Introduction to NHLBI BioData Catalyst

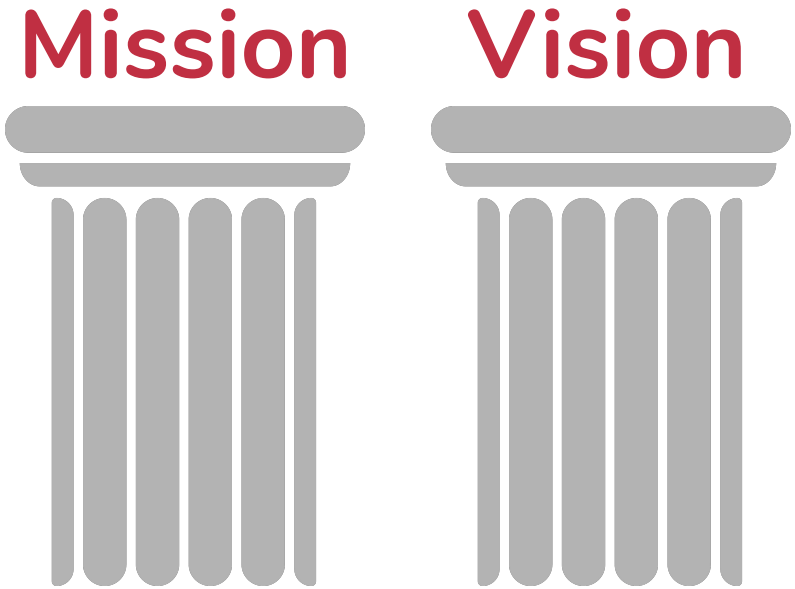
Ingrid Borecki, Ph.D.
BioData Catalyst Steering Committee Chair
October 21, 2020

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

Agenda

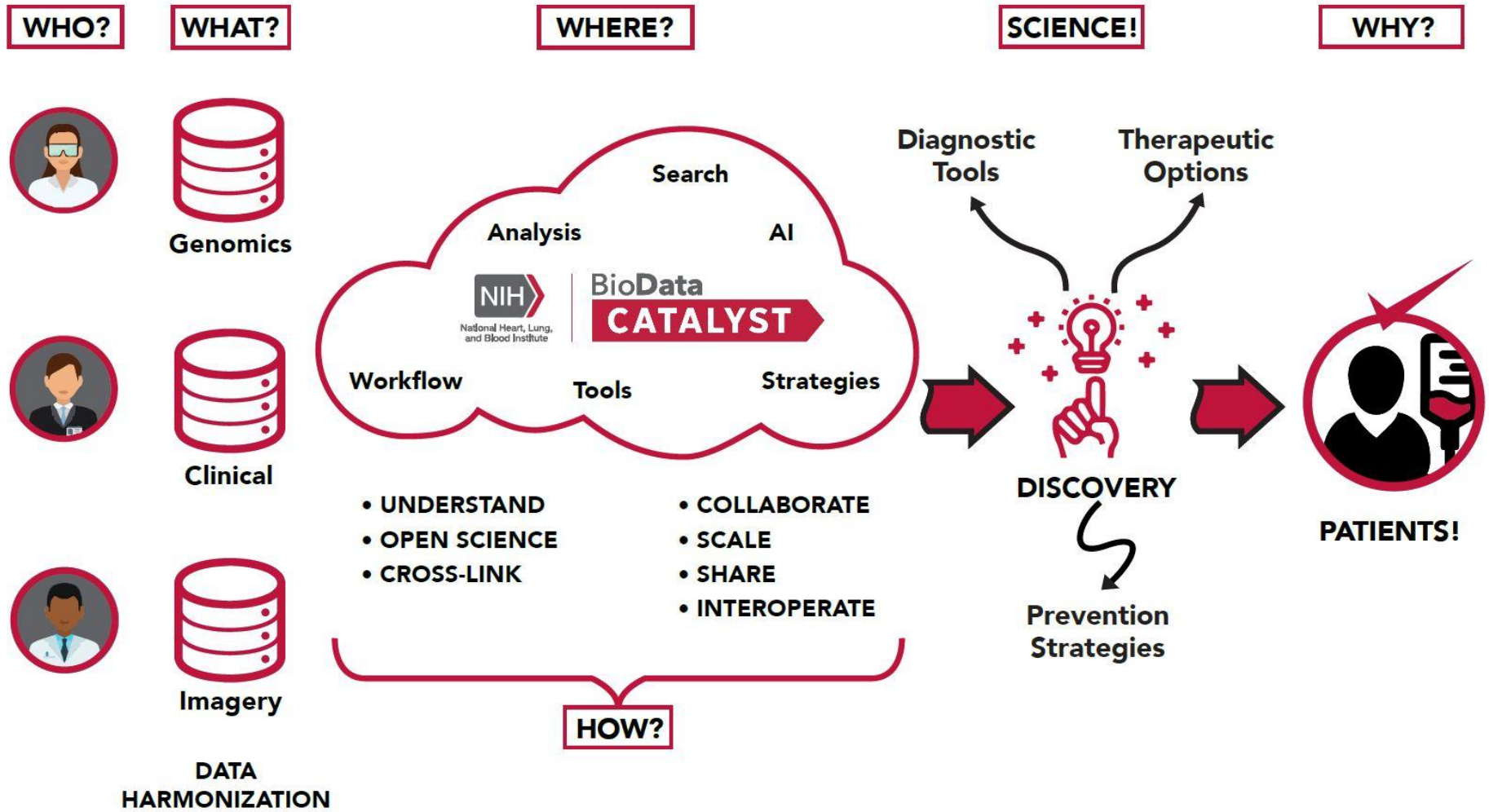
- NHLBI BioData Catalyst Mission
- Ecosystem Data, Tools, and Workflows
- Potential Users
- BioData Catalyst Fellows Program

NHLBI BioData Catalyst



The *mission* is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.

The *vision* is to be a community-driven ecosystem implementing data science solutions to democratize data and computational access to advance Heart, Lung, Blood, and Sleep science.


WHO?
WHAT?
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 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.

Types of Data

Phenotypic

Harmonized data

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

Non-harmonized data

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

Genomic

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

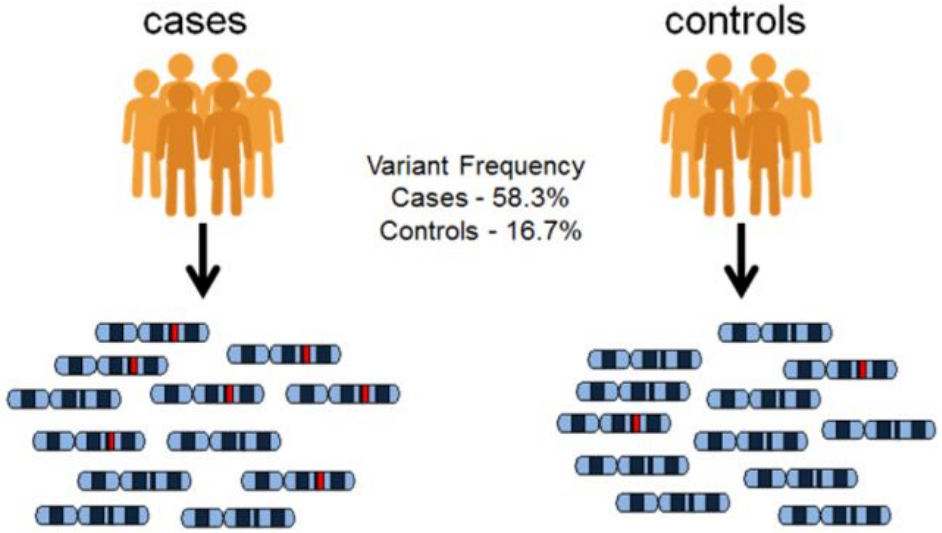
Bring-Your-Own Data

- 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.

Who are our potential users?

Dr. Phil Phelps is a **physician scientist** looking at the interaction of dietary fat with known genes in the lipid metabolism pathway and the effect on measured LDL-cholesterol levels.

Stevie Statler is a **statistician** looking to test multiple correlated traits for genetic association, leveraging the large-scale data in the BioData Catalyst Consortium.

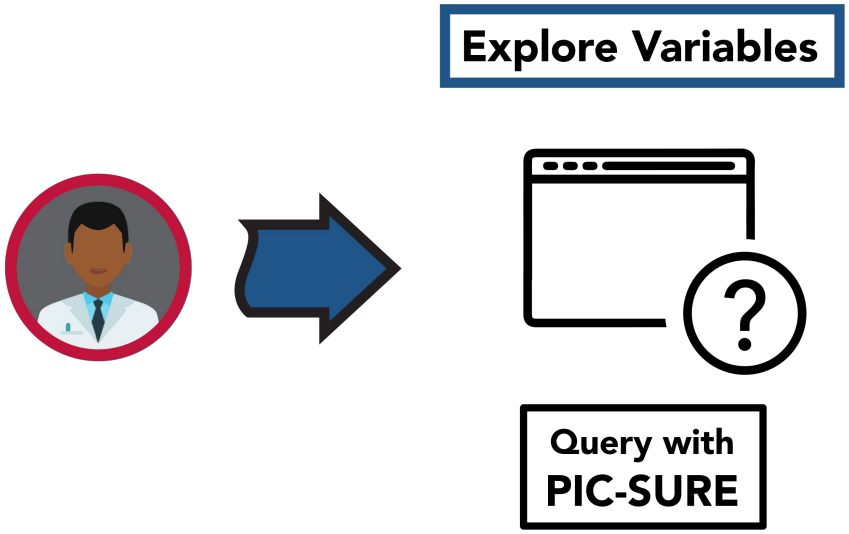
Dr. Imogen Imhoff is a **researcher** interested in explaining the genetic association to subtypes of COPD (emphysema, chronic bronchitis, and chronic obstructive asthma) characterized by image data.

Clinical Genetic Epidemiology

Dr. Phelps wants to ask:

Does intake of dietary fat interact with genes in the lipid metabolism pathway and affect measured LDL-cholesterol levels?

Clinical Genetic Epidemiology

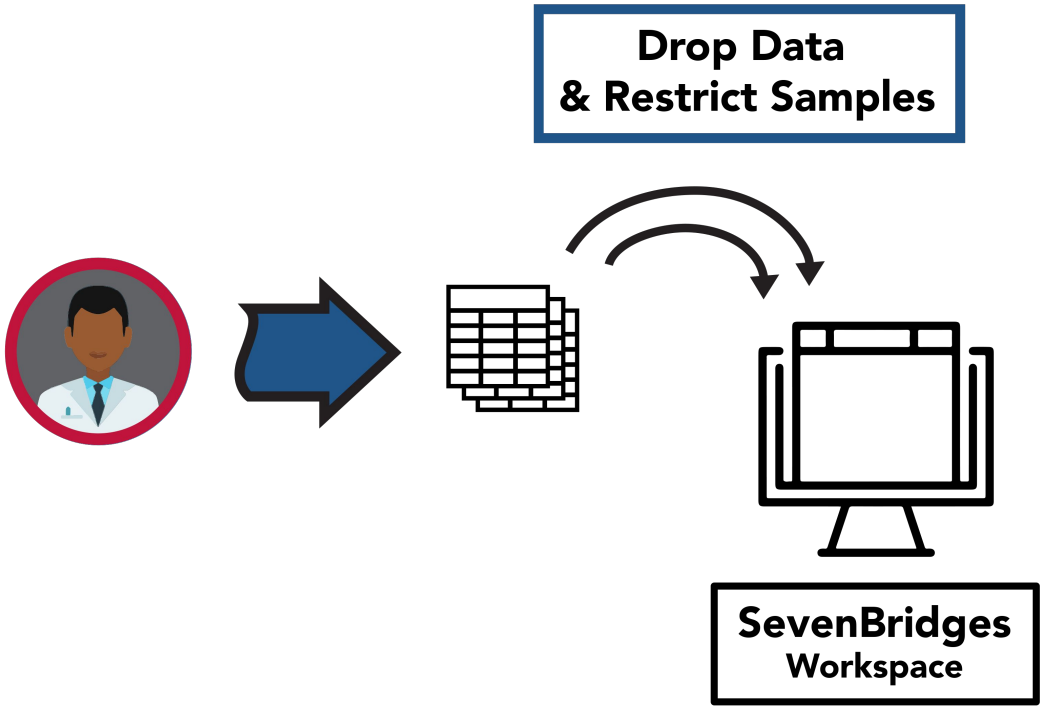


Dr. Phelps uses PIC-SURE to find all studies with lipids measured. Dr. Phelps identifies studies that also have dietary measures and asks:

What types of variables are there?

Are there measures of macronutrient intake, specifically, fat intake?

Clinical Genetic Epidemiology

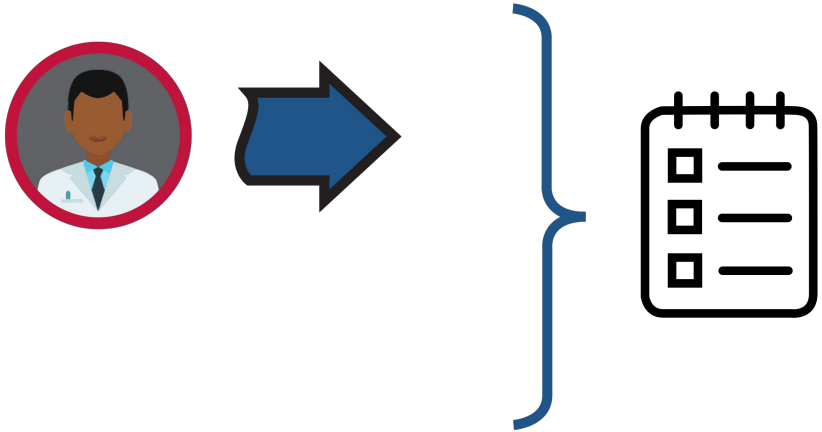


Dr. Phelps drops data into a secure workspace on Seven Bridges or Terra and restricts samples by ancestry using PIC-SURE.

What are the sample sizes?

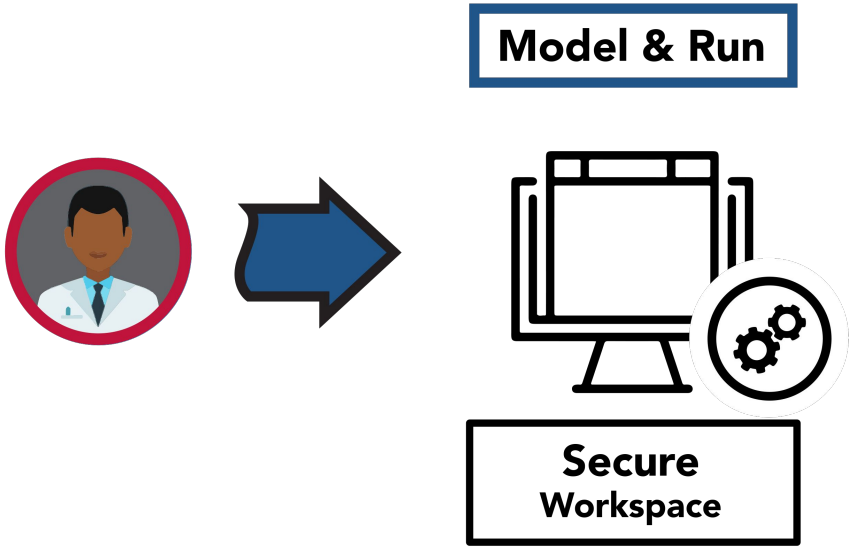
Clinical Genetic Epidemiology

Prepare List of Biomarkers



Dr. Phelps prepares a list of harmonized phenotypes and dietary measures for analysis

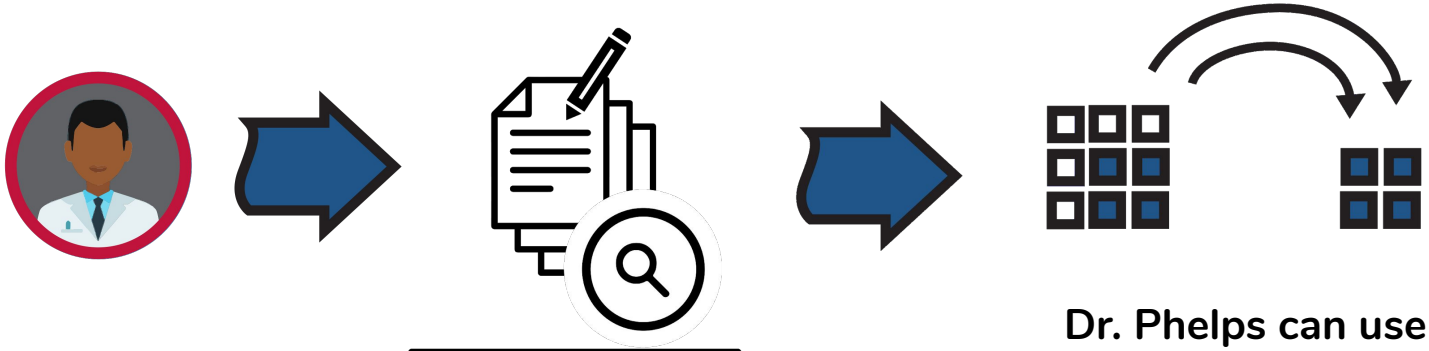
Clinical Genetic Epidemiology



Dr. Phelps searches for workflows to conduct G x E interaction test, sets up models, and launches run.

Clinical Genetic Epidemiology

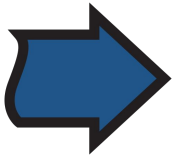
Annotate & Subset

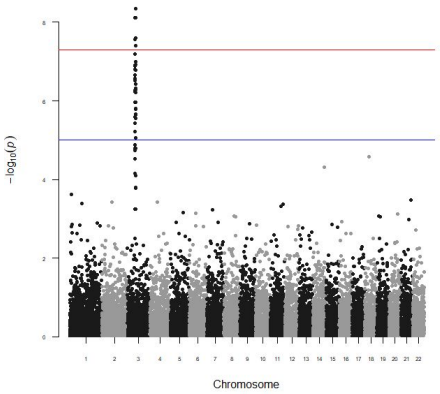


Dr. Phelps can use Seven Bridges Annotation Explorer to annotate significant variants by gene and function, and subsets results by genes in lipid metabolism pathways.

Clinical Genetic Epidemiology

Display Results





Dr. Phelps displays results in a Manhattan plot.

Dr. Phelps can now perform follow-up analyses.

The Fellows Program

For the researcher:

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

For BDCatalyst:

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

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 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.

Lung

Sarah Gerard, PhD
Lung Image AI

Pietro Nardelli, PhD
COPD Phenotyping

Dandi Qiao, PhD
COPD GWAS

Einat Granot-Hershkovitz, PhD
Transcriptomics in Asthma

Pranav Rajpurkar, MS
ML Chest X Rays

Jia Wen, PhD
Genetic Modifiers Cystic Fibrosis

Yonghua Zhuang, PhD
Multiomics in COPD

Fellow: Kenneth Westerman, PhD

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

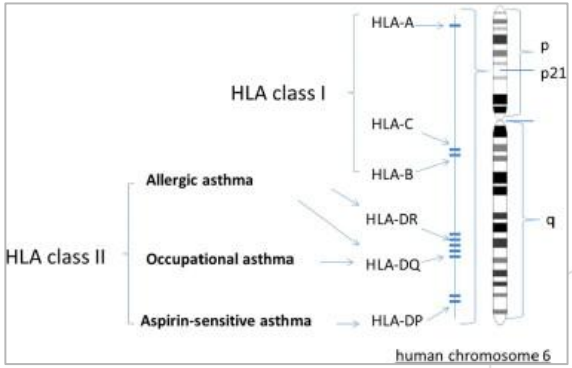
Project goal: Conduct gene-diet interaction analysis in TOPMed cohorts using a multi-exposure approach and characterize loci using metabolomics.

How BioData Catalyst has helped:

- Direct use of genotype files from 11 TOPMed cohorts in 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.

Fellow: Michelle Daya, PhD

HLA and Genome-Wide Association Studies of Total Serum IgE (tIgE) Levels



Project goal: Use state-of-the-art tools to call HLA alleles and test for association with tIgE in subjects of diverse ancestry.

How BioData Catalyst has helped

- **Cloud access to CRAM** files for HLA typing saved time (no need to download) and cost (no need to store)
- **Compute availability** that is not limited to local cluster
- **TOPMed imputation** on BioData Catalyst (6x faster than a similar project 4 years ago)
- Platform encourages development of **well-designed, well-documented, source-controlled, and re-usable pipelines**

Fellows Program Cohort 3 Deadlines

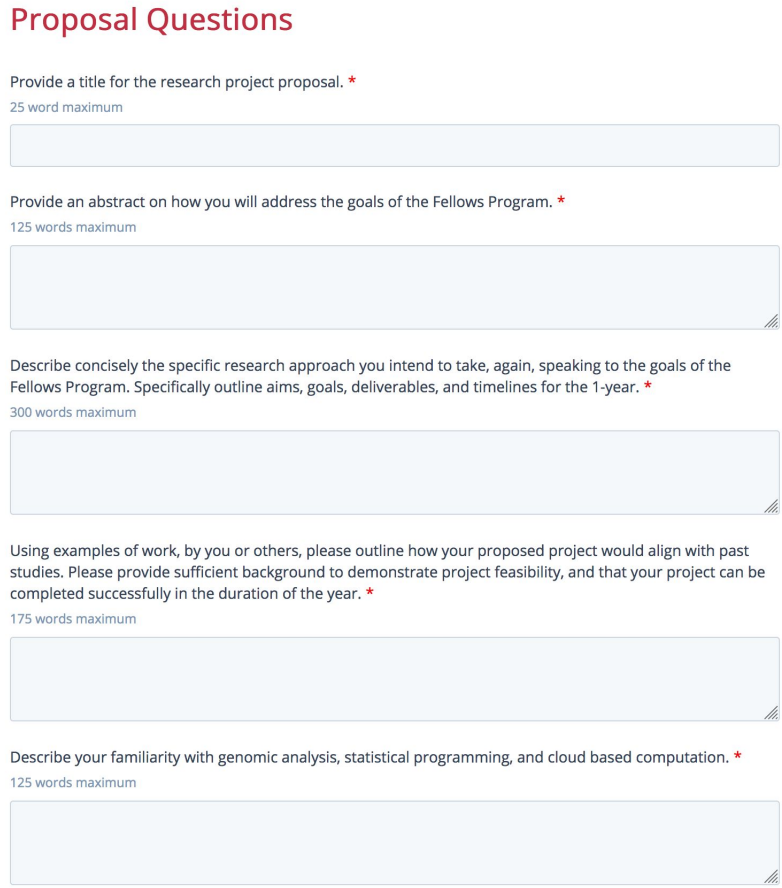
Fellows Applications Open	Oct 16 2020
Fellows Applications Close	Dec 04 2020
Award Notification	Jan 25 2021
1st Period of Performance	Mar 2021 - Feb 2022

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

Please share with your community!

Fellows Application Elements

- **Background:** Profile information, Letters of support, Biographical sketch
- **Abstract:** description of the research
- **Project aims:** specific aims, goals, deliverables, and timelines
- **Prior studies:** how proposed project would extend prior work.
- **Expertise:** genomic analysis, statistical programming, & cloud based computation.
- **Methods and materials**
- **Engagement** with Consortium members, other Fellows, broader community
- **Budget:** costs for salary, travel, training, publication, & conference fees up to ~$69k



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.

Other Ways to Try Out the BDCatalyst Ecosystem

- Currently, the ecosystem is by invitation only, but we are working toward opening our doors in 2021.
- PRIDE members can request access at https://biodatacatalyst.nhlbi.nih.gov/contact. Make sure to include your association with PRIDE in your message.
- Each researcher receives cloud credits to explore resources, pilot runs, and initiate their project.

Learn More on the BioData Catalyst Website

BioData Catalyst Website (main page): 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.

Questions?