NHLBI

WHOLE GENOME SEQUENCING

Module 17: Computational Pipeline for WGS Data

TOPMed Data Coordinating Center

July 24-26, 2019

Introduction

Schedule

Each day:

8.30	- 10.00am	Session 1
10.00am	- 10.30am	break
10.30am	- noon	Session 2
noon	- 1.30pm	lunch on your own
1.30pm	- 3.00pm	Session 3
3.00pm	- 3.30pm	break
3.30pm	- 5.00pm	Session 4

Weds 5-6pm: Social hour, South Campus Center

NHLBI
WHOLE GENOME SEQUENCING

Schedule

Wednesday (3 hours)

- Introduction
- Sequencing data formats
- Intro to Genomic Data Storage
- Phenotype harmonization
- Association tests
  - Methods and motivation

Thursday (6 hours)

- Association tests
  - GENESIS for association tests
  - Aggregate tests
- Population structure and relatedness
  - Population structure inference
  - Relatedness inference
- Mixed model association testing
- Variant annotation

Schedule

Friday (6 hours)

- Variant annotation (again)
- Pipeline design and examples
  - Analysis pipeline design
- Cloud platforms
  - Analysis Commons
  - Seven Bridges Genomics
  - Terra

Connectivity

- Wireless Connection: TBA
- Slides and schedule: https://uw-gac.github.io/SISG_2019/index.html#schedule
- Hands-on exercises: https://uw-gac.github.io/SISG_2019/
- Slack channel: (sign up!) https://sisg2019module17.slack.com

  …contact bheavner@uw.edu for help with slack

Workshop Outline and People

Wednesday pm

Intro, association testing

Ken Rice
Professor
TOPMed DCC PI

Genotypes and data formats

Stephanie Gogarten
Research Scientist

Phenotype harmonization

Leslie Emery
Research Scientist

Workshop Outline and People

Thursday:
Association testing

Ken Rice
Professor
TOPMed DCC PI

Population structure and
Relatedness, mixed models

Tim Thornton
Associate Professor
TOPMed DCC PI

Variant annotation – Thursday pm and Friday am

Deepti Jain
Research Scientist

Ben Heavner
Research Scientist

Workshop Outline and People

Friday am:

UW GAC Pipeline

Stephanie Gogarten
Research Scientist

Analysis Commons

Jen Brody
Research Scientist

Friday pm:

Seven Bridges: Dave Roberson

Terra: Allie Hajian, Tim Majarian

Workshop Outline and People

Additional GAC faculty/staff on hand to help and advise:

Cecelia Laurie
Research Scientist

Prof Bruce Weir
TOPMed DCC PI

Workshop Outline and People

And you? Please tell us – very briefly:

- Who you are
- Where you work
- What you would like to get from the module

Other essentials

- This room will not be locked
- Restrooms are available down the hallway
- Lunch options – to follow! Or follow a local…
- Bags on final day (Light Rail to airport beats taxis…)
- Questions?

NHLBI
WHOLE GENOME SEQUENCING

TOPMed overview

- Goals/structure of the TOPMed program
- What TOPMed data is available
- How to access it

Goals of the TOPMed program

- Sponsored by the NHLBI; focus on heart, lung, blood and sleep traits
- Primary goal is to identify genetic variants with effects on subclinical-disease measures, clinical disease events, disease severity and response to treatment
- Facilitate personalized approaches to prevention, diagnosis and treatment of disease

NHLBI
WHOLE GENOME
SEQUENCING

The TOPMed Program

- Provide whole-genome sequencing (WGS) and other omics measures to pre-existing studies
- WGS well advanced, several datasets freely available via dbGaP/SRA
- Other omics assays just beginning, not yet available
- **Extensive** phenotypic and exposure data for participating studies available on dbGaP

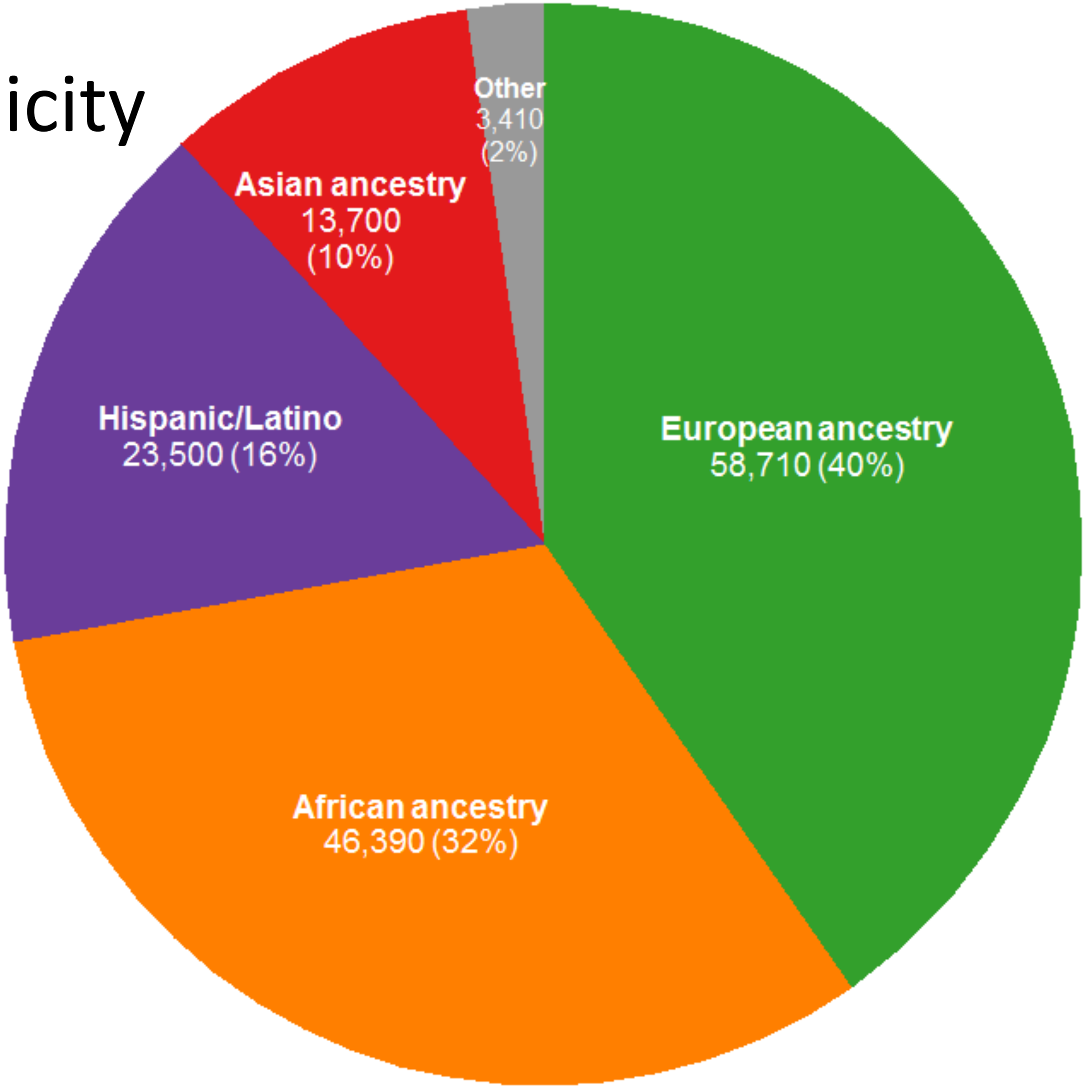
Who’s in TOPMed?

At time of writing:

- Up to 144K participants – largest WGS resource
- 72 studies (may contribute >1 subject group)
- 7 sequencing centers
- Informatics Research Center (Umich) focusing on genotype data, e.g. joint calling & analysis
- Data Co-ordinating Center (UW) focusing on genotype data, e.g. harmonization & analysis

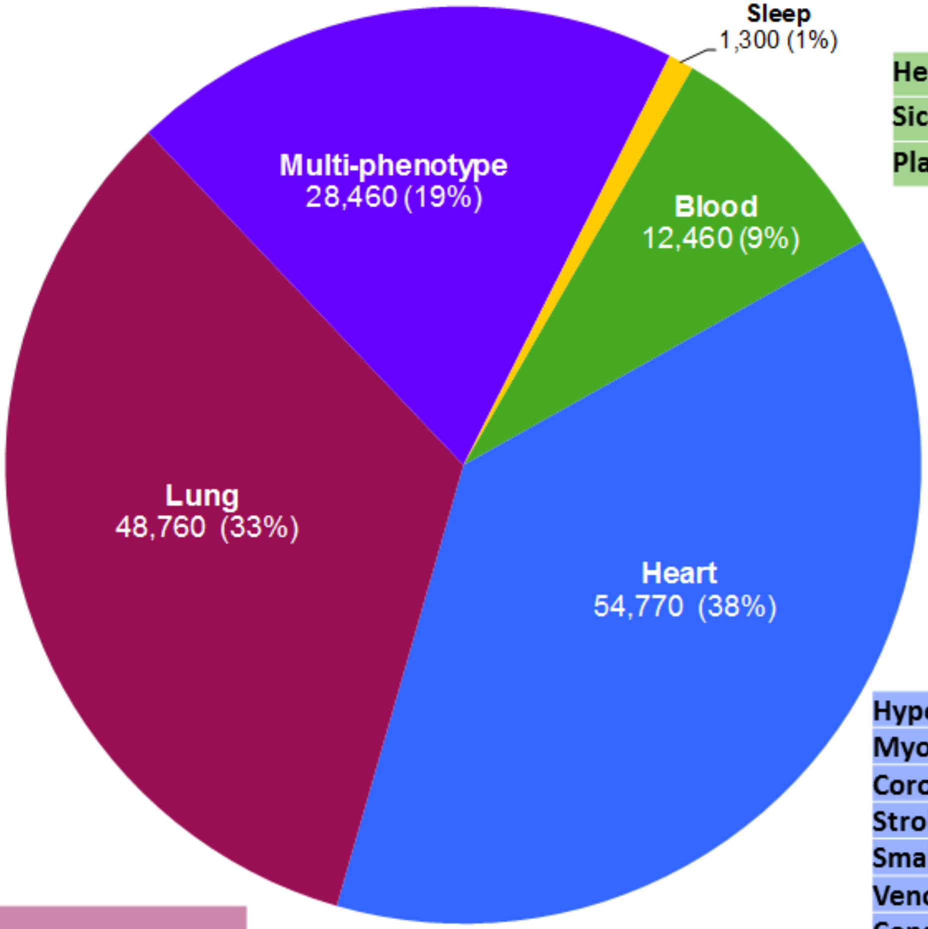
Who’s in TOPMed?

Major ancestry/ethnicity groups:



Who’s in TOPMed?

Areas of Phenotype focus:



- Hemophilia
- Sickle Cell Disease
- Platelets

- Hypertension
- Myocardial Infarction
- Coronary Artery Disease
- Stroke
- Small Vessel Disease
- Venous Thromboembolism
- Congenital Heart Disease
- Atrial Fibrillation
- Adiposity
- Congestive Heart Failure
- Lipids

- Asthma
- Chronic Obstructive Pulmonary Disease
- Idiopathic Pulmonary Fibrosis
- Sarcoidosis
- Interstitial Lung Disease

TOPMed data availability

TOPMed data are made available to the scientific community via the database for **Genotypes and Phenotypes (dbGaP)** and the **Sequence Read Archive (SRA)**

- The SRA and dbGaP are separate data archives. Both have controlled-access and open-access components. Controlled-access SRA data are restricted to approved dbGaP users.
- SRA contains DNA sequence data (CRAM files) and single-sample genotype calls (VCF) – more on these later
- dbGaP contains phenotypic data and various types of molecular data (including multi-sample VCF files)
- Today we will focus on dbGaP and SRA data structures

TOPMed data availability

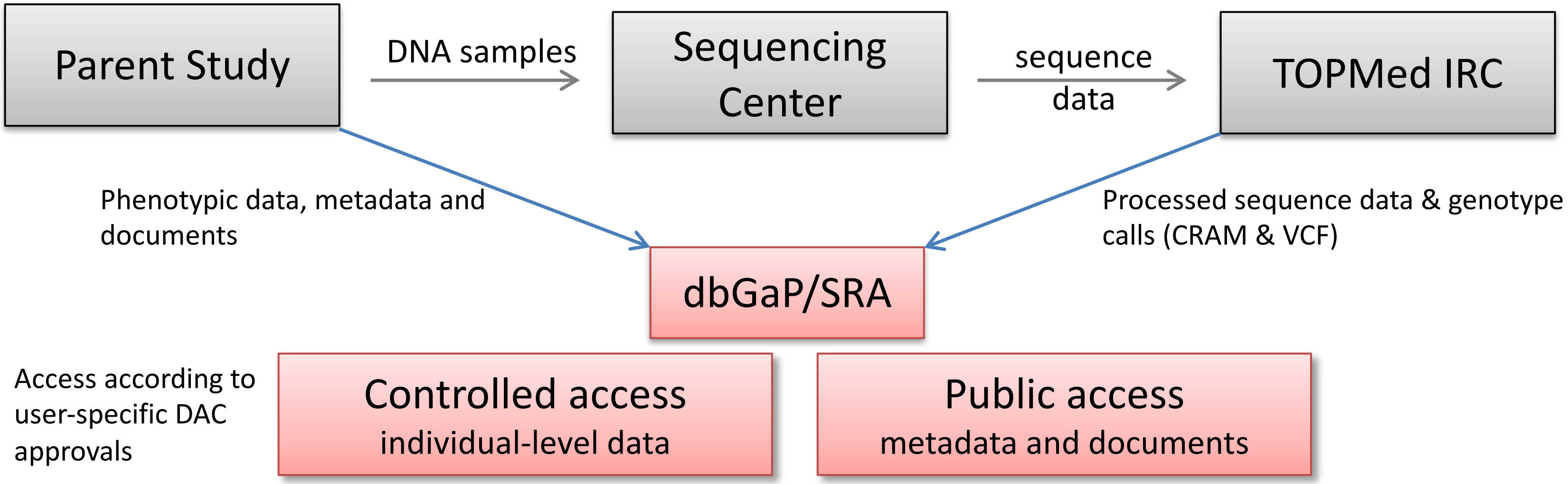
- Individual-level TOPMed data are **controlled-access** – i.e. must apply to NIH Data Access Committee, and get approval
- **Exactly** which data provided depends on **what it is used for**, because participants consent to some uses and not others
- Access via DataSTAGE (data & compute resources) is coming
- Our examples use simulated/1000G data, and (for speed) are much smaller than real WGS

TOPMed Study/Parent Study

These are currently organized as separate dbGaP accessions:

- **Parent study** = pre-existing study that recruited subjects, obtained informed consent, collected biosamples and data (including phenotypic data and various types of molecular data); provides DNA samples for TOPMed WGS. Some have been collecting data for decades.
- **TOPMed Study** = TOPMed-funded study consisting of DNA samples and phenotypic data from one or more Parent studies; some are focused on a specific disease area, while others are very broad in phenotypic characterization.

TOPMed Data Flow



- Phenotypes available through dbGaP…
- …also within-study, we are harmonizing across TOPmed (more later)

Parent study designs

Study designs reflect original "Epi" goals:

- Prospective cohort studies – focus on risk factors, longitudinal trends and incident disease
- Case-control studies – usually cross-sectional, cases and controls from the same population(s)
- Randomized trials for interventions (causation)
- Family-based genetic studies
- Case-only studies – disease severity and/or response to treatment

TOPMed study designs

- Some derived from **single Parent study** by selecting according to various criteria – e.g. relatedness, having phenotypes of interest or extent of phenotypic characterization
- Some are a **consortium of multiple Parent studies** that each contribute a common phenotype of interest – e.g. atrial fibrillation cases from several parent studies, along with controls from same/other studies

Yes, this all gets complex! But designs ***do*** matter when choosing appropriate analyses.

Watch out for…

- *dbGaP accession numbers* identify studies, sub-studies and their subordinate digital objects
- *SRA RUN identifiers* for each DNA sample as a set of CRAM, CRAI and sample-specific VCF files
- A *subject* is a person, a *sample* is an analyte or biological specimen sampled from a subject (e.g. DNA from blood)
- One subject in one study can have multiple samples! Mappings are available…

dbGaP file types (controlled access)

- Subject consent – submitted Subject IDs with associated consent group types
- Subject-Sample mapping – correspondence between subject and sample IDs
- Sample attributes – e.g. analyte type, specimen body site
- Pedigree – documented familial relationships
- Subject phenotype data
- Molecular data
- Medical imaging
- Phenotype-genotype association test results

dbGaP file types (public access)

- Data dictionaries – variable names, descriptions, encoded values, etc
- Variable reports – generated by dbGaP – variable summaries (counts, ranges, etc)
- Study documents – e.g. study design, methods of molecular data acquisition, methods of phenotypic data acquisition (including protocols and questionnaires)
- These files can be downloaded from dbGaP’s ftp site

dbGaP file structure

No one format is specified by dbGaP (!) – here are two very different examples

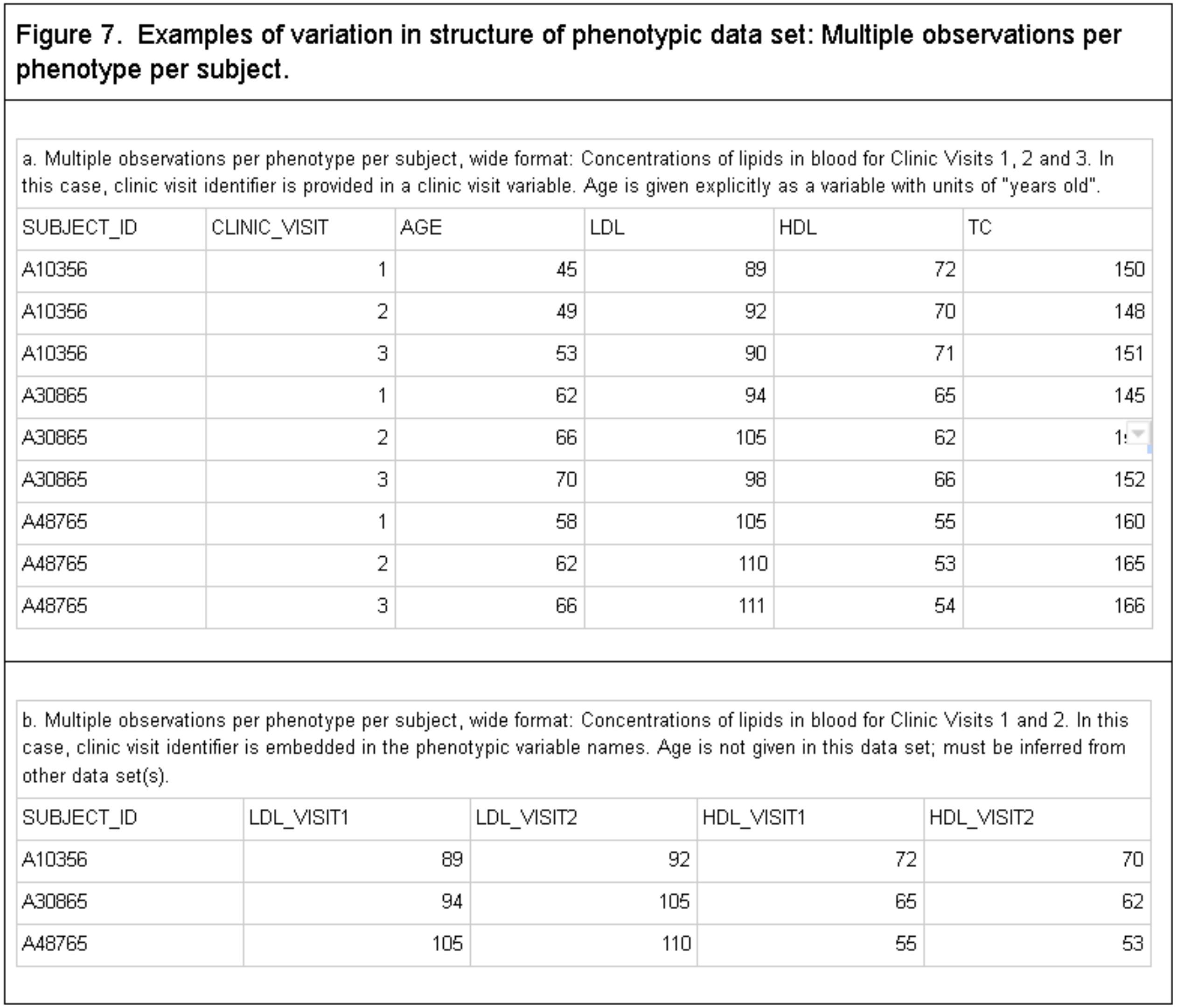


Figure 7. Examples of variation in structure of phenotypic data set: Multiple observations per phenotype per subject.

a. Multiple observations per phenotype per subject, wide format: Concentrations of lipids in blood for Clinic Visits 1, 2 and 3. In this case, clinic visit identifier is provided in a clinic visit variable. Age is given explicitly as a variable with units of "years old".

SUBJECT_ID	CLINIC_VISIT	AGE	LDL	HDL	TC
A10356	1	45	89	72	150
A10356	2	49	92	70	148
A10356	3	53	90	71	151
A30865	1	62	94	65	145
A30865	2	66	105	62	1[illegible]
A30865	3	70	98	66	152
A48765	1	58	105	55	160
A48765	2	62	110	53	165
A48765	3	66	111	54	166

b. Multiple observations per phenotype per subject, wide format: Concentrations of lipids in blood for Clinic Visits 1 and 2. In this case, clinic visit identifier is embedded in the phenotypic variable names. Age is not given in this data set; must be inferred from other data set(s).

SUBJECT_ID	LDL_VISIT1	LDL_VISIT2	HDL_VISIT1	HDL_VISIT2
A10356	89	92	72	70
A30865	94	105	65	62
A48765	105	110	55	53

dbGaP file structure

- This may seem messy/awkward
- It is, but most of those cleaning it up are volunteers, and resources are limited. If you're a trait expert affiliated to a TOPMed study, **please** join the relevant TOPMed Working Group
- More on DCC's harmonization work with Leslie, this afternoon

Discovering genetic risk factors for disease

- This is a primary goal of TOPMed
- TOPMed investigators are performing Genome-Wide Association Studies (GWAS) using genotype calls from whole-genome sequencing across multiple studies
- The process consists of several steps outlined in the following slides

TOPMed GWAS: Step 1 – Planning

Develop analysis plan, including specification of all variables needed, specifically:

- Primary outcome phenotype (e.g. HDL level in serum); if a derived variable is to be used (e.g. diabetes status), define derivation algorithm and required component variables
- Covariates to be adjusted for (e.g. age at measurement, sex, and study) or otherwise allowed for (relatedness, measurement accuracy info)
- Ancillary variables for modifying phenotypes in the model (e.g. medication use) and/or selection of subjects to include/exclude (e.g. fasting status)

TOPMed GWAS: Step 2 – Prepare the data

- Identify the necessary variables in dbGaP and construct data sets
    - Search dbGaP phenotype files for variables related to the required phenotypes
    - Decide which ones are relevant
    - Determine which subjects have both relevant phenotypes and genotypic data (from TOPMed WGS)
    - Determine which subjects with pheno/genotype also gave **consent for this analysis**
- Harmonize phenotypes across studies
    - Evaluate similarities and differences among studies and develop harmonization plan
    - QC source variables
    - Write and run harmonization code on each study
    - QC harmonized phenotypes
    - Identify subject exclusions (e.g. non-fasting, <18 years old, outliers, etc.)

TOPMed GWAS: Step 3 – Prepare genotypes

Assuming genotype QC has been done, still need to:

- Start with a genotype call set constructed from joint calling of all subjects to be included in the analysis
- For subjects to be analyzed, calculate and analyze relatedness and population structure; decide on any further exclusions
- Calculate Genetic Relatedness Matrix for samples to be included
- Define genomic aggregation units (i.e. genomic ranges for genes, regulatory elements, etc.)
- Define variant filtering (e.g. minor allele count, conservation score, loss-of function, etc.)

TOPMed GWAS: Step 4 – analyze & interpret

- Select analysis pipeline (e.g. ENCORE, GENESIS, etc.)
- Select computing environment for the analysis pipeline, including I/O, memory requirements and parallelization strategy
- Perform association tests, visualize results
- Evaluate and interpret association test results
  - Evaluate model fit, type I error rate control, heteroscedasticity
  - Modify analysis plan as needed – possibly rerun, or filter out worst behavior
  - Check for novel hits (typically using follow-up conditional analysis)
  - Develop hypotheses about causal variants and affected gene(s)
    - Compare results to genomic annotations for variants, including eQTL (e.g. using GTEx)
    - Examine possible functions of implicated genes (e.g. using MODs)

Questions?

- Ask one of us, or use the Slack channel
- Visit the TOPMed website (some material restricted to TOPMed investigators)

Acknowledgments

- The TOPMed program supported by NHLBI
- TOPMed investigators and their Parent Studies
- Participants of Parent studies
- TOPMed sequencing centers
- Members of the TOPMed DCC and IRC