NHLBI BioData Catalyst: Focusing on Users

Rebecca Boyles
Co-PI, BioData Catalyst Coordinating Center
Senior Scientist, Research Computing, RTI

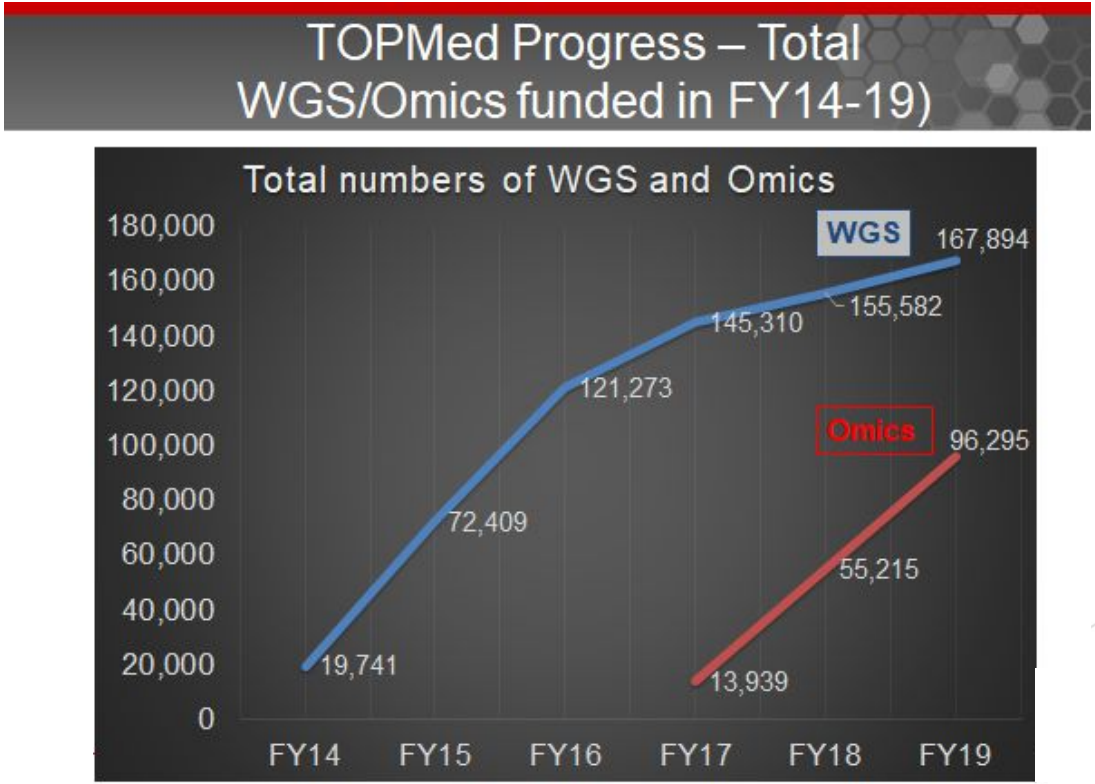
October 26, 2020

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

Agenda

- The Data Challenge
- BioData Catalyst Mission/Vision
- Ecosystem Components and Architecture
- BioData Catalyst Users and Fellows

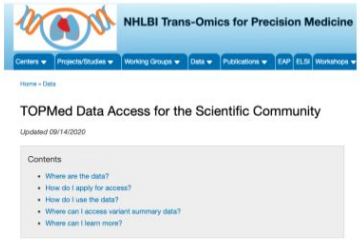
Explosion of Available Data



Limitations on Data Use

How to access to TOPMed data

- ➔ Available through **study-specific accessions**
  - ◆ phsXXXXXX
  - ◆ Phenotype data may be in TOPMed or pre-existing accessions
- ➔ Submit **dbGaP application** for access
- ➔ **Data Use Limitations** (DULs) vary by study
  - ◆ Some require local IRB approval (-IRB) or letter of collaboration (-COL)
- ➔ dbGaP applications reviewed by **NHLBI Data Access Committee** (DAC)

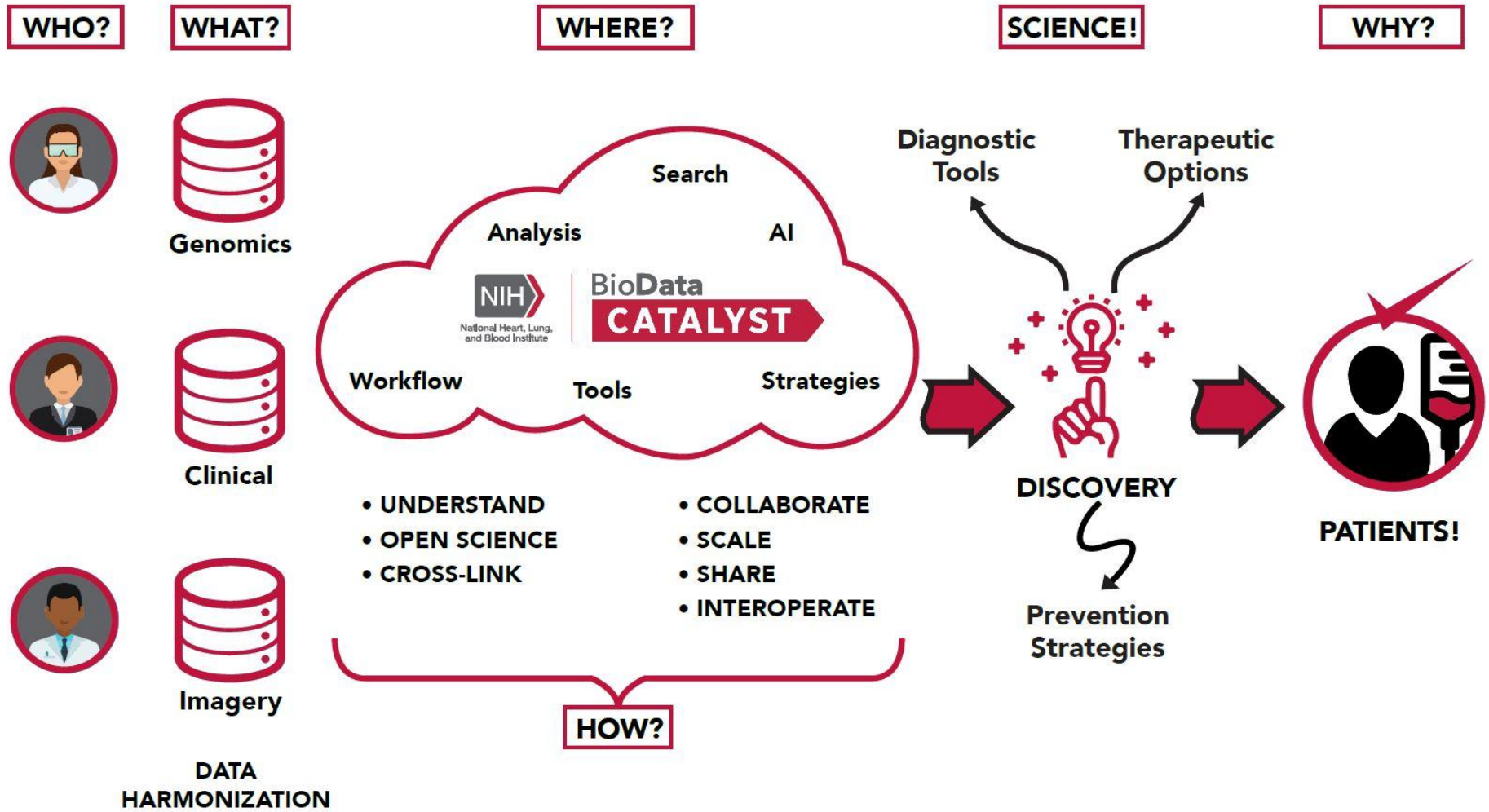


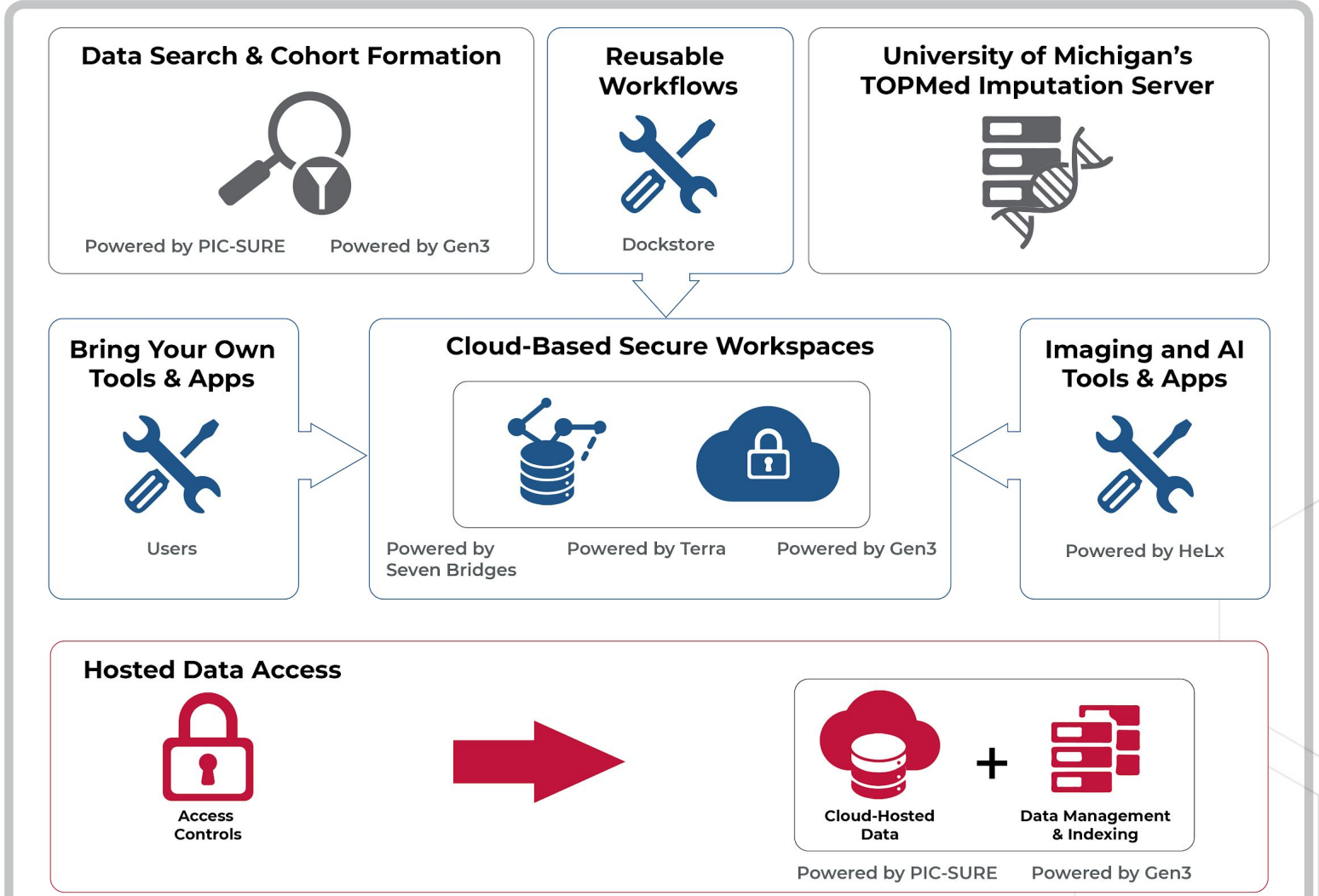
More info: https://www.nhlbiwgs.org/topmed-data-access-scientific-community

TOPMed Data Use Limitations

- Heterogeneity of DULs across TOPMed
- Compilation of diverse studies with unique histories, source populations, and informed consent processes
- Proposed research uses must align with DULs and participant consents
- Some studies require
  - documentation of local IRB approval (-IRB)
  - letter of collaboration (-COL)
- Look for “Data Use Certification (DUC) Agreement” on dbGaP study pages

Image: NHGRI Media Gallery


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


Data Search & Cohort Formation
Powered by PIC-SURE
Powered by Gen3
Reusable Workflows
Dockstore
University of Michigan's TOPMed Imputation Server
Bring Your Own Tools & Apps
Users
Cloud-Based Secure Workspaces
Powered by Seven Bridges
Powered by Terra
Powered by Gen3
Imaging and AI Tools & Apps
Powered by HeLx
Hosted Data Access
Access Controls
Cloud-Hosted Data
Data Management & Indexing
Powered by PIC-SURE
Powered by Gen3

Data Available in BioData Catalyst

The Trans-omics for Precision Medicine (**TOPMed**) initiative (https://www.nhlbiwgs.org/)

- Available now:
  - 45 TOPMed Freeze 8 studies (16 new studies to BioData Catalyst)
    - Genomic and Phenotypic Data
- Coming soon:
  - Additional TOPMed Freeze 8 studies
  - 1000 Genomes Project
  - BioLINCC Training Datasets
  - Pediatric Cardiac Genomics Consortium (PCGC) data
  - COVID data

For more detailed information, see About BioData Catalyst Datasets.

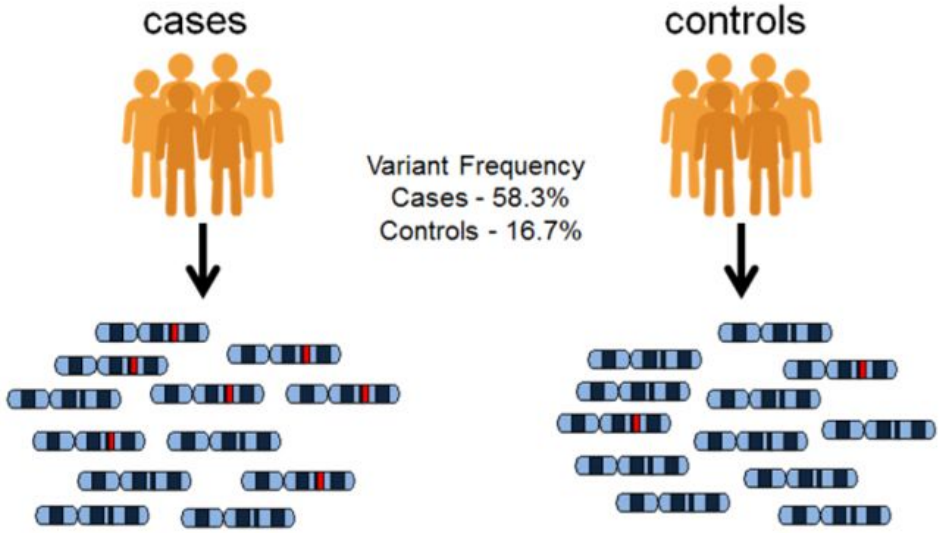
Bring-Your-Own Data

- 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

Ecosystem Tools

Key tools include

- **Genome-wide Association Study** (GWAS)
- **Genetic Variant Calling**
- **Structural Variant Calling**
- **Annotation Explorer**: for variant functional annotation
- **TOPMed Imputation Server:** Yields phased and imputed genotypes based on multiethnic reference panel



Reproducible Workflows

The BioData Catalyst ecosystem leverages Docker-based reproducible tools that can be discovered in Dockstore's open-access catalog and used in secure workspaces.

Two descriptor languages for Docker-based reproducible pipelines are supported:

- Common Workflow Language (CWL) supported on Seven Bridges
- Workflow Description Language (WDL) supported on Terra.

Dockstore hosts both languages, and you can launch tools and workflows directly from Dockstore into cloud workspace environments.

BioData Catalyst Cloud Credits

- The NHLBI currently provides **$500 in cloud credits** to new users of NHLBI BioData Catalyst via a billing group on either *BioData Catalyst Powered by Seven Bridges* or *BioData Catalyst Powered by Terra*.
- If the anticipated costs are in excess of $500:
  - Users can cover those costs using their own AWS and/or Google accounts which can be brought to BioData Catalyst
  - Users can apply for additional credits via the NHLBI BioData Catalyst Cloud Credit Program

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

The Fellows Program

For the researcher:

- Offers early career researchers **funding for novel and innovative research**

For BioData Catalyst:

- **Improves the the ecosystem** based on Fellow feedback.

Methodology

Sheila Gaynor, PhD

Functional Rare Variant Analysis

Stephanie Gogarten, PhD

Population Diversity

Caitlin McHugh, PhD

Multiple Phenotypes

Jean Monlong, PhD

Structural Variant Characterization

Jinling Liu, PhD

Identification of Causal Variants

Ravi Mathur, PhD

Matching Controls for Cases

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

Fellow: Sheila Gaynor, PhD

Genomic Variants

Functional Rare Variants

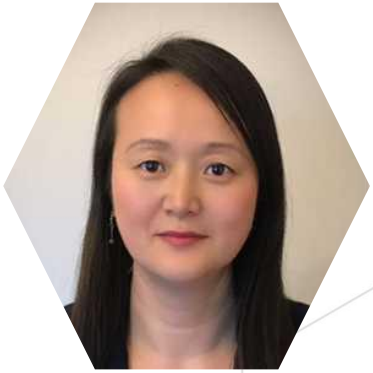
Scalable Statistical and Computational Methods for Integrating Functional Data in Rare Variant Analysis of Large Whole Genome Sequencing Data

Project goal: Implement a workflow to analyze rare variants incorporating functional annotations and apply the workflow to analyze TOPMed pulmonary function, glycemic, and inflammation traits.

How BioData Catalyst has helped:

- Developing **efficient, reproducible pipelines** for sharing for end-to-end analyses with **well-developed, open-source tools**
- Supporting **effective collaboration** to perform phenotype harmonization across institutions in TOPMed working groups and share results efficiently
- **Accelerating ability to run analyses**, return results, follow-up with collaborators

Fellow: Fayuan Wen, PhD



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

Blood/ Inflammation

Sickle Cell

Project goal: Identify genomic variation associated with iron overload in sickle cell disease (SCD) patients and to determine the pathway and functional relationship of iron overload related genes.

How BioData Catalyst has helped:

- Furnishes **secure data storage** space to store large dataset
- Facilitates **user-friendly Apps** for efficient genomics data analysis.
- Allows for close **collaboration with developers and bioinformaticians**.

Fellow: Kenneth Westerman, PhD

Methodology

Gene-diet Interaction in Glycemia

Identification and Characterization of Diet-Responsive Genetic Loci for Glycemic Traits

Project goal: Conduct gene-diet interaction analysis in TOPMed cohorts using a multi-exposure approach and characterize loci using metabolomics.

How BioData Catalyst has helped:

- Direct import of genotype files from 11 TOPMed cohorts using Gen3 **avoided a time-intensive and error-prone manual download from dbGaP.**
- The combination of Notebooks and Data Tables in Terra has made my **GWAS pipeline cleaner and more reproducible.**
- BioData Catalyst acts as a **secure and fully-featured space to share and harmonize phenotype datasets** within the TOPMed T2D Working Group.

Fellows Program Cohort 3 Deadlines

Fellows Applications Close	December 4, 2020
Award Notification	January 25, 2021
1st Period of Performance	March 2021-February 2022

More information at
https://biodatacatalyst.nhlbi.nih.gov/fellows/program

Please share with your community!

Fellows Application Criteria

A successful applicant

- Addresses a **scientific topic that can be answered** using BioData Catalyst
- Conducts analysis toward **publication**
- **Contributes to the functionality** of the ecosystem
- Obtains appropriate **data use agreements**
- **Demonstrates the necessary capabilities** to accomplish the proposed work
- **Engages and collaborates** with the BioData Catalyst community
- Contributes to **diversity** across fields of study, institutions, geography, and investigators.

Heart

Alexander Bick, MD, PhD

CHIP expansion and CVD

Melissa Cline, PhD

Genetics of Cardiomyopathies

Jacqueline Dron, PhD

Genetics CAD Lifecourse

Einat Granot-Hershkovitz, PhD

Ancestry-enriched Variants and CVD

Jamie Murkey, MPH

Psychosocial stress and CVD

Yaling Tang, MD

Transcriptomics in Heart Failure

Xuefang Zhao, PhD

Structural Variants and Lipids

Fellows Collaborative Projects

- It is possible to propose a project involving a pair of investigators with complementary skills to collaborate on a project.
- For example: clinical investigator with knowledge of disease collaborating with a statistical investigator. Both must contribute substantially to the project.
- It is not permissible to fund a senior investigator as a mentor using this mechanism.

Fellows at ASHG

Harriet Dashnow

Poster Session, Bioinformatics and Computational Approaches, 10/26
STRling: a k-mer approach for detecting short tandem repeat expansions at both known and novel loci from short-read sequencing data.

Caitlin McHugh

Poster Session, Statistical Genetics and Genetic Epidemiology, 10/26
A mixed model approach to testing multiple correlated traits in large samples: An application to the Trans-Omics for Precision Medicine (TOPMed) program hematology phenotypes.

Ravi Mathur

Poster Session, Statistical Genetics and Genetic Epidemiology, 10/26
A generalizable protocol for leveraging whole genome sequenced cohorts as population controls for new genome wide association studies.

Randi Johnson

Poster Session, Molecular Phenotyping and Omics Technologies, 10/26
Discovering metabolite quantitative trait loci in asthma using a genetically isolated founder population

Platform Talk, Novel Statistical Genetics Methods for Complex Traits, 10/28
Powerful and resource-efficient rare variant meta-analysis for large-scale whole genome sequencing studies using summary statistics and functional annotations

Zilin Li

Poster Session, Statistical Genetics and Genetic Epidemiology, 10/26
MultiSTAAR: Powerful rare variant multi-trait analysis incorporating functional annotations for large-scale whole genome sequencing studies, with application to TOPMed lipid data.

Michelle Daya

Poster Session, Complex Traits and Polygenic Disorders, 10/26
An HLA association study of total serum IgE levels using whole-genome sequence data from TOPMed

Other Ways to Access the BioData Catalyst Ecosystem

- Currently, the ecosystem is by invitation only, but we are working toward opening our doors in 2021.
- We will be rolling out another round of invitations soon. Contact us to request an invitation and reference this event.
- Each researcher receives orientation materials and cloud credits to explore resources, pilot runs, and initiate their project.

Why BioData Catalyst

- Gives users controlled access to TOPMed
- Enables team collaboration working with data
- Allows users to bring data, tools, and workflows **to** the data
- Democratizes access to data and compute resources
- Provides support, tutorials, and documentation to help users navigate the system
- Provides $500 in cloud credits to new users of the system
- Will to continue to develop according to user community needs!

Learn More on the BioData Catalyst Website

biodatacatalyst.nhlbi.nih.gov

Main access point for all things BioData Catalyst

BioData Catalyst Data: Description of data availability and access limitations.

BioData Catalyst Services: Overview and links to ecosystem, platforms, services and workflows.

BioData Catalyst Learn: Access tutorials and other documentation.

BioData Catalyst Documentation: Documentation server.

Biodata Catalyst Help Desk and Knowledge Base: Help desk environment featuring searchable knowledge base, FAQs, and contact forms.

National Heart, Lung, and Blood Institute

Providing strategic leadership and funding the researchers and other professionals developing the ecosystem.

Director: Gary Gibbons
CIO: Alastair Thomson
Program Officer: Jon Kaltman

Steering Committee

Providing strategic decision-making and achieving consensus for the Consortium.

Ingrid Borecki (Chair), Principal Investigators, NHLBI Working Group

External Expert Panel

Independently informing and advising the work of the Consortium.

Donna Arnett
Mark Craven
Jason Williams
David Mendelson
Warren Kibbe

Coordinating Center

Coordinating project management, communications, project reporting, and collaboration standards.

Ahalt, Boyles

Data Stewards

Partnering with the Consortium on data accessibility and interoperability.

TOPMed, COPDGene

The Broad Institute, University of Chicago, University of California, Santa Cruz

Grossman, Manning, Paten, Philippakis

Providing authorized access and faceted search of harmonized data across studies, genomic analysis and visualization in virtual workspace, and high-quality Docker-based research tools.

Harvard Medical School

Avillach

Exploring data with interactive search and visualizations for feasibility assessment and providing data science tools to access and analyze clinical and genomic data.

RTI International, UNC-CH/RENCI

Krishnamurthy, Bradford

Developing tools and apps for machine learning; deep learning models; semantic search; and visualizing, annotating and analyzing biomedical images. Developing methods for tool and app creation to enhance the ecosystem.

Seven Bridges Genomics

Davis-Dusenbery

Finding, accessing, analyzing TOPMed genomics data at scale; bringing your workflows or choosing from hosted CWL tools; performing association studies with tooling for variant aggregation.

NIH National Heart, Lung, and Blood Institute
BioData CATALYST

Welcome!

TOPMed Ancillary Session
October 26, 2020
11am-12:30 pm ET

ASHG VIRTUAL MEETING 2020 OCTOBER 27-30

This session will be recorded.

Schedule

11:00	Program overview NHLBI TOPMed Program	Weiniu Gan
11:15	Data overview & access TOPMed Data Coordinating Center	Sarah Nelson
11:30	Genomic variation & imputation server TOPMed Informatics Research Center	Albert Smith
11:45	NHLBI BioData Catalyst Focusing on Users	Rebecca Boyles
	Audience Q&A	

www.nhlbiwgs.org/ashg-2020-ancillary-session