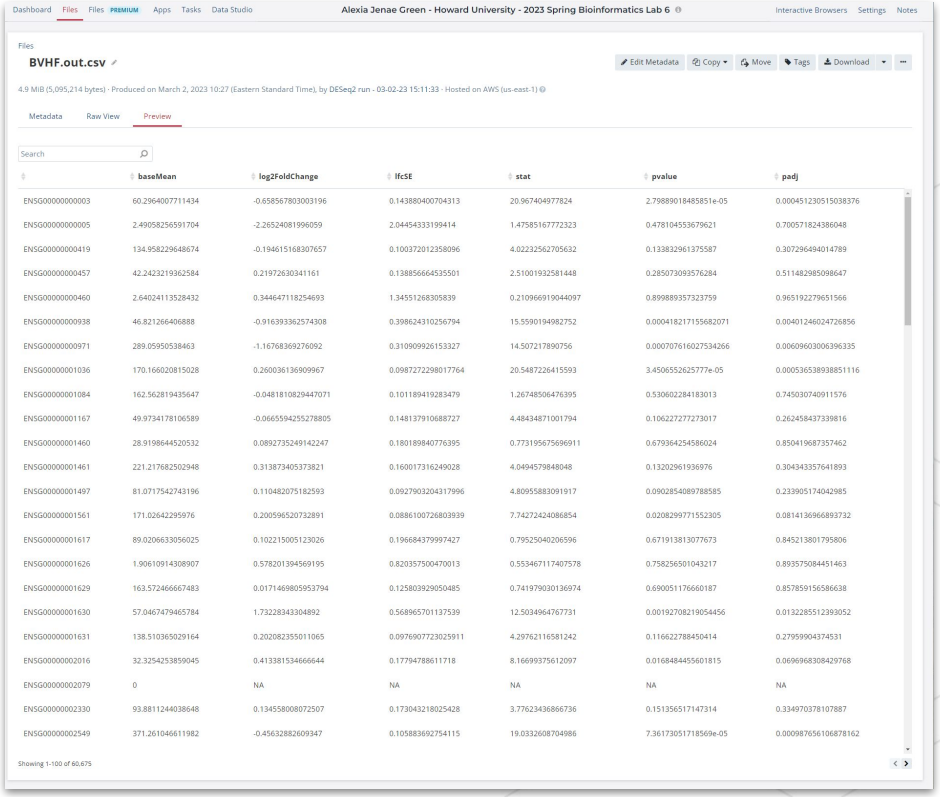


Enable students to do complex analysis workflows with a GUI



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





	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000000003	60.2964007711434	-0.658567803003196	0.143880400704313	20.967404977824	2.79889018485851e-05	0.000451230515038376
ENSG00000000005	2.49058256591704	-2.26524081996059	2.04454333199414	1.47585167772323	0.478104553679621	0.700571824386048
ENSG00000000419	134.958229648674	-0.194615168307657	0.100372012358096	4.02232562705632	0.133832961375587	0.307296494014789
ENSG00000000457	42.2423219362584	0.21972630341161	0.138856664535501	2.51001932581448	0.285073093576284	0.511482985098647
ENSG00000000460	2.64024113528432	0.344647118254693	1.34551268305839	0.210966919044097	0.899889357323759	0.965192279651566
ENSG00000000938	46.821266406888	-0.916393362574308	0.398624310256794	15.5590194982752	0.000418217155682071	0.00401246024726856
ENSG00000000971	289.05950538463	-1.16768369276092	0.319090926153327	14.507217890756	0.000707616027534266	0.00609603006396335
ENSG00000001036	170.166020815028	0.260036136909967	0.0987272298017764	20.5487226415593	3.45065526257777e-05	0.000536538938851116
ENSG00000001084	162.562819435647	-0.0481810829447071	0.101189419283479	1.26748506476395	0.530602284183013	0.745030740911576
ENSG00000001167	49.9734178106589	-0.0665594255278805	0.148137910688727	4.48434871001794	0.106227277273017	0.262458437339816
ENSG00000001460	28.9198644520532	0.0892735249142247	0.180189840776395	0.773195675696911	0.679364254586024	0.850419687357462
ENSG00000001461	221.217682502948	0.313873405373821	0.160017316249028	4.0494579848048	0.13202961936976	0.304343357641893
ENSG00000001497	81.0717542743196	0.110482075182593	0.0927903204317996	4.80955883091917	0.0902854089788585	0.233905174042985
ENSG00000001561	171.02642295976	0.200596520732891	0.0886100726803939	7.74272424086854	0.0208299771552305	0.0814136966893732
ENSG00000001617	89.0206633056025	0.102215005123026	0.196684379997427	0.795250402065956	0.671913813077673	0.845213801795806
ENSG00000001626	1.90610914308907	0.578201394569195	0.820357500470013	0.553467117407578	0.758256501043217	0.893575084451463
ENSG00000001629	163.572466667483	0.0171469805953794	0.125803929050485	0.741979030136974	0.690051176660187	0.857859156586638
ENSG00000001630	57.046747946784	1.73228343304892	0.568965701137539	12.5034964767731	0.00192708219054456	0.0132285512393052
ENSG00000001631	138.510365029164	0.202082355011065	0.0976907723025911	4.29762116581242	0.116622788450414	0.27959904374531
ENSG00000002016	32.3254253859045	0.413381534666644	0.17794788611718	8.16699375612097	0.0168484455601815	0.0696968308429768
ENSG00000002079	0	NA	NA	NA	NA	NA
ENSG00000002330	93.8811244038648	0.134558008072507	0.173043218025428	3.77623436866736	0.151356517147314	0.334970378107887
ENSG00000002549	371.261046611982	-0.456328826093347	0.105883692754115	19.0332608704986	7.36173051718569e-05	0.000987656106878162

Planning for Support

- BDC has a focused group on user engagement
  - Meets weekly
  - Can bring in additional support and planning
    - When you are ready please reach out
- Questions for discussion
  - Timeline
  - Background level of students
  - Data used in courses (what needs to be hosted ahead of time?)
  - Any student or faculty research projects BDC can help to support?

Use Case: University of Washington Summer Institute in Statistical Genetics

Matt Conomos, PhD
University of Washington

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

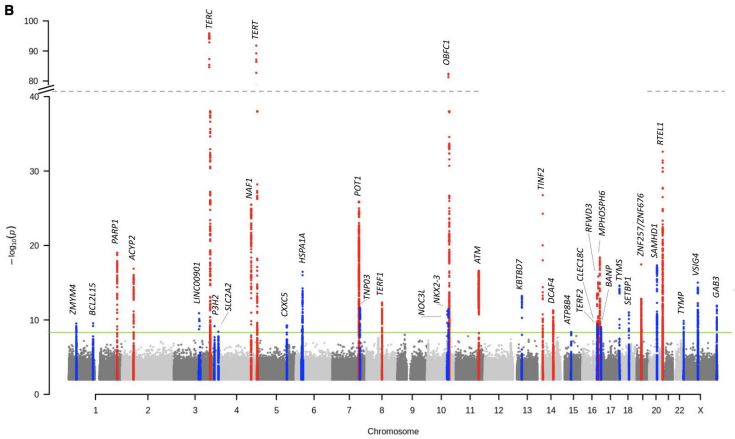
Summer Institute in Statistical Genetics (SISG)

- Three week institute held each summer at the University of Washington in Seattle: https://si.biostat.washington.edu/institutes/sisg
  - 28th SISG is happening now (July 10th - 28th, 2023)
- ~20 short course modules (2.5 days each) on specialized stat-gen topics
  - Focus: modern methods of statistical analysis and challenges posed by modern genetic data
- Instructors and attendees from across the world
- Goal: strengthen the statistical and genetic proficiency and career preparation of scholars from all backgrounds, especially those from groups historically underrepresented in STEM

SISG: Computational Pipeline for WGS Data

- Hands-on introduction to pipelines for whole genome sequence (WGS) data analysis
  - Genome Wide Association Studies (GWAS) and aggregate rare-variant association testing
  - Analysis tools written in R, with pipelines implemented in the cloud
  - Informed by instructors' experience in NIH-funded consortia (e.g. TOPMed project)

Example Manhattan Plot from GWAS



- 6th year offering this course, and 4th year using BioData Catalyst Powered by Seven Bridges platform for tutorials & exercises

Ken Rice
Univ. of Washington

Laura Raffield
UNC Chapel Hill

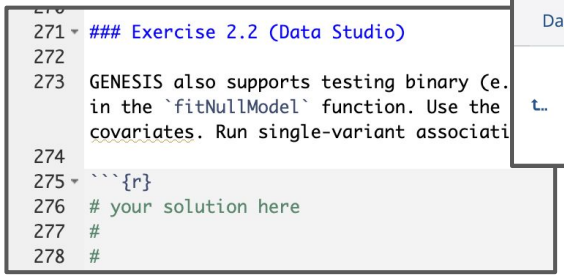
Matthew Conomos
Univ. of Washington

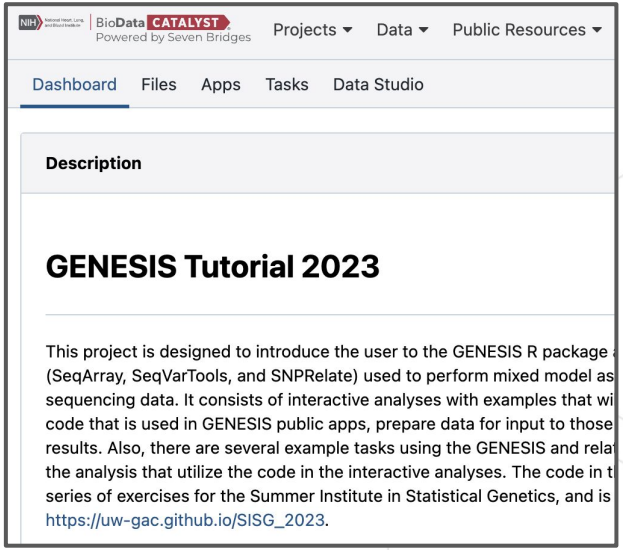
NIH National Heart, Lung, and Blood Institute | BioData CATALYST

Image: Taub, M. A., Conomos, M. P., Keener, R., Iyer, K. R., Weinstock, J. S., Yanek, L. R., ... & Mathias, R. A. (2022). Genetic determinants of telomere length from 109,122 ancestrally diverse whole-genome sequences in TOPMed. *Cell Genomics*, *2*(1).

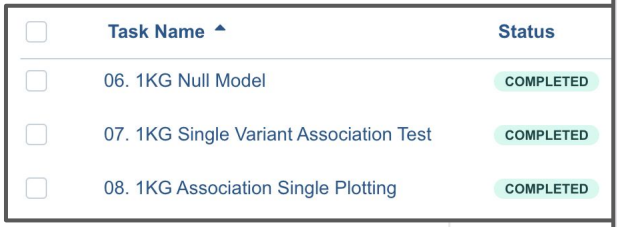
Course Structure and Materials

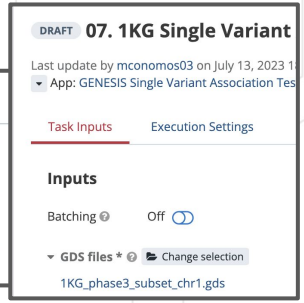
- Lectures paired with hands-on tutorials & exercises
- Course materials available as a “Public Project” on BioData Catalyst Powered by Seven Bridges
  - Students make a copy of the project
- Run interactive tutorials in Data Studio
  - Hands-on coding in R





- Run workflow application tasks
  - Learn to run analyses at scale in the cloud





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

Benefits of Using BioData Catalyst for SISG

- Easy to share analysis code, data, and tutorials with entire class
  - No need to worry about transferring large files
- Students only need a basic laptop with internet access
  - No concerns about laptop performance
- Students do not need to install specialized software on their laptops
  - Everyone has access to the same version; no compatibility issues
- Public project with course materials remains available on the platform
  - Students can revisit and re-do exercises after the course ends
  - Students can adapt tutorials for their own research
- Students get exposure to cloud computing
  - Opportunity to run realistic analyses at scale
- Supports virtual or in-person class
  - 2020-2022 the course was virtual; 2023 will be in-person

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

Breakout room

Discuss:

1. What topics do you normally teach?
2. How are student projects organized?
3. Are your courses remote?
4. Where do students struggle the most?

Feel free to ask TAs any questions.

→ Answer now on Jamboard

Breakout rooms - converge

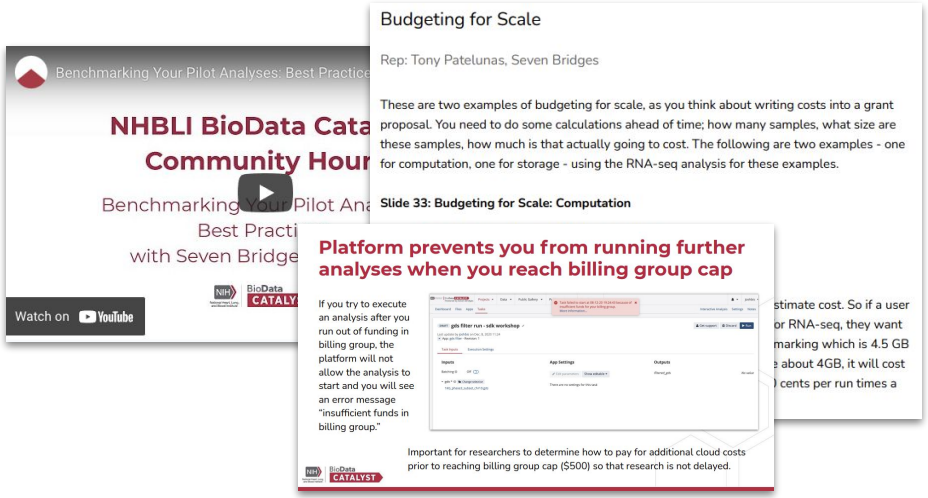
Report back

1. What topics do you normally teach?
2. How are student projects organized?
3. Are your courses remote?
4. Where do students struggle the most?

Final Closing

Community Hours

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



Topics of interest:

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

...and more!

Join us next week, July 19 at 1 pm ET

Linking Phenotypic Data to Genomic Data Files

Register: https://bit.ly/BDC-July

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

Learning Resources

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

Our Gitbook documentation includes:

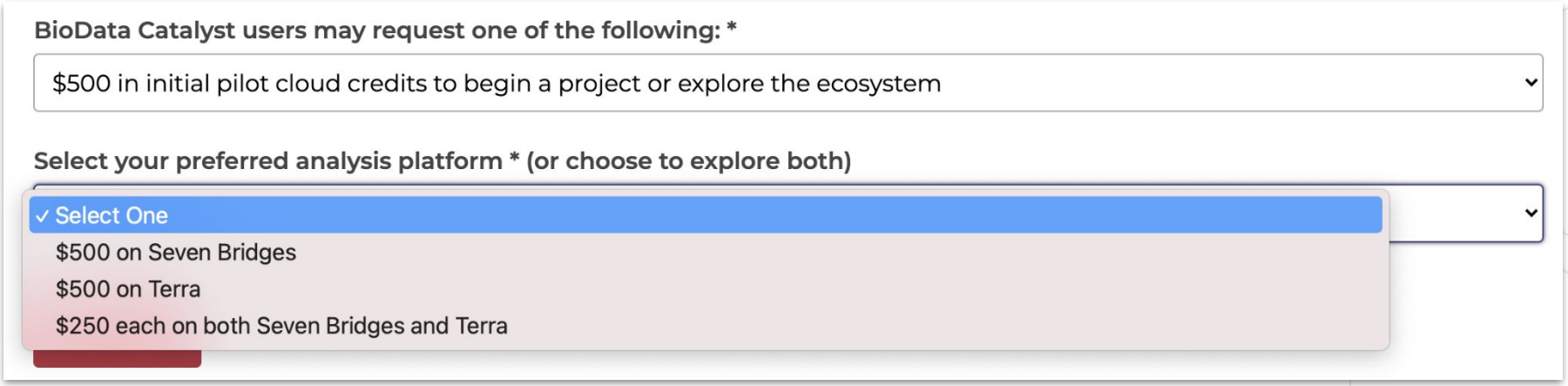
- Instructions on approvals and accounts needed to access BDC and how to check data access
- User Guides for PIC-SURE, Gen3, Seven Bridges, Terra, and Dockstore

Website resource: Learn

Documentation Resource: BioData Catalyst Documentation

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

Reminder: Try out BDC with Pilot Credits



Day 2 Close

Follow up with us at Office Hours

Drop in next Friday, July 21 between 11 am-noon ET for exclusive office hours for registrants of this workshop

https://bit.ly/BDC-NDSA-Office-Hours

Post-event Survey

https://www.surveymonkey.com/r/NDSAPostWorkshop