

Breakout Session 3: Track A

Building a Cross-study Data Set for the PRIMED Consortium

Dr. Ben Heavner

Senior Research Scientist, University of Washington

Building a cross-study data set for the PRIMED Consortium



PRIMED

Polygenic Risk
Methods in
Diverse
Populations

2024 NIH/ODSS Cloud
Supplement Program
PI Meeting
January 17-18, 2024

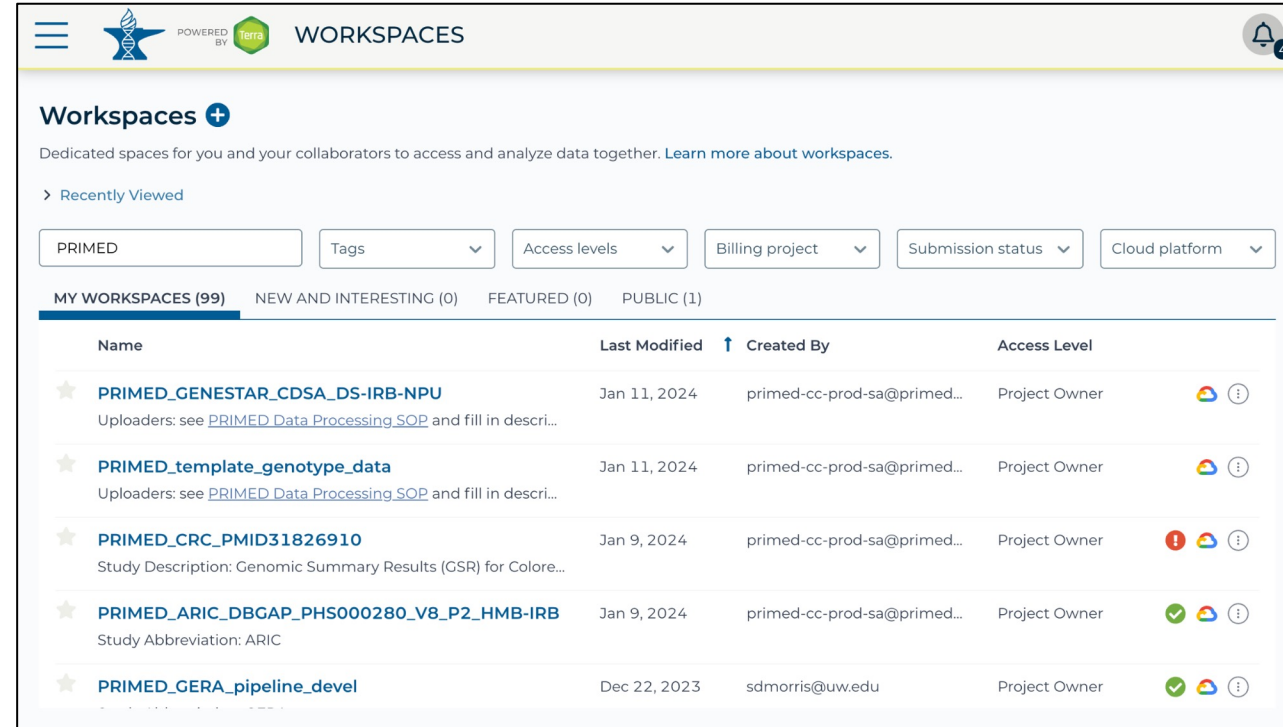
@PRSDiversity



The PRIMED Consortium gratefully acknowledges the support provided by the NIH/ODSS Cloud Supplement Program.

This support has enabled Consortium data storage on NHGRI's AnVIL.

Through this support, we have gained valuable experience with cloud-based collaborative work. The PRIMED Consortium is well positioned to continue to further advance this work.



The screenshot shows the 'Workspaces' page in the AnVIL interface. It features a header with the AnVIL logo, 'POWERED BY terra', and 'WORKSPACES'. Below the header, there's a section for 'Workspaces' with a plus icon and a link to 'Learn more about workspaces'. A 'Recently Viewed' section is followed by several filter buttons: 'PRIMED', 'Tags', 'Access levels', 'Billing project', 'Submission status', and 'Cloud platform'. Below these filters, there are tabs for 'MY WORKSPACES (99)', 'NEW AND INTERESTING (0)', 'FEATURED (0)', and 'PUBLIC (1)'. The main content is a table with columns for 'Name', 'Last Modified', 'Created By', and 'Access Level'. The table lists five workspace entries, each with a star icon, a name, a description, a last modified date, a creator email, an access level, and a set of icons (Google Cloud, info, and status).

Name	Last Modified	Created By	Access Level
★ PRIMED_GENESTAR_CDSA_DS-IRB-NPU Uploaders: see PRIMED Data Processing SOP and fill in descri...	Jan 11, 2024	primed-cc-prod-sa@primed...	Project Owner
★ PRIMED_template_genotype_data Uploaders: see PRIMED Data Processing SOP and fill in descri...	Jan 11, 2024	primed-cc-prod-sa@primed...	Project Owner
★ PRIMED_CRC_PMID31826910 Study Description: Genomic Summary Results (GSR) for Colore...	Jan 9, 2024	primed-cc-prod-sa@primed...	Project Owner
★ PRIMED_ARIC_DBGAP_PHS000280_V8_P2_HMB-IRB Study Abbreviation: ARIC	Jan 9, 2024	primed-cc-prod-sa@primed...	Project Owner
★ PRIMED_GERA_pipeline_devel	Dec 22, 2023	sdmorris@uw.edu	Project Owner



About PRIMED

The NIH-funded **P**olygenic **R**isk **M**ethods in **D**iverse Populations (PRIMED) Consortium is developing and evaluating methods to improve the use of [polygenic risk scores](#) (PRS) to predict disease and health outcomes in diverse ancestry populations.



1. Gather Diverse Datasets

Bring together large datasets with genomic and health measures from diverse ancestry populations



2. Develop New Methods

Develop new methods to improve genetic risk prediction across diverse populations for a broad range of health and disease outcomes



3. Enable Collaboration

Enable collaborative analysis by sharing PRS-related data, software, and other resources with the scientific community



4. Improve Health

Leverage existing precision medicine partner programs to develop, test, and refine PRS in diverse populations to improve health outcomes



PRIMED's use of AnVIL: Overview



- Data Storage and Sharing
 - Centralized data storage: no need to store (and pay for) multiple copies of large data files
 - Shared, harmonized data available to all PRIMED members (with appropriate approved data access)
 - **Note:** PRIMED researchers are generally secondary users of released data
- Workflow Development
 - Collaborative development, implementation, and testing of software tools/workflows
 - Share code via Dockstore and GitHub – usable by PRIMED members and broader community
- Interactive Analysis
 - RStudio and Jupyter
 - Easily write and share analysis code in R or Python
 - The PRIMED Coordinating Center provides [example notebooks](#) illustrating how to work with PRIMED data in AnVIL
 - available in example AnVIL workspaces shared with PRIMED members



PRIMED AnVIL Workspace Organization

Shared Data Storage Workspaces

- CC sets-up and pays for **shared workspaces for data storage**
 - Organized by study, consent, and data access mechanism
- Study Sites upload study data and format it to PRIMED data model
 - Data uploaders run data validation workflows maintained by CC
- CC manages Consortium member access
 - Based on approved DARs; aligned with NIH & Consortium policy

dbGaP Study Data Workspaces

Access Mechanism: PRIMED Coordinated dbGaP Application

phs000001.v1.p1.c1

phs000001.v1.p1.c2

phs000002.v3.p2.c1

Consortium Data Sharing Agreement (CDSA) Workspaces

Access Mechanism: Consortium Data Sharing Agreement

Study A Consent 1

Study A Consent 2

Study B Consent 1

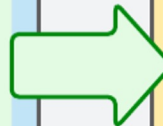
Open Access Data Workspaces

1000 Genomes Data

Open Access GSR

Simulated Data

- **All Data in Shared Data Storage Workspaces are accessible in Analysis & Dev Workspaces** (with proper data access approvals)
 - Does *not* require copying large molecular data files (use pointers)



Analysis Workspaces

- Study Sites set-up, manage access to, and pay for their **own workspaces to perform analyses (compute)**
 - Also used to store outputs/results

Study Site Analysis Workspaces

CAPE

CARDINAL

D-PRISM

EPIC-PRS

FFAIRR-PRS

PREVENT

PRIMED-Cancer

- CC sets-up and pays for workspaces for Consortium-wide analyses in benefit of the entire Consortium

Consortium-Wide Analysis Workspaces

Collaborative Project 1

Collaborative Project 2

Collaborative Project 3

Development Workspaces

- CC sets-up and pays for **workspaces for workflow development and testing** upon request

Workflow Dev 1

Workflow Dev 2

Workflow Dev 3

PRIMED dbGaP data sharing circle

Applicants with approved DARs via PRIMED Coordinated dbGaP Applications

PRIMED Steering Committee

PRIMED-SAG data sharing circle

CDSA Signatories

SAG = Sharing Agreement Group

NIH DACs

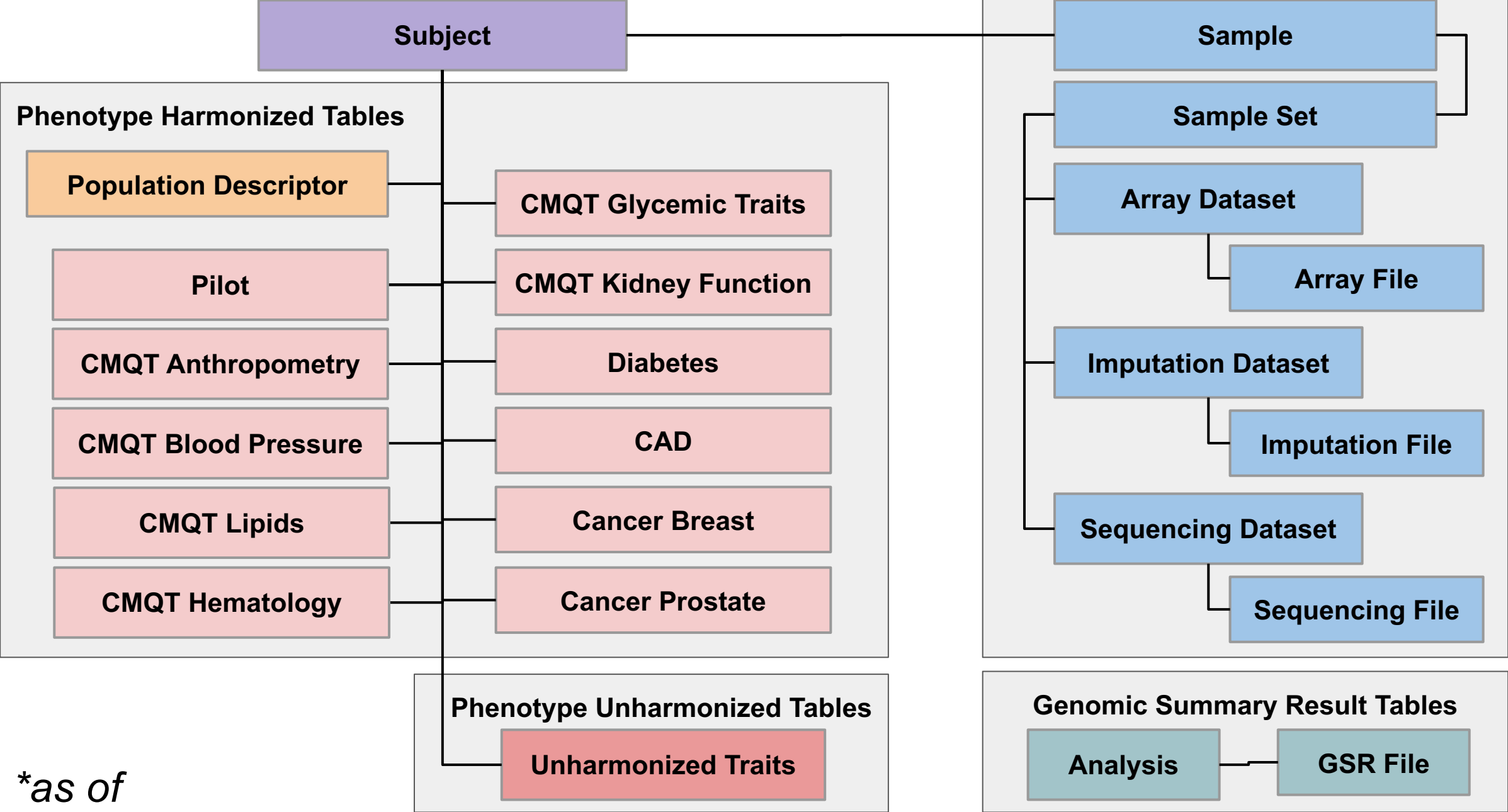
Additional Coordinated Access data sharing circle(s)

Site-Specific data sharing circle(s)

Creation and management specified in PRIMED Data Sharing Policy



PRIMED Data Model Tables



**as of*

Summary of Study Data Availability for PRIMED in AnVIL

Access Mechanism	Study	Phenotype Data													Genotype Data						
		Population Desc.	Breast Cancer	Colorectal Cancer	Lung Cancer	Prostate Cancer	Anthropometry	Blood Pressure	Glycemic	Hematology	Kidney Function	Lipids	CAD	Diabetes	Unharmonized	Array Genotypes	TOPMed WGS	TOPMed Imputed			
dbGaP	ARIC	S					S	S	S	S	S	S	S	S			S	S			
	CARDIA	A					A	A	A		A	A		A			S				
	HCHS/SOL	A					A	A	A		A	A		A			S				
	JHS	P					P	P	P	P			P				S				
	RPGEH (GERA)	S												S				S			
	WHI	A	A					A	A	A		A	A		A		P	S	P		
CDSA	GeneSTAR	A						A	A		A	A					P				
		Genomic Summary Results																			
Open	GWAS Catalog	S		S	S												S: Shared P: In Preparation A: Anticipated				
	QBB	S											S								
	UKBB	S	S	S		S	S	S	S	S	S	S	S	S							

Shared:
data uploaded;
passed PRIMED data
model validation;
shared with PRIMED
members with
approved access

In Preparation:
data uploaded;
contributors working
on harmonizing and
formatting to PRIMED
data model

Anticipated:
workspace created;
contributors have not
created data tables



Additional Data Available in AnVIL

- Publicly Available reference WGS data from 1000 Genomes that conforms to the PRIMED data model
 - Joint call sets in multiple file formats: VCF, PLINK
 - Data and metadata formatted to PRIMED data model
 - Used for genetic ancestry inference, data simulation, testing methods and workflows
 - Available at https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_1000G_PRIMED-data-model
- Simulated individual-level and summary data
 - Provided by the Simulation & Benchmarking sub-WG (esp. Haoyu Zhang)
 - Genotypes based off of 1000 Genomes “super populations”
 - 120K samples each
 - Simulated phenotypes and association summary statistics
 - Used for testing methods and workflows
 - Workflows available to generate additional simulated data:
<https://dockstore.org/organizations/PRIMED/collections/simulation>



Workflow Development: Current Status



- Released Workflows (available in [PRIMED Dockstore](#))
 - **Data validation** (n = 6): data and metadata checks against PRIMED data model
 - **Data import** (n = 2): fetch data from external sources (e.g. dbGaP) and upload into AnVIL
 - **Genotype file conversion** (n = 4): convert genotype data among popular formats
 - **Genetic ancestry inference** (n = 2): calculate genetic ancestry PC SNP loadings and project samples into genetic ancestry PC space
 - **Genotype imputation** (n = 2): submit and retrieve imputation jobs using TOPMed server
 - **Simulation** (n = 3): simulate realistic genotypes and phenotypes from admixed populations
- In Development
 - **Genetic ancestry inference**: global and local genetic ancestry inference methods
 - **PRS calculation**: compute PRS values from individual geno data using provided PRS models
 - **PRSmix**: method for developing PRS models

**CC provides AnVIL workspaces to Consortium members for workflow development/testing upon request*



Lessons Learned (1/2)

The growth rate of cloud-hosted data in the PRIMED Consortium has been slower than we anticipated for at least two reasons: **Scale** and **Incentive**

- 1) **Scale:** PRIMED Sites have proposed using data from ~70 studies. These studies have **very heterogeneous data and access control requirements**
 - To enable data contributions and sharing, PRIMED has developed new policy frameworks and specific Consortium Data Sharing policies
 - Obtaining access approval remains cumbersome and tedious at this scale
 - We have developed a new web-based tool to manage the complexity of workspaces and user/group permissions on AnVIL



Lessons Learned (2/2)

The growth rate of cloud-hosted data in the PRIMED Consortium has been slower than we anticipated for at least two reasons: **Scale** and **Incentive**

- 2) **Incentive:** Incentives for gathering data in a Cloud Environment may not be obvious to PRIMED Researchers yet – they have on-prem computing resources and can use them to analyze data they can download. Concerns about a steep learning curve and the risk of high cloud costs may currently outweigh perceived benefits
 - Looking forward, we anticipate some PRIMED data that will be exclusively available in AnVIL
 - As analysis tools/pipelines become available in AnVIL, we expect the incentive to become more apparent
 - The Consortium is considering a strategy of enabling distributed analysis that is collectively planned. This approach will enable collaborative research executed in either on cloud-based or on-prem computing resources.



Acknowledgements

Study Sites

Center for Admixed Populations and Health Equity (CAPE)

Bogdan Pasaniuc (University of California Los Angeles)

Eimear Kenny (Mount Sinai)

Leslie Lange (University of Colorado)

[U01HG011715](#)

Diabetes Polygenic Risk Scores in Multiple ancestries (D-PRISM)

Josep Mercader (Broad Institute)

Alisa Manning (Broad Institute)

Maggie Ng (Vanderbilt University Medical Center)

[U01HG011723](#)

Functional and Fine-Mapping Approach to Improve Responsible Risk-modeling of Polygenic Risk Scores (FFAIRR-PRS)

Pradeep Natarajan (Broad Institute)

[U01HG011719](#)

Leveraging Diversity in Cancer Epidemiology Cohorts and Novel Methods to Improve Polygenic Risk Scores (PRIMED-Cancer)

David Conti (University of Southern California)

John Witte (Stanford University)

[U01CA261339](#)

CARDiometabolic Disorders IN African-ancestry popuLations (CARDINAL)

Sally Adebamowo (University of Maryland Baltimore)

Bamidele Tayo (Loyola University of Chicago)

[U01HG011717](#)

EndoPhenotype InCorporated PRS (EPIC-PRS)

Yun Li (University of North Carolina at Chapel Hill)

Nancy Cox (Vanderbilt University Medical Center)

Alex Reiner (Fred Hutchinson Cancer Center)

[U01HG011720](#)

Polygenic Risk Estimation and Validation to ENhance Treatment - Coronary Heart Disease (PREVENT)

Iftikhar Kullo (Mayo Clinic)

Dan Schaid (Mayo Clinic)

[U01HG011710](#)

Coordinating Center (CC)

Ken Rice (University of Washington)

[U01HG011697](#)

National Human Genome Research Institute (NHGRI)

Iman Martin - Program Director, Division of Genomic Medicine

Robb Rowley - Program Director, Division of Genomic Medicine

Erin Ramos - Deputy Director, Division of Genomic Medicine

Teri Manolio - Director, Division of Genomic Medicine

Riley Wilson - Program Analyst, Division of Genomic Medicine

National Cancer Institute (NCI)

Leah Mechanic - Program Director, Epidemiology and Genomics Research Program

Elizabeth Gillanders - Branch Chief, Genomic Epidemiology Branch, Epidemiology and Genomics Research Program

Rachel Hanisch - Program Director, Genomic Epidemiology Branch, Epidemiology and Genomics Research Program

