Challenges and New Approaches for Complex Trait Mapping in Ancestrally Diverse Populations

Timothy Thornton, PhD
Robert W. Day Endowed Professor of Public Health
Department of Biostatistics
University of Washington

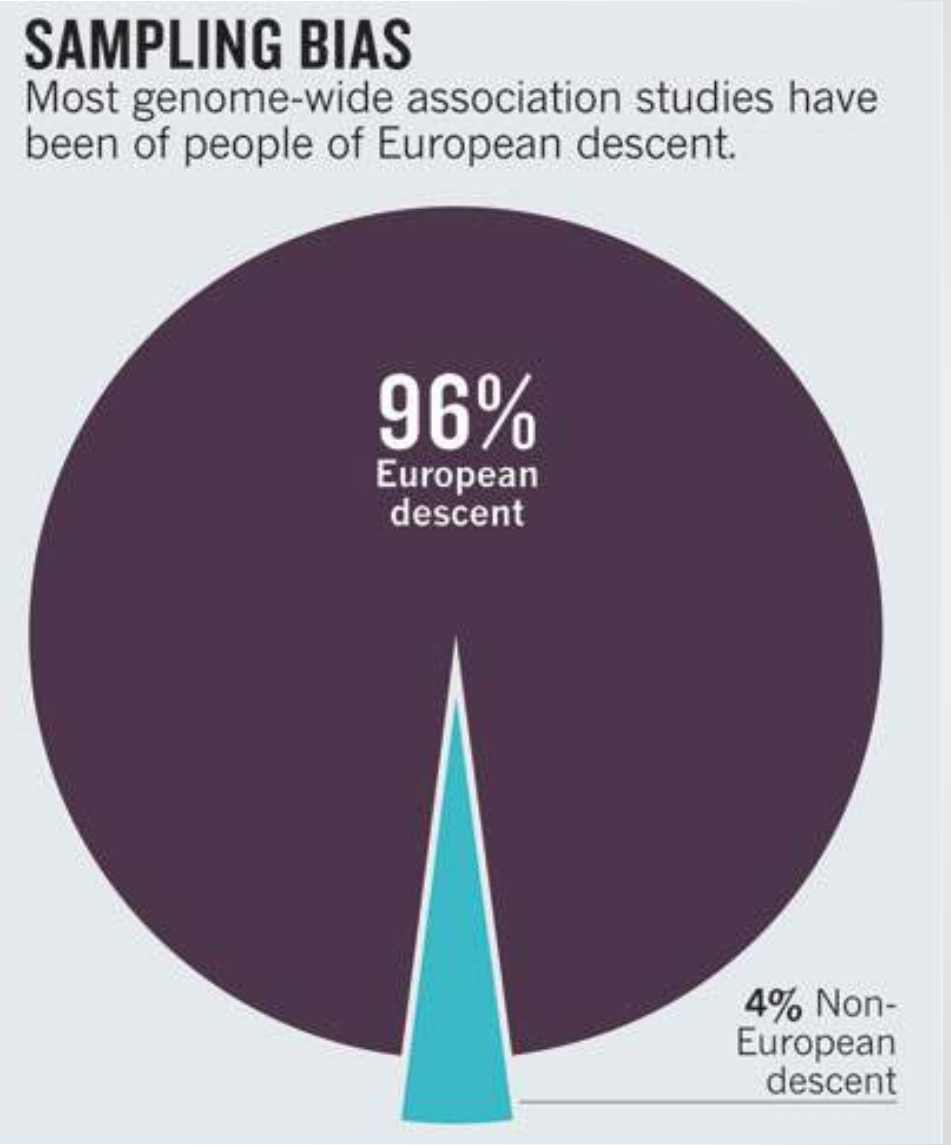
SAGES 2017

June 9, 2017

SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

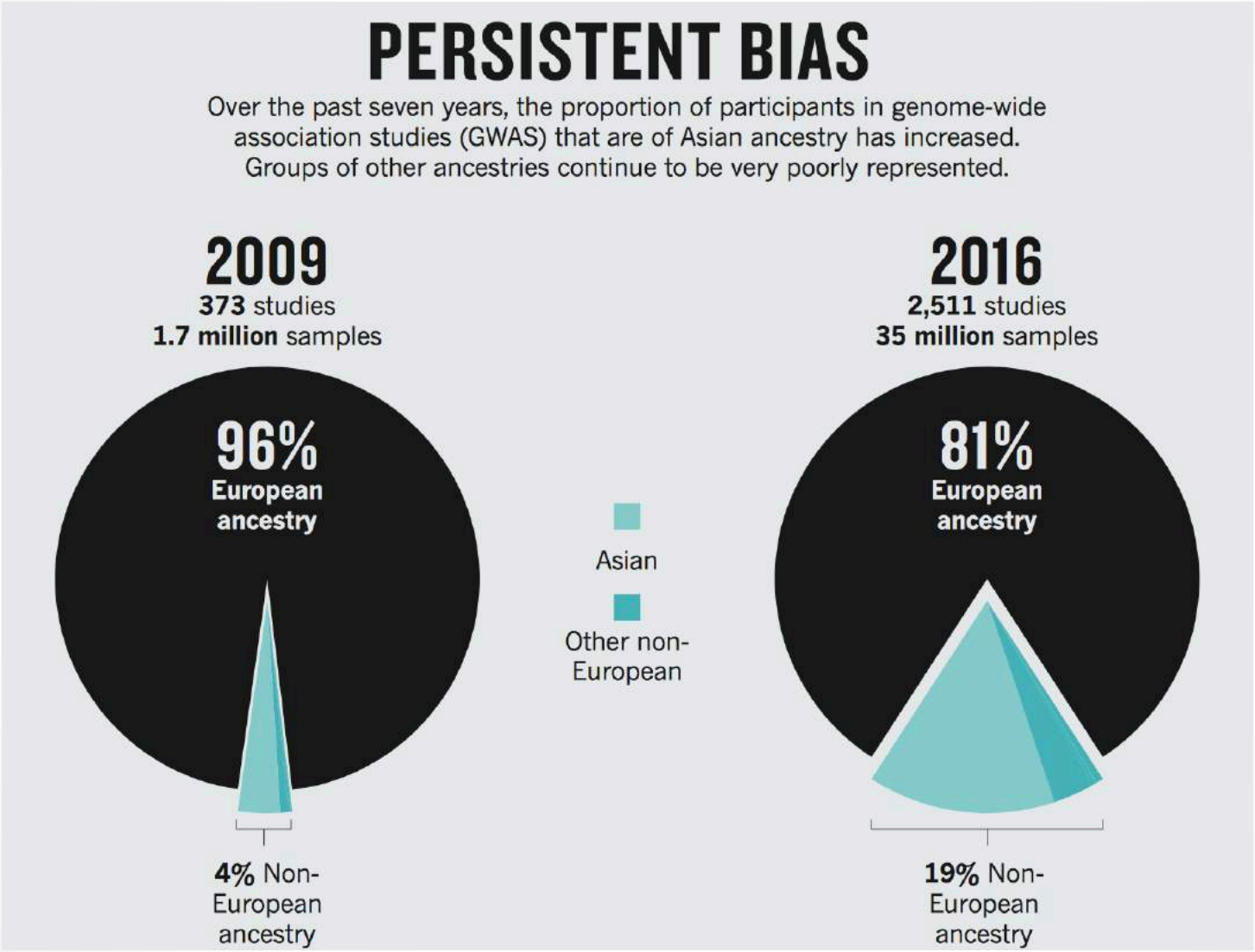
Introduction

- To date, the genomes of tens of millions of individuals have been interrogated in GWAS and sequencing association studies for the mapping of complex traits.
- The vast majority of these studies, however, have been conducted in populations of European ancestry



Bustamante et al. (Nature, 2011)

Current State of Affairs



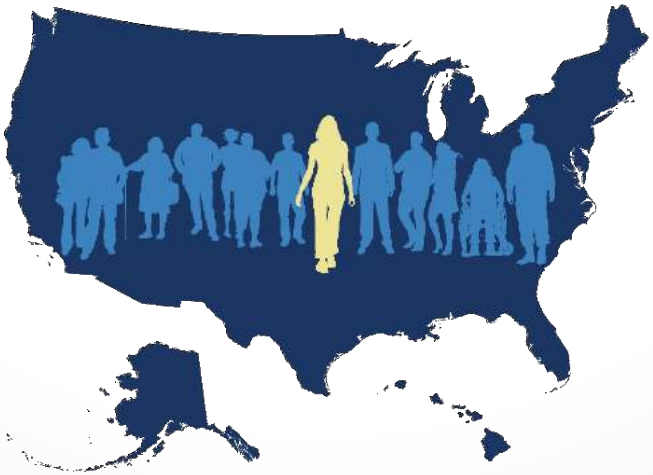
Popejoy and Fullerton (Nature, 2016)

Need for Genetic Studies in Diverse Populations

- Medical genomics has focused almost entirely on those of European descent.
- Other race and ethnic groups must be studied to ensure that more people benefit

Bustamante et al. (Nature, 2011)

The U.S. Precision Medicine Initiative® Cohort Program



"And that's why we're here today.
Because something called ***Precision Medicine*** … gives us one of the greatest opportunities for new medical breakthroughs that we have ever seen."

State of the Union Address
January 20, 2015

Precision Medicine Initiative

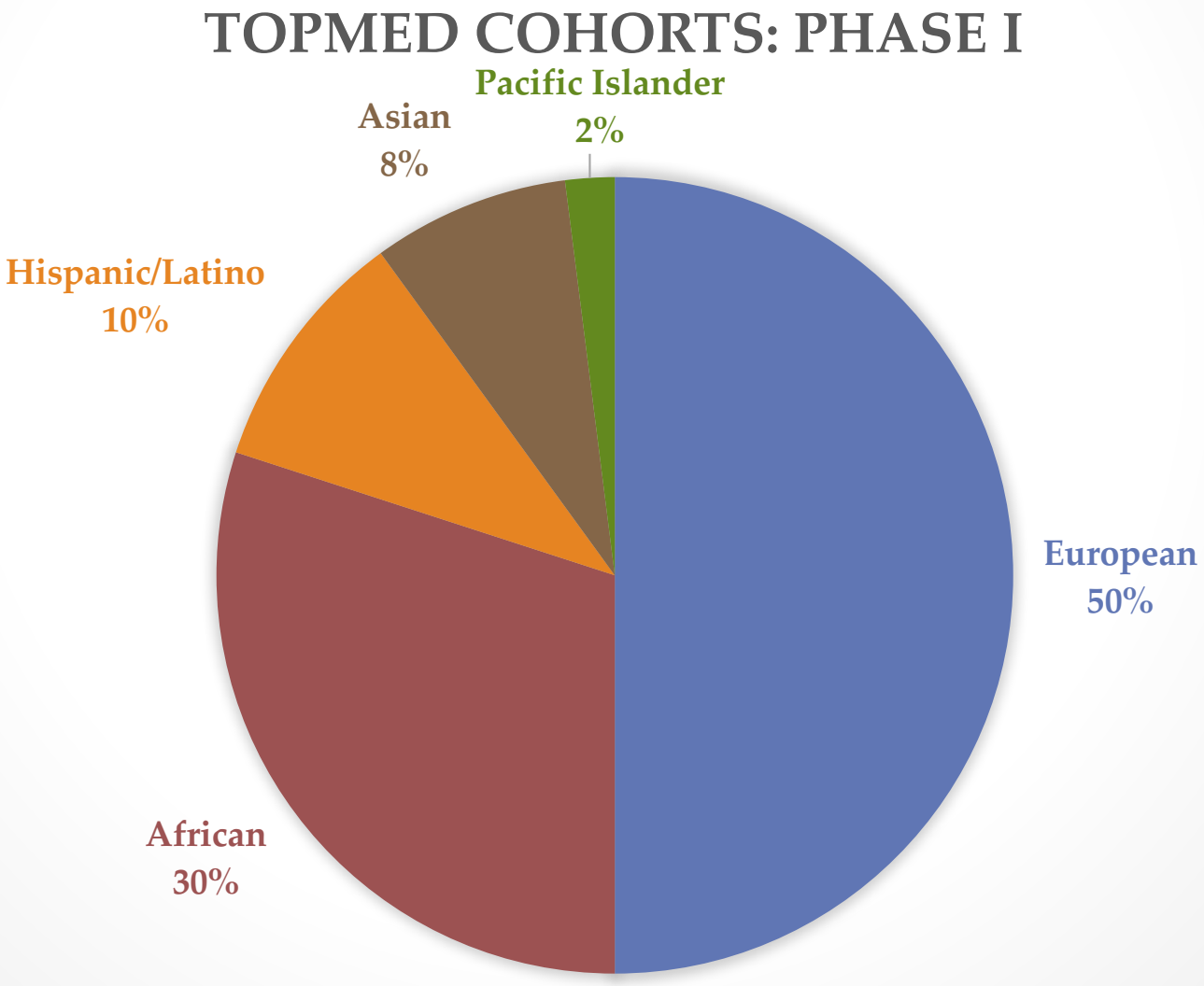
- NIH launched the Precision Medicine Initiative (PMI) in 2015
  - PMI Cohort Program will build a large research cohort of **one million or more** Americans
  - Goal is to support and advance the targeted prevention and treatment strategies that take an individual's unique characteristics into account, **including individual genome sequences**, environmental factors and lifestyles.

TOPMed WGS Project

- NIH/NHLBI Trans-Omics for Precision Medicine (TOPMed) Program is a component of the PMI
- TOPMed Whole-genome-sequence (WGS) project currently generating deep WGS data for over **120,000 individuals**
- More than 30 cohorts studies with well-defined phenotypes and existing clinical outcomes data:
- Primary aims is to identify genetic variants for increased or decreased risk of disease, as well as those that help define disease subtypes.
- As of January 2017, 62,000 whole genomes have been completed
- **University of Washington Genetic Analysis Center** is the Data Coordinating Center for the TOPMed WGS Project
-

Multi-ethnic TOPMed Cohorts

- Concerted effort to be reflective of the diverse ancestries of the U.S. population.



TOPMed WGS Project: Opportunities

- Identification of novel low frequency and rare genetic variants underlying phenotypic diversity
- Potential to provide new insights in human health and health disparities of minority populations for many complex diseases

TOPMed WGS Project: Challenges

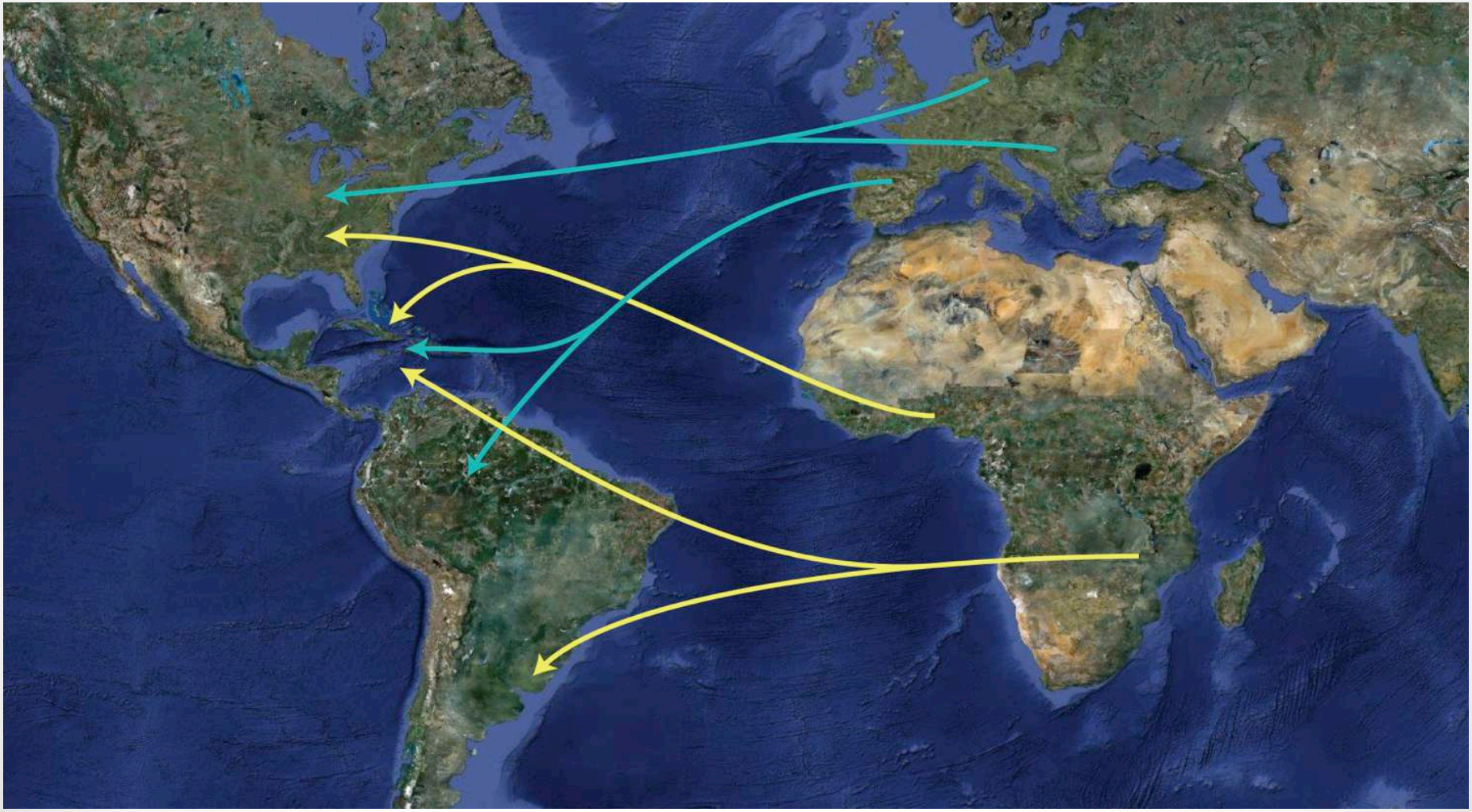
- Some of the challenges for TOPMed WGS data:
  - Multi-ethnic populations
  - Variety of study designs: family, case-control, cohort based designs, founder populations (Amish).
  - **Confounding due to highly heterogeneous genetic and environmental backgrounds**
  - Computational burden for analysis of deep whole genome sequence data for 120,000+ individuals
  - Population structure inference and correction with whole genome sequencing data: common and rare variants

Genetic Relatedness in Multi-ethnic Populations

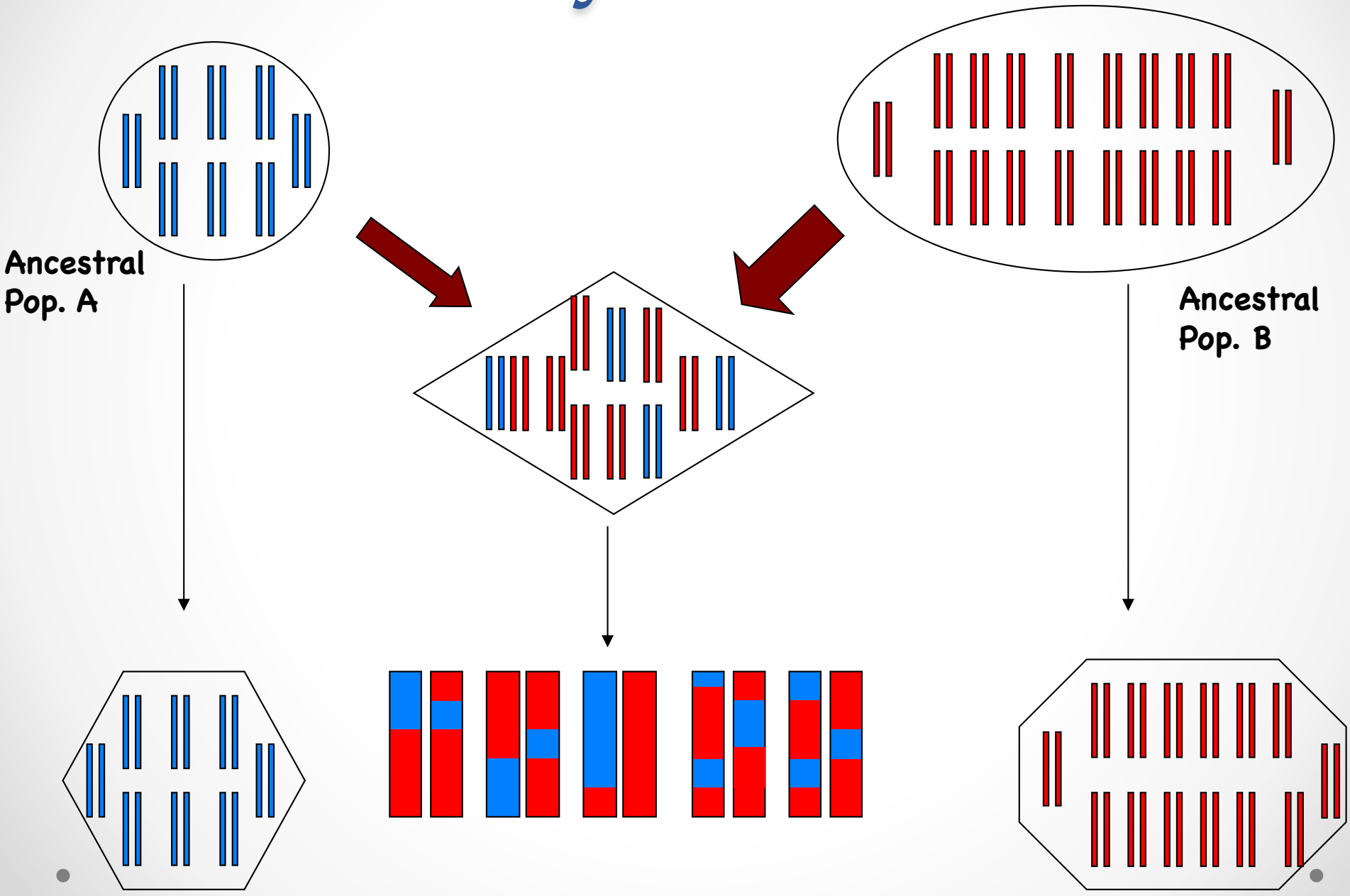
- The genealogy of individuals in a sample consists of:
  - Distant genetic relatedness, such as population structure
  - Recent genetic relatedness: pedigree relationships of close relatives (known and cryptic!)

- Samples from ancestrally diverse populations have complex genealogy due to ancestry admixture and both recent and distant genetic relatedness
- Distinguishing familial relatedness from ancestry using genotype data in diverse populations is difficult, as both manifest as genetic similarity through the sharing of alleles.

Complex Genealogy of Multi-Ethnic Admixed Populations

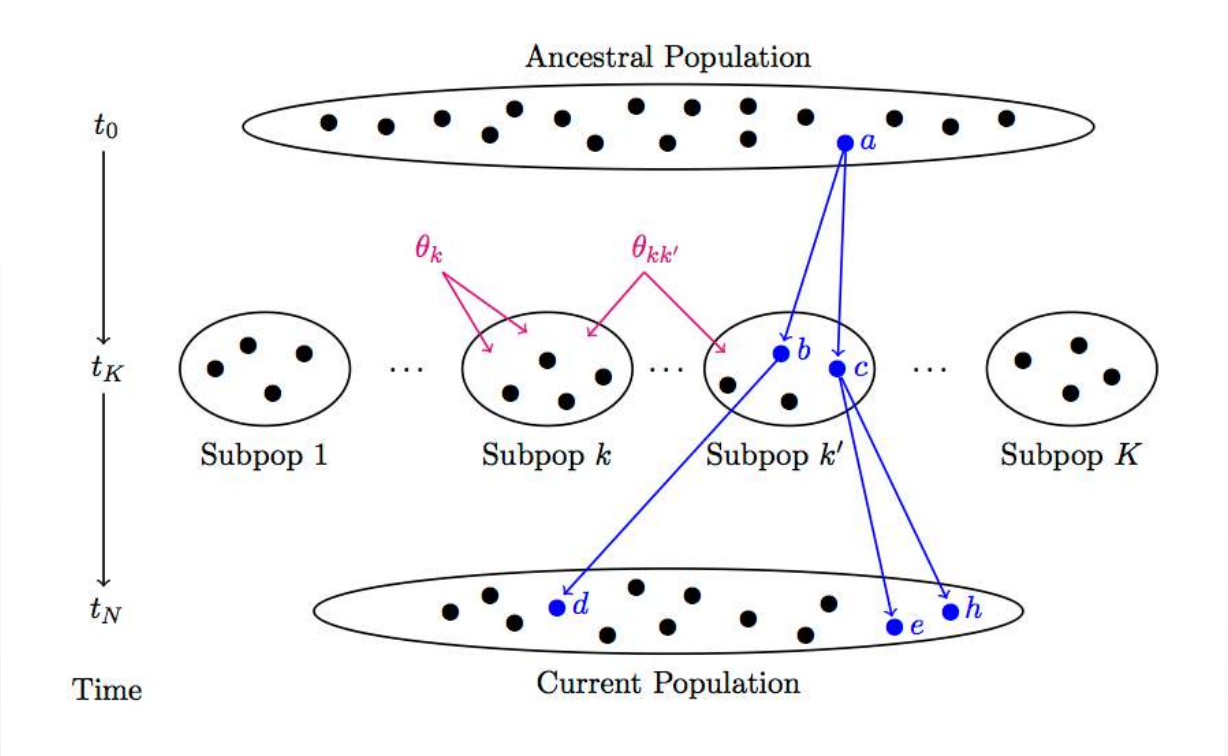


Ancestry Admixture



Recent versus Distant Genetic Relatedness

- Distinguishing familial relatedness from ancestry using genotype data in diverse populations is difficult, as both manifest as genetic similarity through the sharing of alleles.



Conomos et al. (AJHG, 2016)

Deconvolution of Genetic Relatedness

- **Conomos et al., *Am J Hum Genet*, 2016**

ARTICLE

Model-free Estimation of Recent Genetic Relatedness

Matthew P. Conomos,[1,*] Alexander P. Reiner,[2,3] Bruce S. Weir,[1] and Timothy A. Thornton[1,*]

- **Conomos et al., Genet Epidemiology, 2015**

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness

Matthew P. Conomos,[1] Michael B. Miller,[2] and Timothy A. Thornton[1]*

- **Thornton et al., *Am J Hum Genet*, 2012**

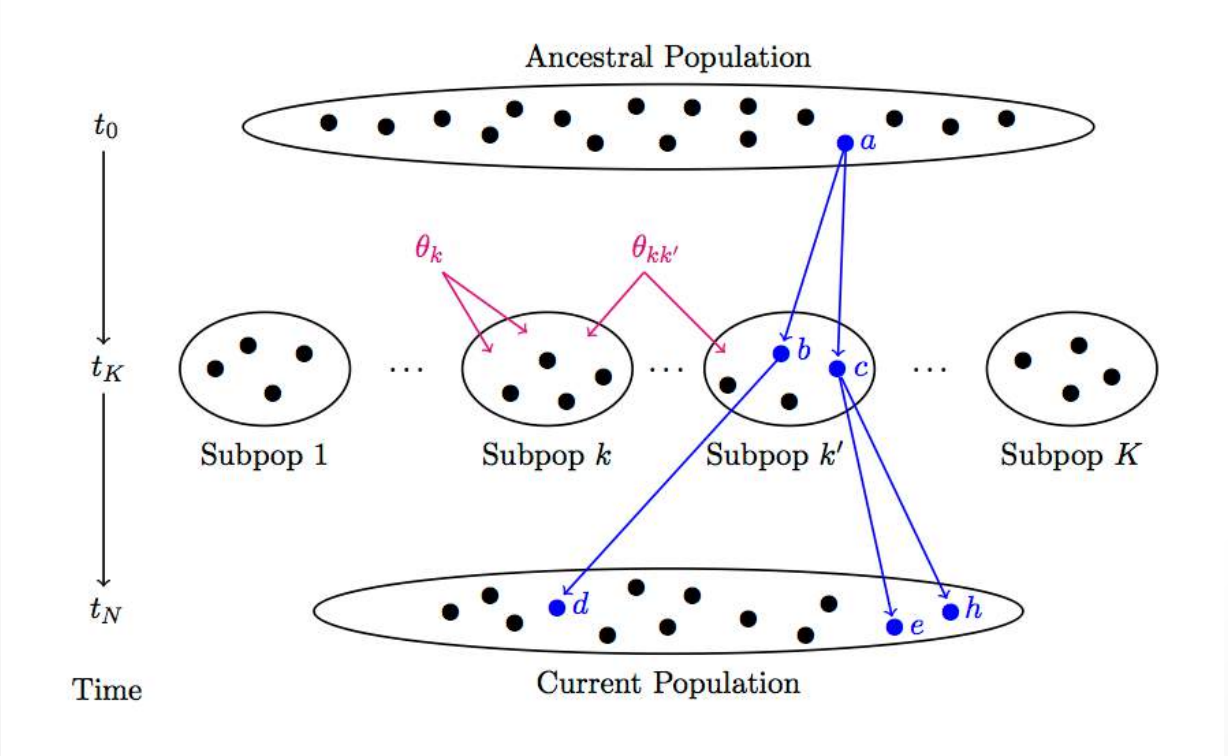
ARTICLE

Estimating Kinship in Admixed Populations

Timothy Thornton,[1,*] Hua Tang,[2] Thomas J. Hoffmann,[3,4] Heather M. Ochs-Balcom,[5] Bette J. Caan,[6] and Neil Risch[3,4,6,*]

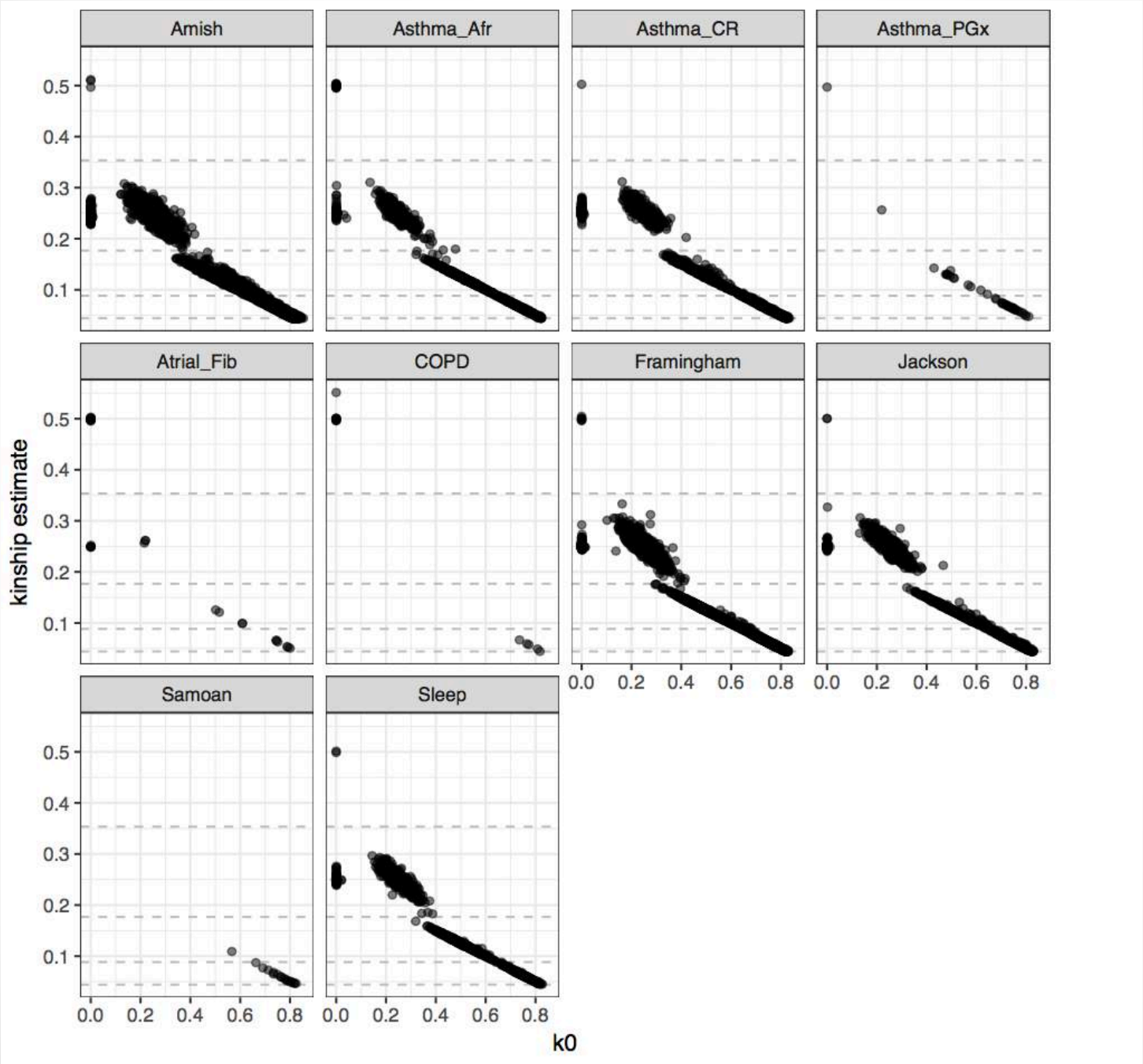
Recent versus Distant Genetic Relatedness

- Distinguishing familial relatedness from ancestry using genotype data in diverse populations is difficult, as both manifest as genetic similarity through the sharing of alleles.



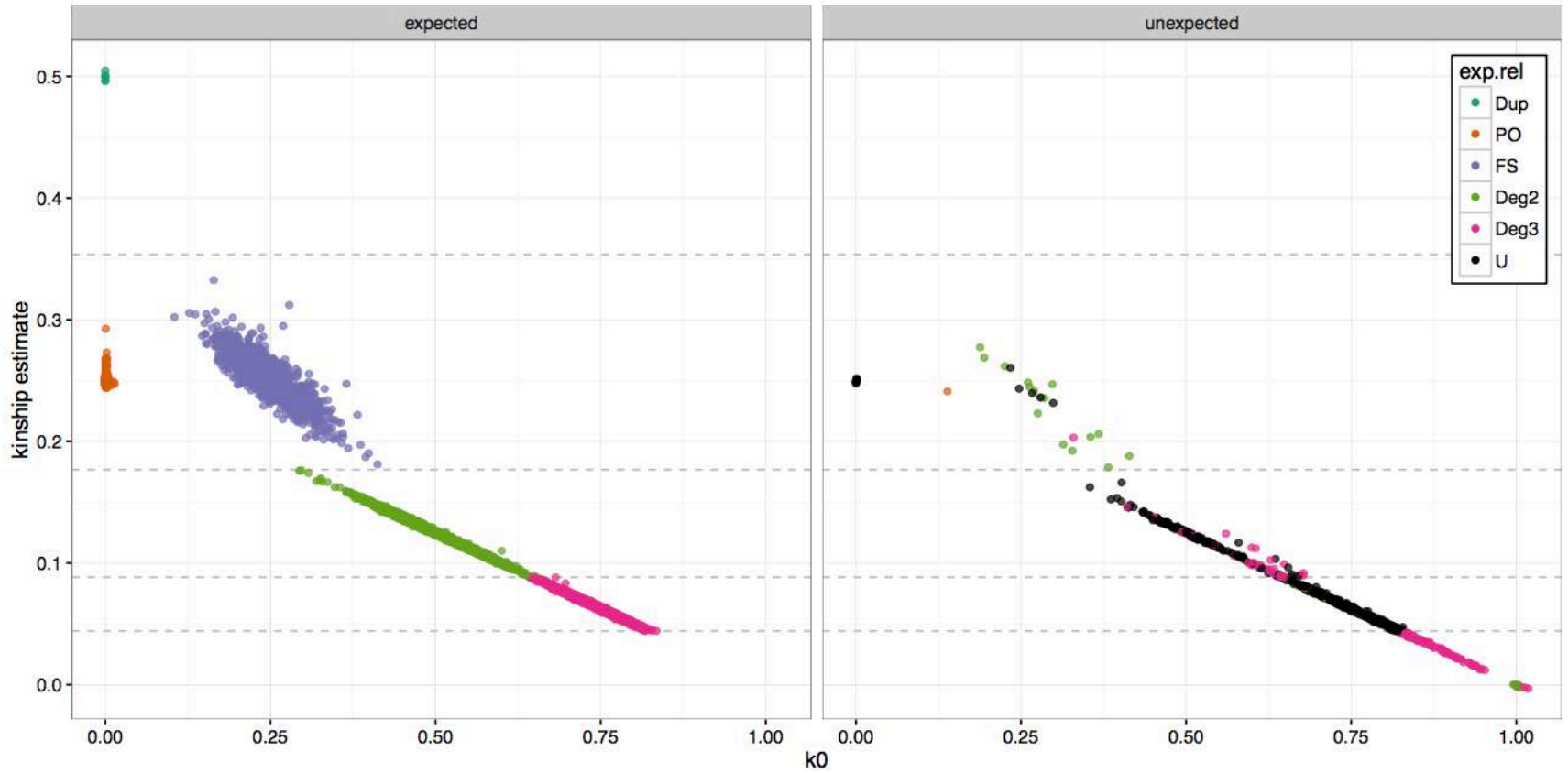
Conomos et al. (AJHG, 2016)

TOPMed Recent Genetic Relatedness



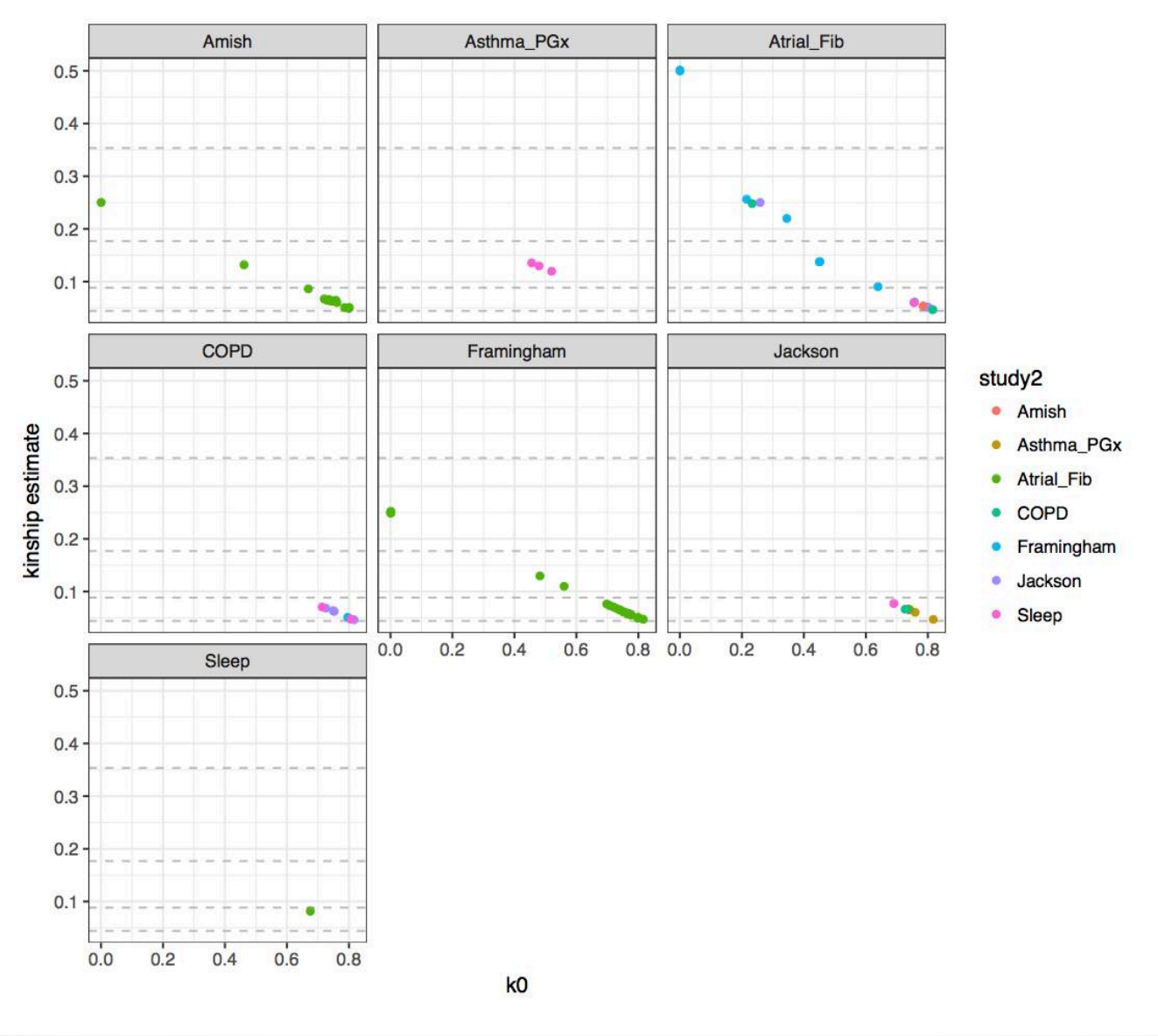
TOPMed Recent Genetic Relatedness

- Unexpected relatedness that does not match "known" pedigrees.
- Example: Framingham Heart Study

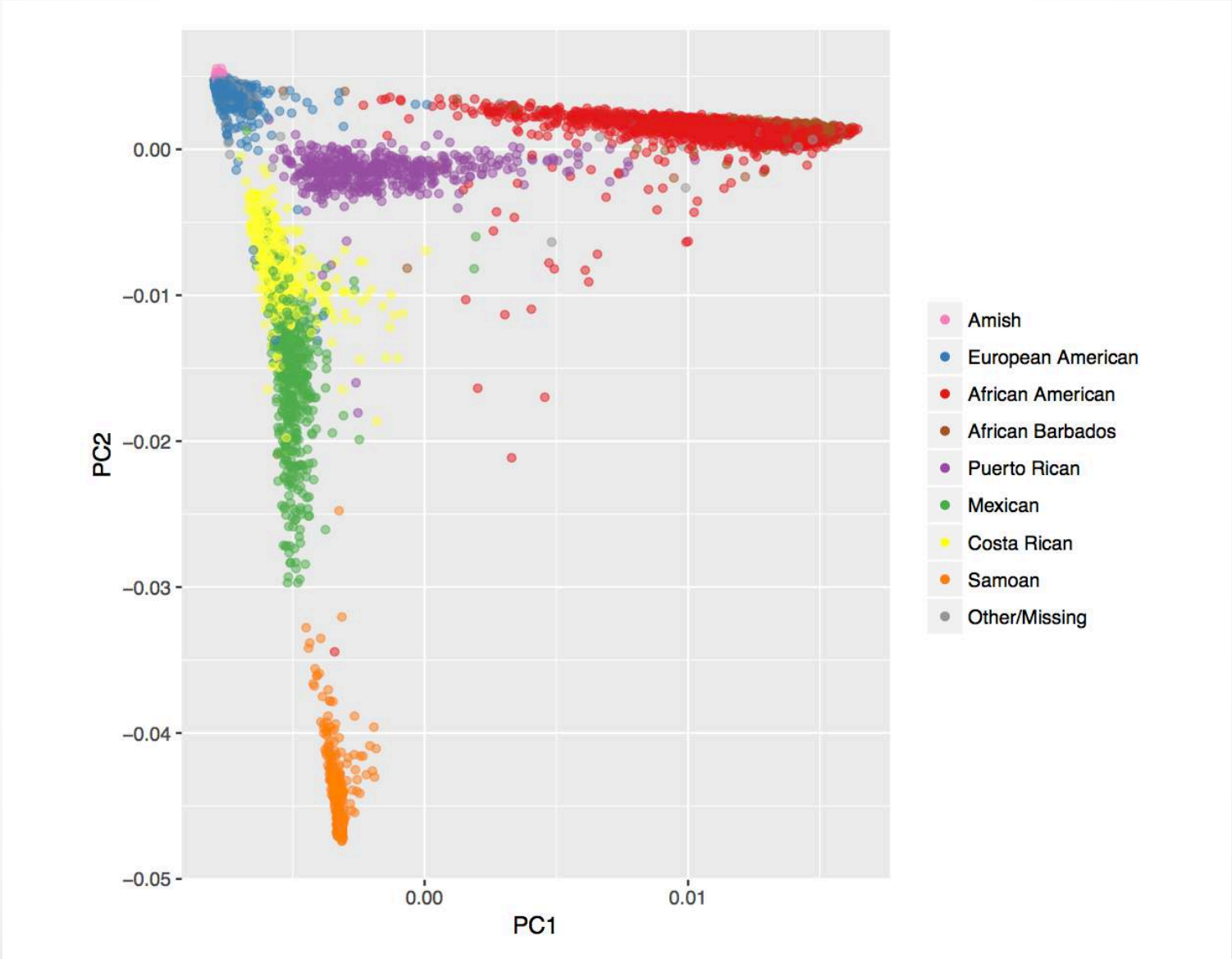


TOPMed Recent Genetic Relatedness

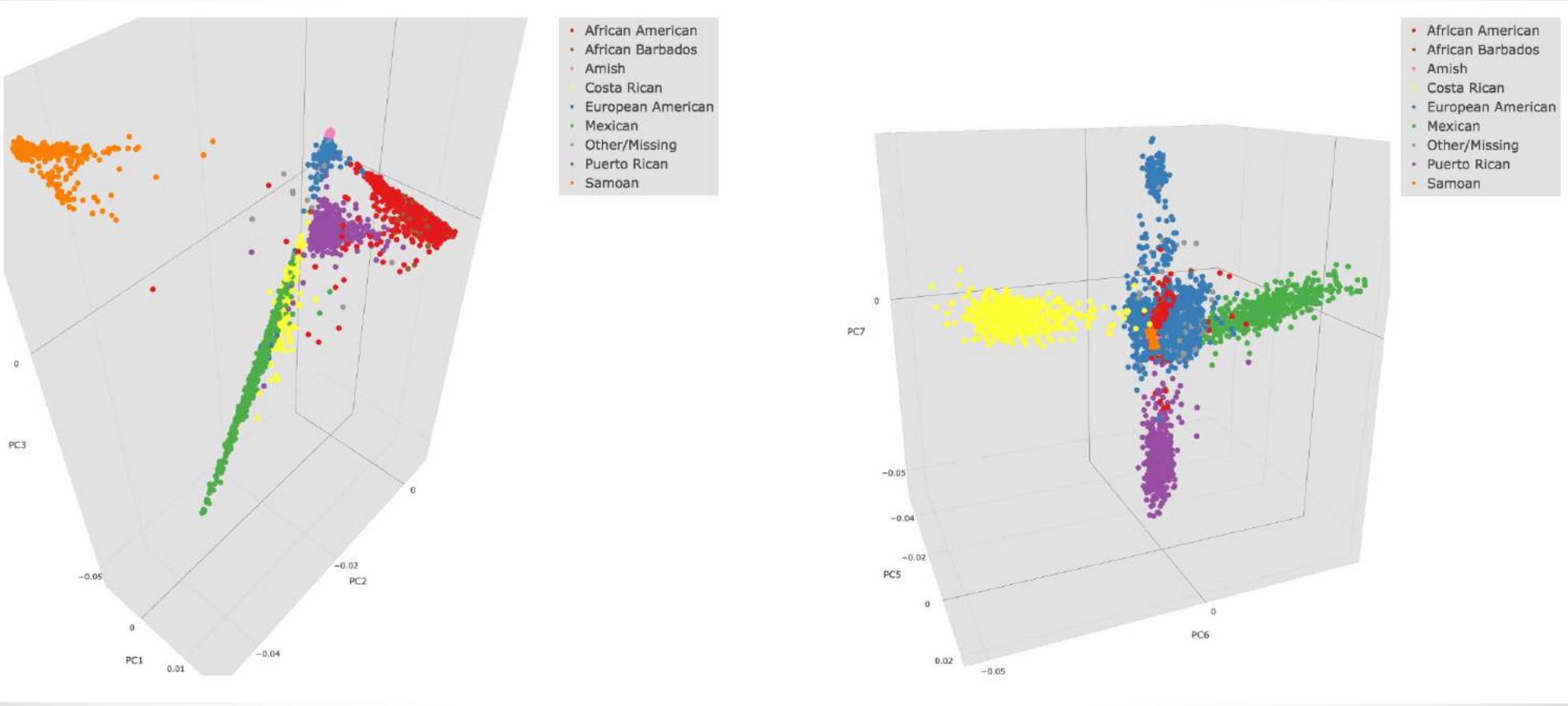
- Finding "unexpected" relatedness across TOPMed cohort studies as well



Genetic Ancestry Inference in TOPMed (with PC-AiR)



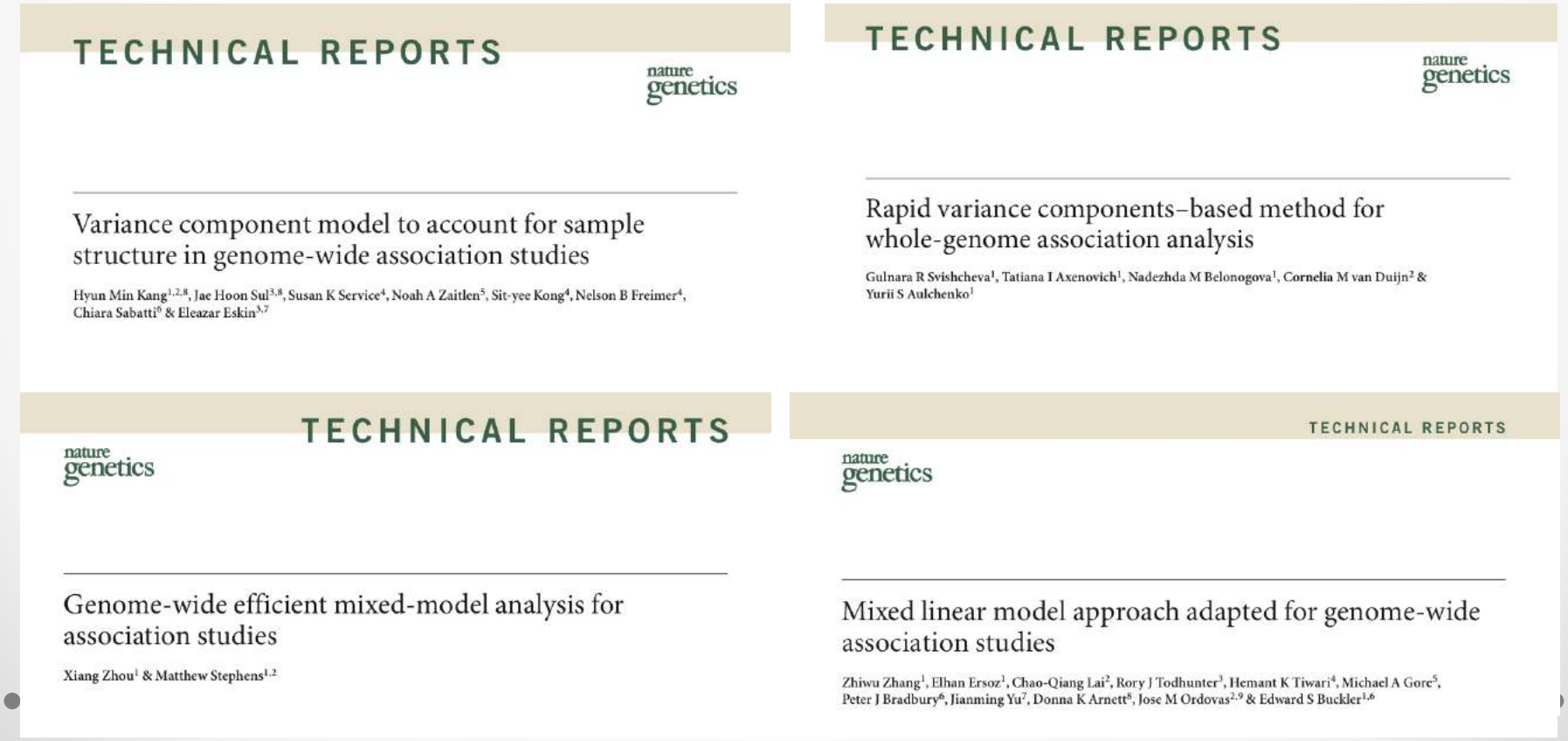
TOPMed: Population Structure Inference



Genetic association mapping in the TOPMed samples: diverse ancestries and complex sample structure.

Linear Mixed Models

- Linear mixed models (LMMs) have emerged as a powerful and effective approach for genetic association testing of single variants in the presence of sample structure



Association Mapping in Multi-Ethnic Populations

- Matt Conomos PhD work developed **LMM-OPS** for association mapping in ancestrally diverse populations
- **LMM-OPS**, linear mixed models **with orthogonal partitioned structure**
- Appropriately accounts for the complex genealogy of ancestrally diverse samples by partitioning sample structure into **two orthogonal components**:
  1. a component for the sharing of alleles inherited identical by descent (IBD) from recent common ancestors, which represents familial relatedness
  2. and another component for allele sharing due to more distant common ancestry, which represents population structure.

New LMM approach for Admixed Populations

- With LMM-OPS, a score test for association is calculated based on the following linear mixed model:

$$\mathbf{Y} = \mathbf{g}_s\beta_s + \mathbf{X}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \qquad \text{with} \qquad \boldsymbol{\epsilon} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma} \equiv \sigma_A^2\boldsymbol{\Phi} + \sigma_\epsilon^2\mathbf{I}\right)$$

where:

- $\boldsymbol{\Phi}$ is an genetic relatedness matrix adjusted for ancestry admixture (via the PCs) with PC-Relate
- $\mathbf{V}$ is a matrix with PCs from PC-AiR, and $\boldsymbol{\gamma}$ is a vector (unknown) ancestry effects on the phenotype
- $\mathbf{X}$ is a matrix of covariate values with vector $\boldsymbol{\alpha}$ of covariate effects

Applications and Discoveries in Hispanic/Latino Populations

ARTICLE

Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

Matthew P. Conomos,[1,14,*] Cecelia A. Laurie,[1,14] Adrienne M. Stilp,[1,14] Stephanie M. Gogarten,[1,14] Caitlin P. McHugh,[1] Sarah C. Nelson,[1] Tamar Sofer,[1] Lindsay Fernández-Rhodes,[2] Anne E. Justice,[2] Mariaelisa Graff,[2] Kristin L. Young,[2] Amanda A. Seyerle,[2] Christy L. Avery,[2] Kent D. Taylor,[3] Jerome I. Rotter,[3] Gregory A. Talavera,[4] Martha L. Daviglus,[5] Sylvia Wassertheil-Smoller,[6] Neil Schneiderman,[7] Gerardo Heiss,[2] Robert C. Kaplan,[6] Nora Franceschini,[2] Alex P. Reiner,[8] John R. Shaffer,[9] R. Graham Barr,[10] Kathleen F. Kerr,[1] Sharon R. Browning,[1] Brian L. Browning,[11] Bruce S. Weir,[1] M. Larissa Avilés-Santa,[12] George J. Papanicolaou,[12] Thomas Lumley,[13] Adam A. Szpiro,[1] Kari E. North,[2] Ken Rice,[1] Timothy A. Thornton,[1] and Cathy C. Laurie[1,*]

ORIGINAL ARTICLE

Genetic Associations with Obstructive Sleep Apnea Traits in Hispanic/Latino Americans

Brian E. Cade[1,2], Han Chen[3], Adrienne M. Stilp[4], Kevin J. Gleason[1], Tamar Sofer[4], Sonia Ancoli-Israel[5,6,7], Raanan Arens[8], Graeme I. Bell[9], Jennifer E. Below[10], Andrew C. Bjonnes[11], Sung Chun[11,12], Matthew P. Conomos[4], Daniel S. Evans[13], W. Craig Johnson[4], Alexis C. Frazier-Wood[14], Jacqueline M. Lane[1,2,15,16], Emma K. Larkin[17], Jose S. Loredo[18], Wendy S. Post[19], Alberto R. Ramos[20], Ken Rice[4], Jerome I. Rotter[21], Neomi A. Shah[22], Katie L. Stone[13], Kent D. Taylor[21], Timothy A. Thornton[4], Gregory J. Tranah[13], Chaolong Wang[3,23], Phyllis C. Zee[24], Craig L. Hanis[10], Shamil R. Sunyaev[11,12,16], Sanjay R. Patel[1,2,25], Cathy C. Laurie[4], Xiaofeng Zhu[26], Richa Saxena[1,15,16], Xihong Lin[3], and Susan Redline[1,2,25]

ARTICLE

Genome-wide Association Study of Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans

Ursula M. Schick,[1,2,3,16] Deepti Jain,[4,16] Chani J. Hodonsky,[5,16] Jean V. Morrison,[4] James P. Davis,[6] Lisa Brown,[4] Tamar Sofer,[4] Matthew P. Conomos,[4] Claudia Schurmann,[2,3] Caitlin P. McHugh,[4] Sarah C. Nelson,[4] Swarooparani Vadlamudi,[6] Adrienne Stilp,[4] Anna Plantinga,[4] Leslie Baier,[7] Stephanie A. Bien,[1] Stephanie M. Gogarten,[4] Cecelia A. Laurie,[4] Kent D. Taylor,[8,9] Yongmei Liu,[10] Paul L. Auer,[11] Nora Franceschini,[5] Adam Szpiro,[4] Ken Rice,[4] Kathleen F. Kerr,[4] Jerome I. Rotter,[8] Robert L. Hanson,[7] George Papanicolaou,[12] Stephen S. Rich,[13,14] Ruth J.F. Loos,[2,3,15] Brian L. Browning,[4] Sharon R. Browning,[4] Bruce S. Weir,[4] Cathy C. Laurie,[4] Karen L. Mohlke,[6] Kari E. North,[5,16] Timothy A. Thornton,[4,16] and Alex P. Reiner[1,16,*]

ARTICLE

Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models

Han Chen,[1,8] Chaolong Wang,[1,2,8] Matthew P. Conomos,[3] Adrienne M. Stilp,[3] Zilin Li,[1,4] Tamar Sofer,[3] Adam A. Szpiro,[3] Wei Chen,[5] John M. Brehm,[5] Juan C. Celedón,[5] Susan Redline,[6] George J. Papanicolaou,[7] Timothy A. Thornton,[3] Cathy C. Laurie,[3] Kenneth Rice,[3] and Xihong Lin[1,*]

ASSOCIATION STUDIES ARTICLE

Genome-wide association study of dental caries in the Hispanic Communities Health Study/Study of Latinos (HCHS/SOL)

Jean Morrison[1], Cathy C. Laurie[1], Mary L. Marazita[2,3,4], Anne E. Sanders[5], Steven Offenbacher[6], Christian R. Salazar[7,8], Matthew P. Conomos[1], Timothy Thornton[1], Deepti Jain[1], Cecelia A. Laurie[1], Kathleen F. Kerr[1], George Papanicolaou[9], Kent Taylor[10], Linda M. Kaste[11], James D. Beck[5] and John R. Shaffer[2,*]

ASSOCIATION STUDIES ARTICLE

Genome-wide association study of iron traits and relation to diabetes in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL): potential genomic intersection of iron and glucose regulation?

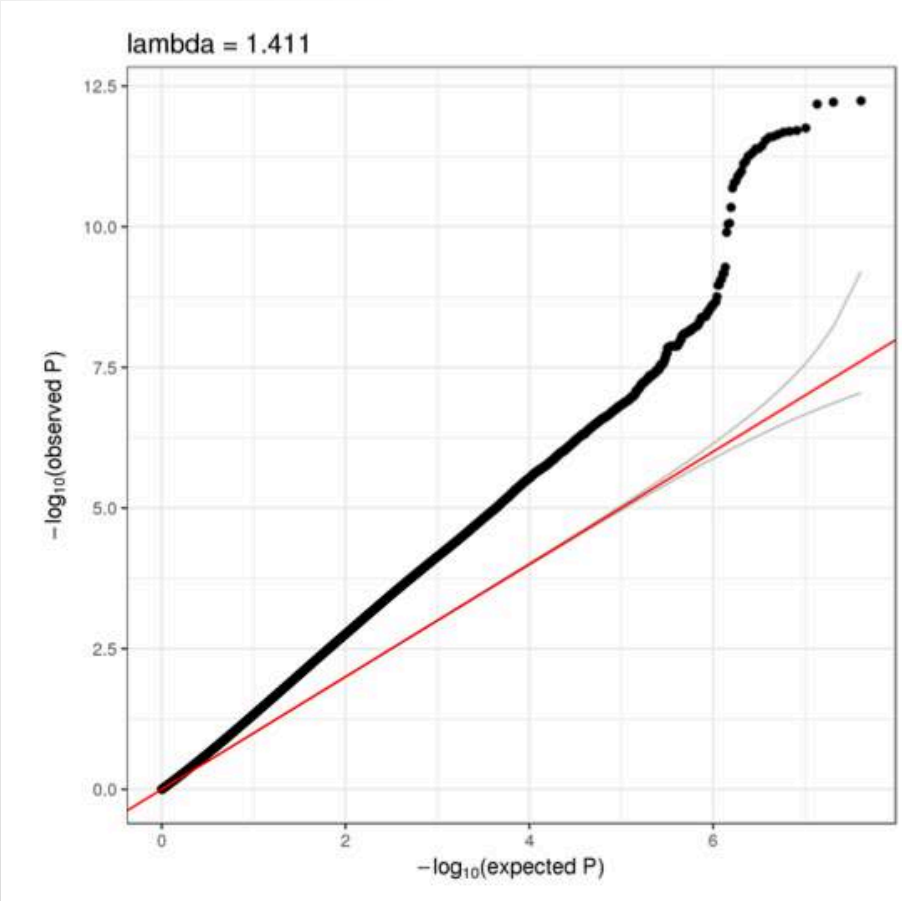
TOPMed Association Analysis with Linear Mixed Models

- LMM-OPS worked well for the Hispanic Community Health Study / Study of Latinos
- Applied LMM-OPS to a few TOPMed phenotypes that are well-defined with previously identified and replicated genome-wide significant variants
- Conducted a mega-analysis for a combined analysis of all studies Included variants in the association analysis that have a minor allele count ≥ 10 in all studies combined

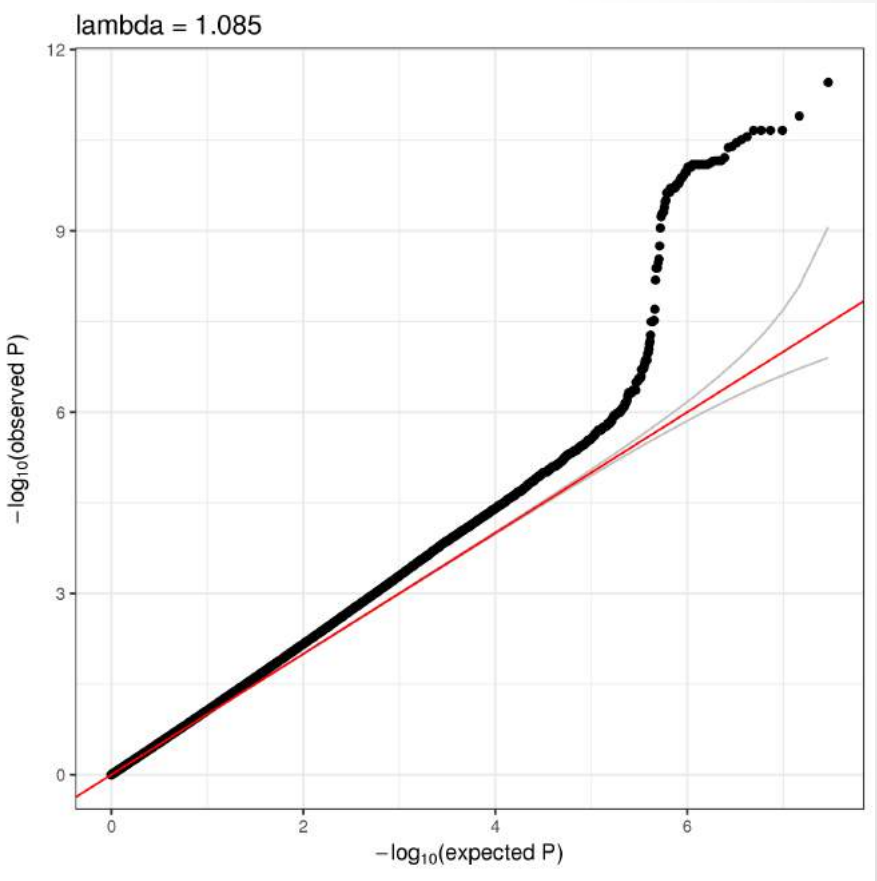
TOPMed: BMI and Hemoglobin analysis with LMM-OPS

- What went wrong in TOPMed?

BMI TOPMed Association



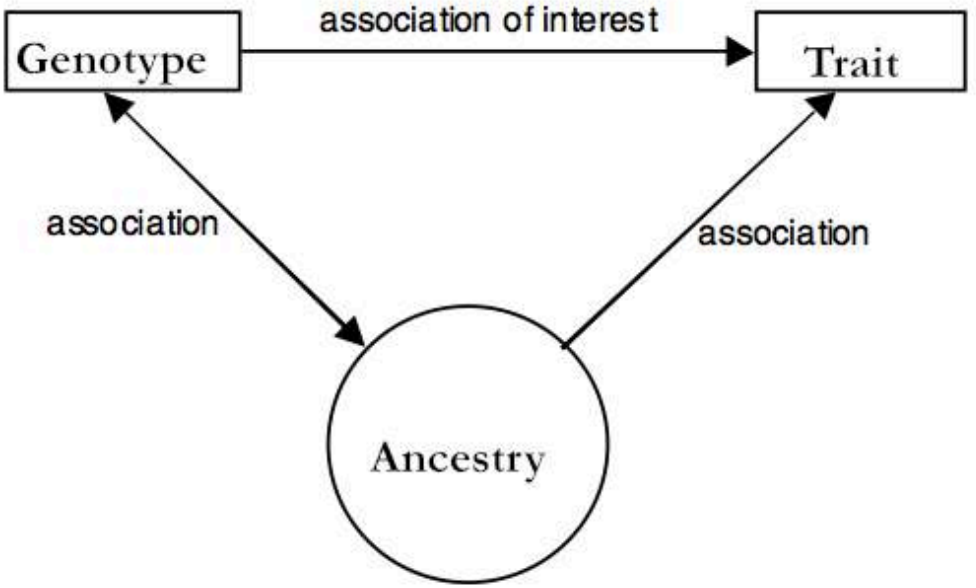
Hemoglobin TOPMed Association



What About Non-Genetic / Environmental Factors?

Confounding in TOPMed due to genetic and non-genetic factors

- Ethnic groups (and subgroups) often share distinct dietary habits and other lifestyle characteristics that result in traits of interest having different distributions that are correlated with genetic ancestry and/or ethnicity.

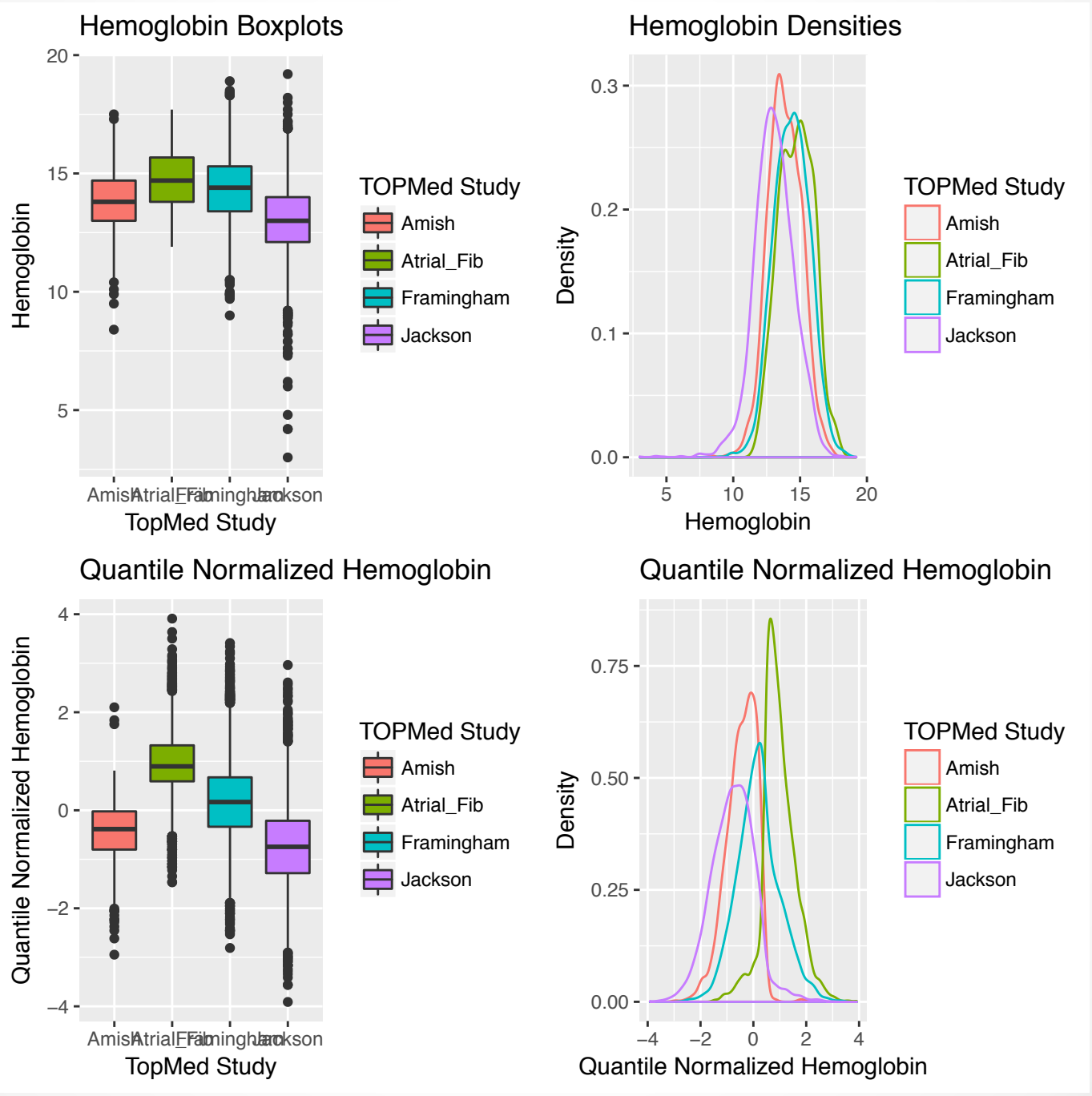


Heterogeneity in Phenotypic Variance in TOPMed

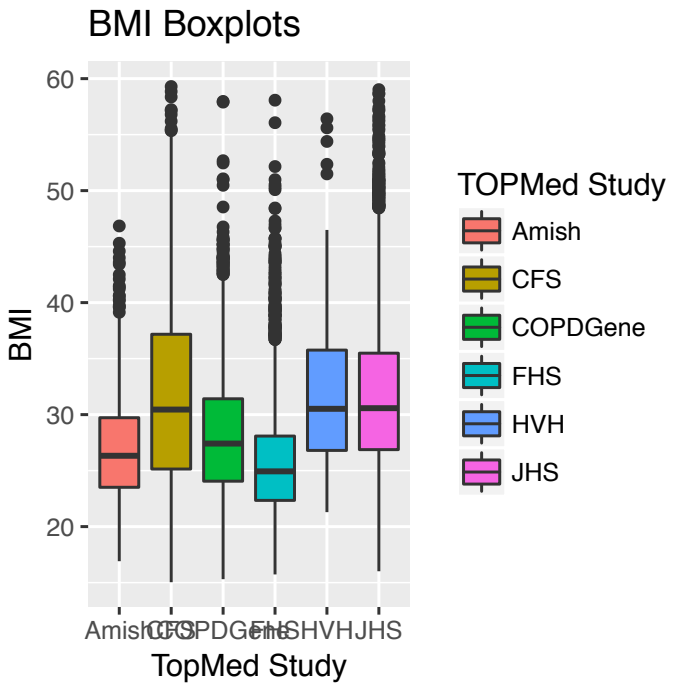
- For some TOPMed traits, we are seeing significant heterogeneity in phenotypic distributions across different ethnic/ancestry groups in TOPMed
- Self reported ethnicity in TOPMed Phase I

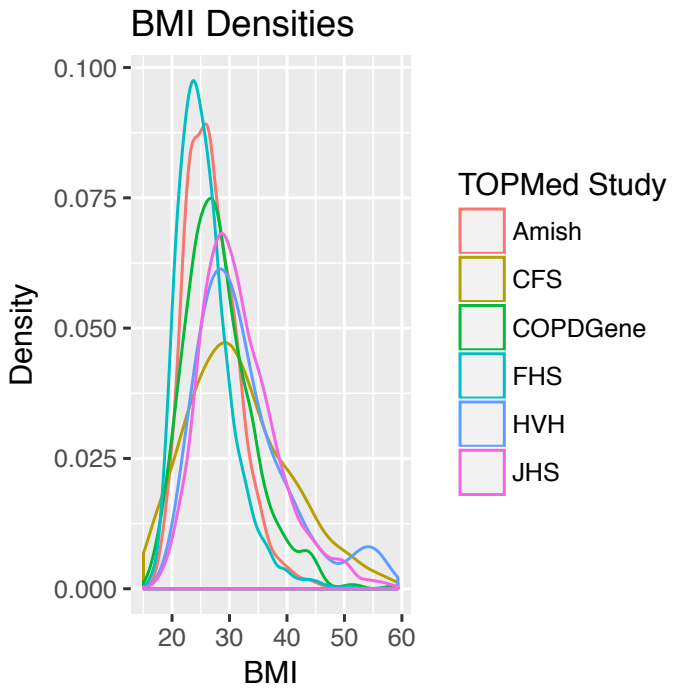
American Indian or Alaska Native	20
Asian	3
Black	5113
More than one race	26
Native Hawaiian or Pacific Islander	381
Other	1987
White	5823

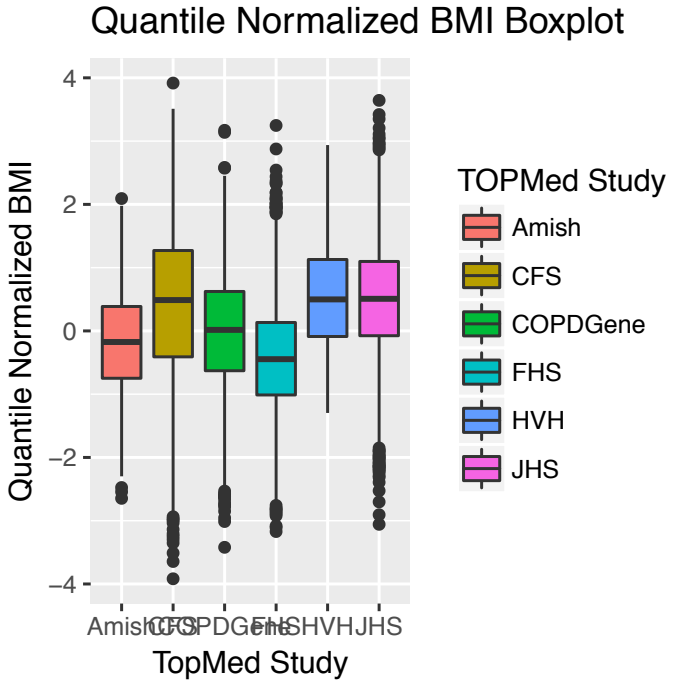
TOPMed Hemoglobin Distributions

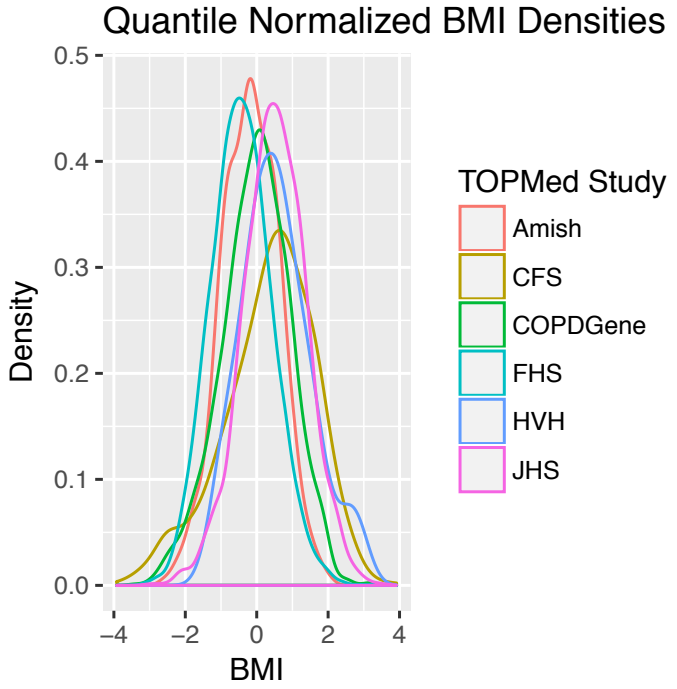


TOPMed BMI Distributions









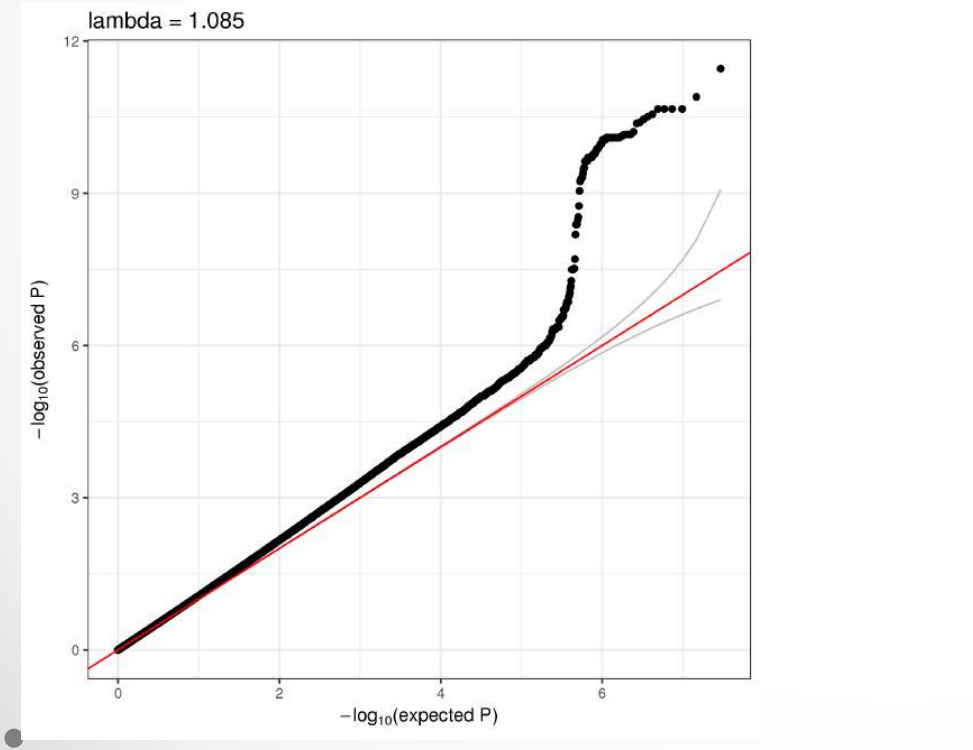
Heterogeneity in Phenotypic Variance in TOPMed

- Extended LMM-OPS to allow for multiple random effects to be included in the model, in addition to a kinship/genetic relatedness matrix.
- Used LMM-OPS with additional **random effects to allow for heterogeneous variances** in TOPMed by study or self-reported race/ ethnicity.

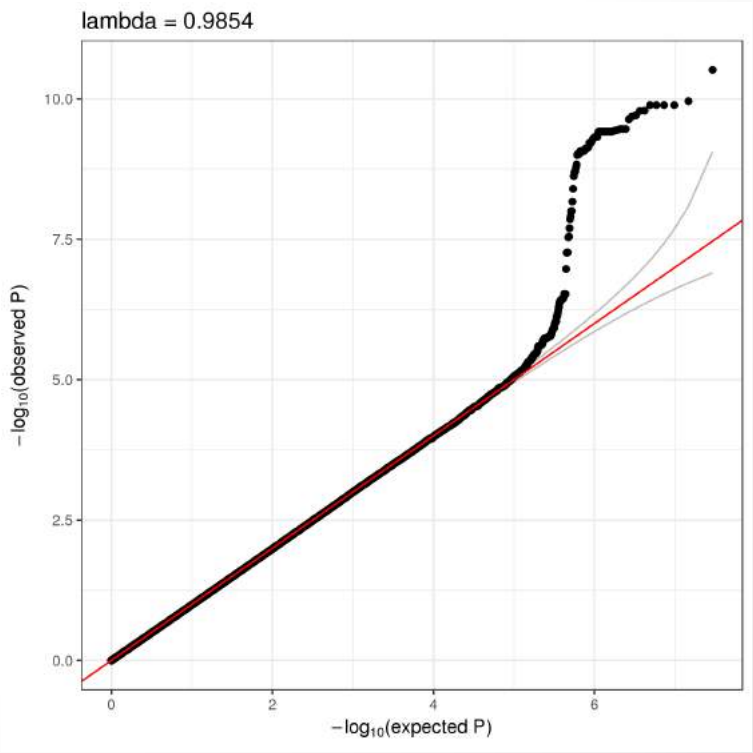
Heterogeneity in Hemoglobin Phenotypic Variances: By Study

- Association results for Hemoglobin allowing for heterogeneous phenotypic variances

Homogenous Residual Variance



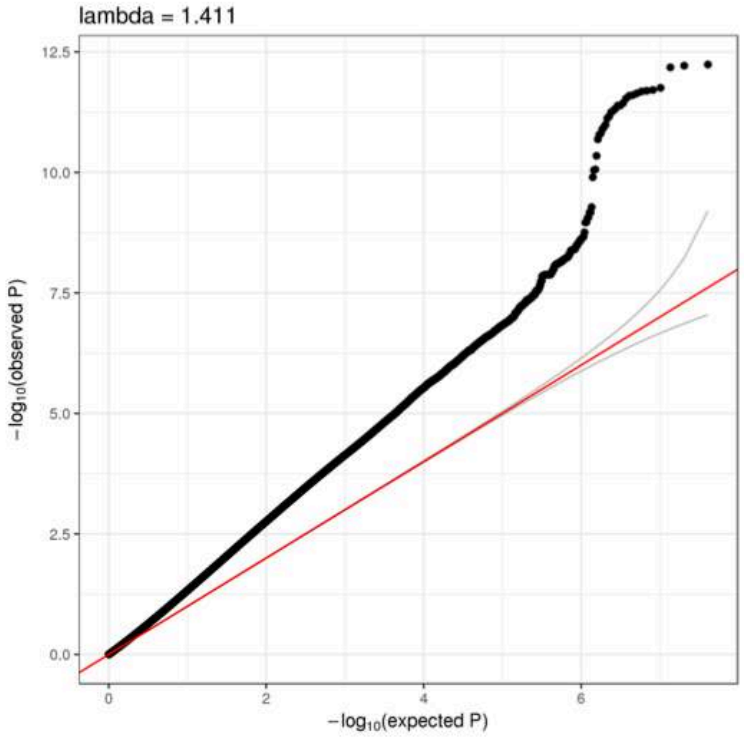
Heterogeneous Residual Variances by Study



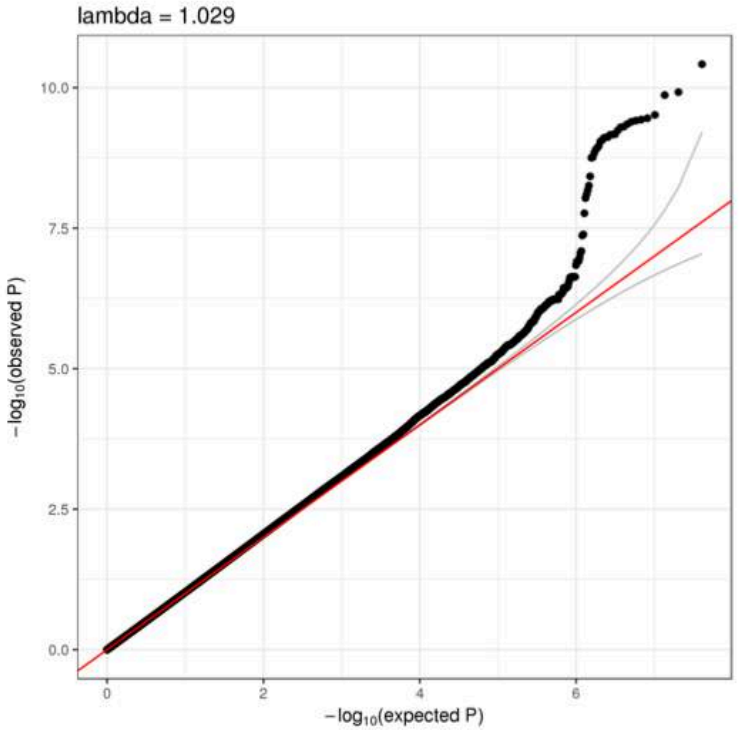
Heterogeneity in BMI Phenotypic Variances

- Association results for body mass index (BMI) allowing for heterogeneous phenotypic variances

Homogenous Residual Variance



Heterogeneous Residual Variances by Study



LMM with heterogenous residual variances for BMI

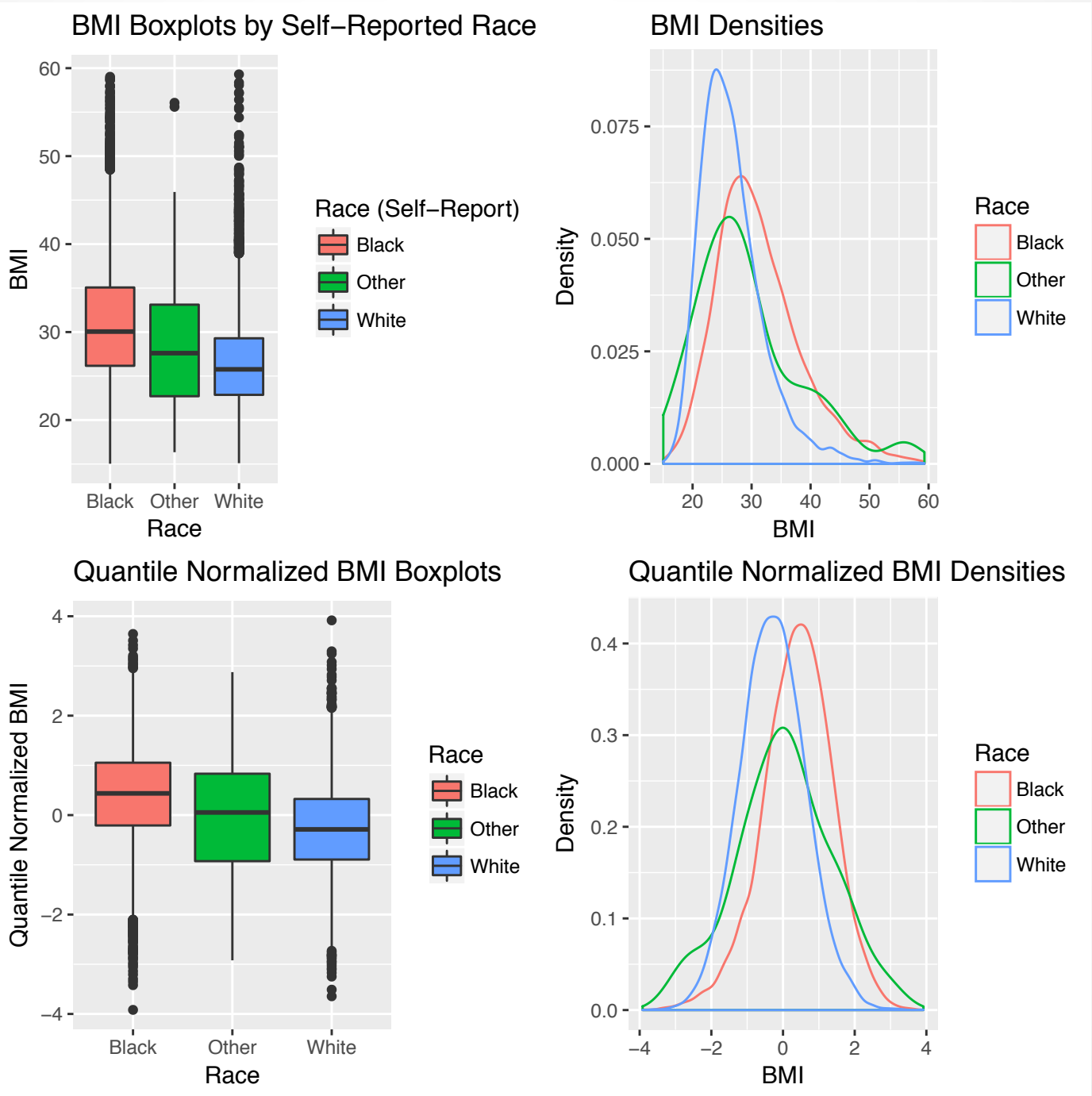
- Residual variance components of BMI for a few studies in TOPMed

TopMed Cohort Study	Phenotypic Residual Variances
Jackson Heart Study	35.44
CFS	52.33
Framingham Heart Study	13.14
Amish	12.19
COPDGene	26.61
HVH	61.31

Allowing for heterogeneity in variances: By Self-Reported Race

- There are limitations with modeling heterogenous variances by study.
- Some TOPMed studies have multiple ethnicities/ancestries.
- Also explored the differences in BMI distribution by self-reported race.

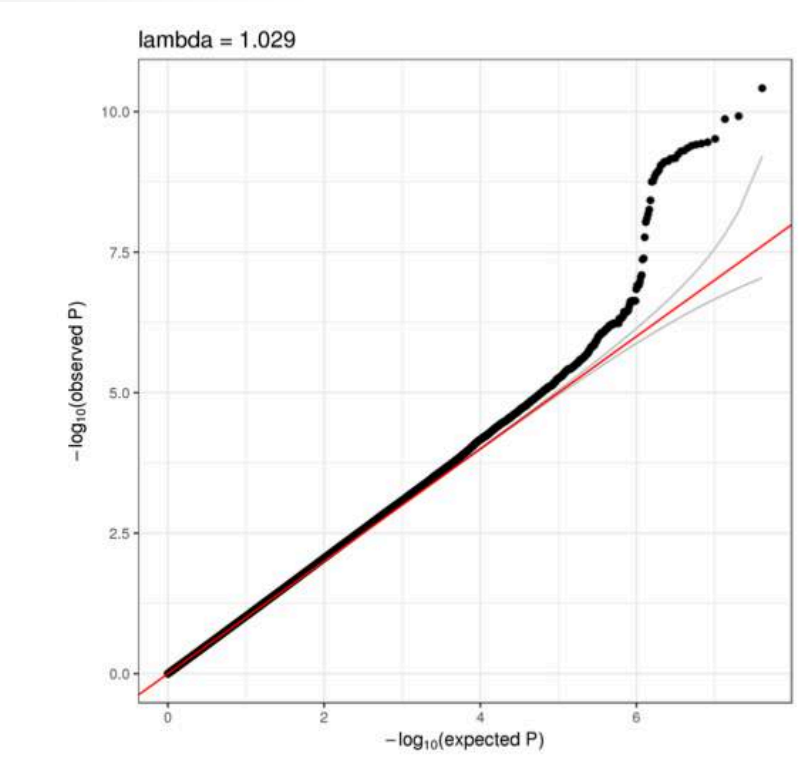
TOPMed BMI Distributions: By Race



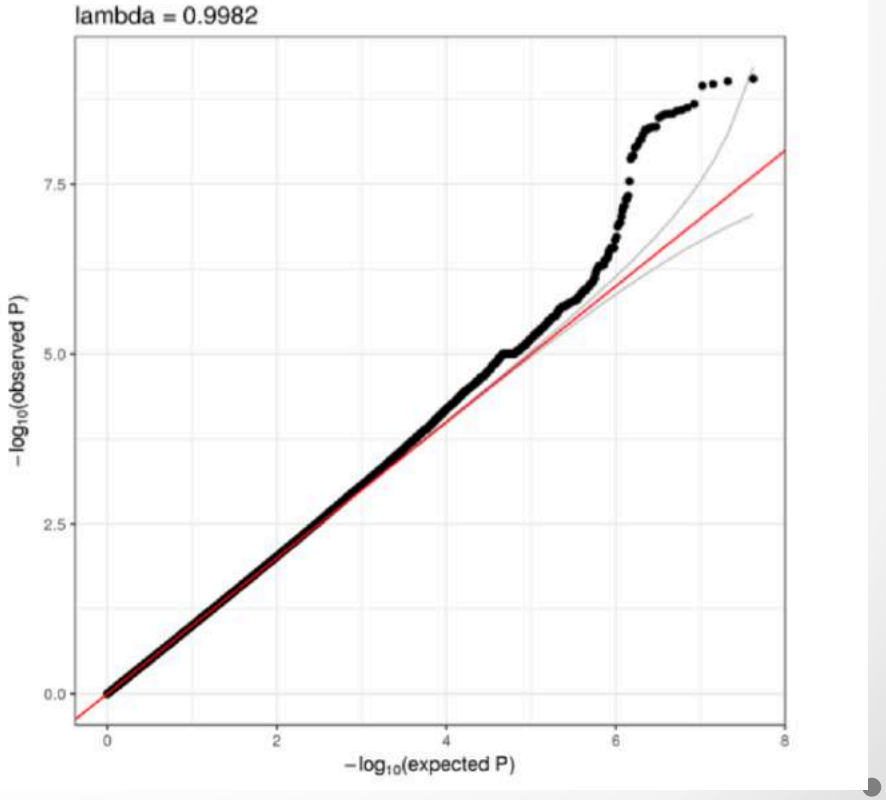
Heterogeneity in BMI Phenotypic Variances: Study vs. Self-Reported Race

- Association results for body mass index (BMI) allowing for heterogeneous phenotypic variances

Heterogenous Residual Variance by Study

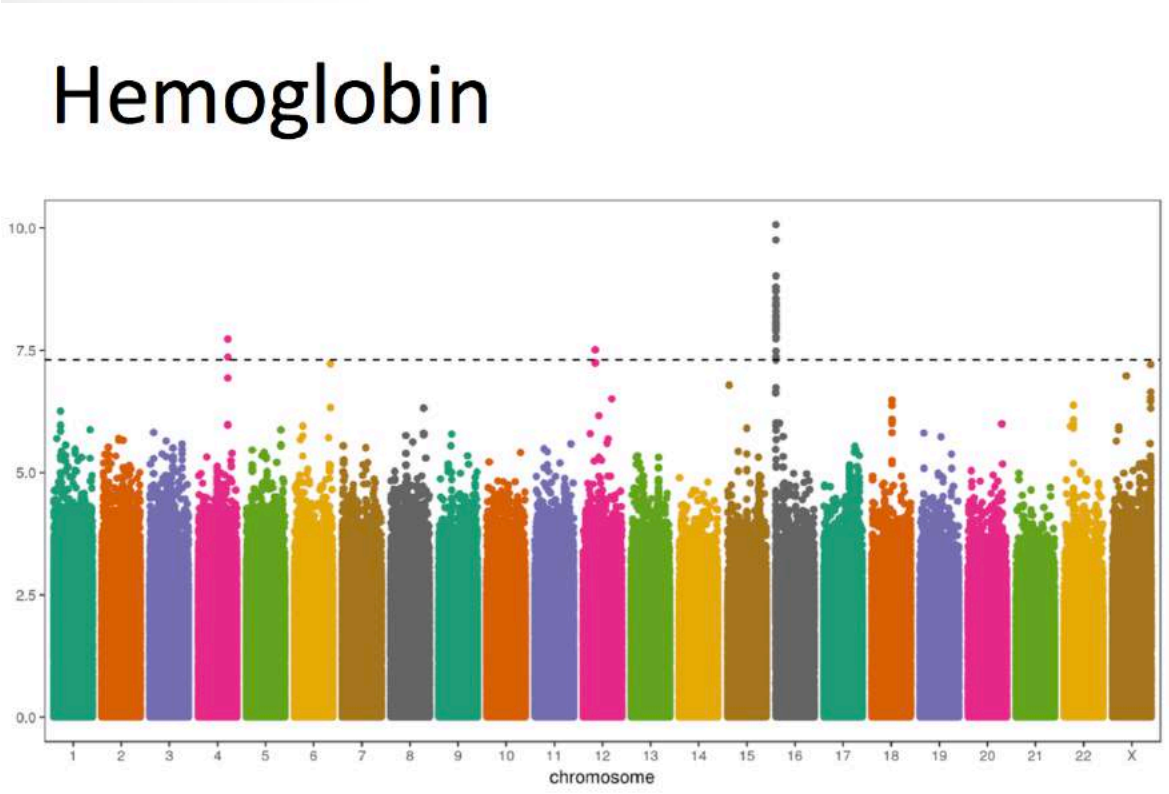


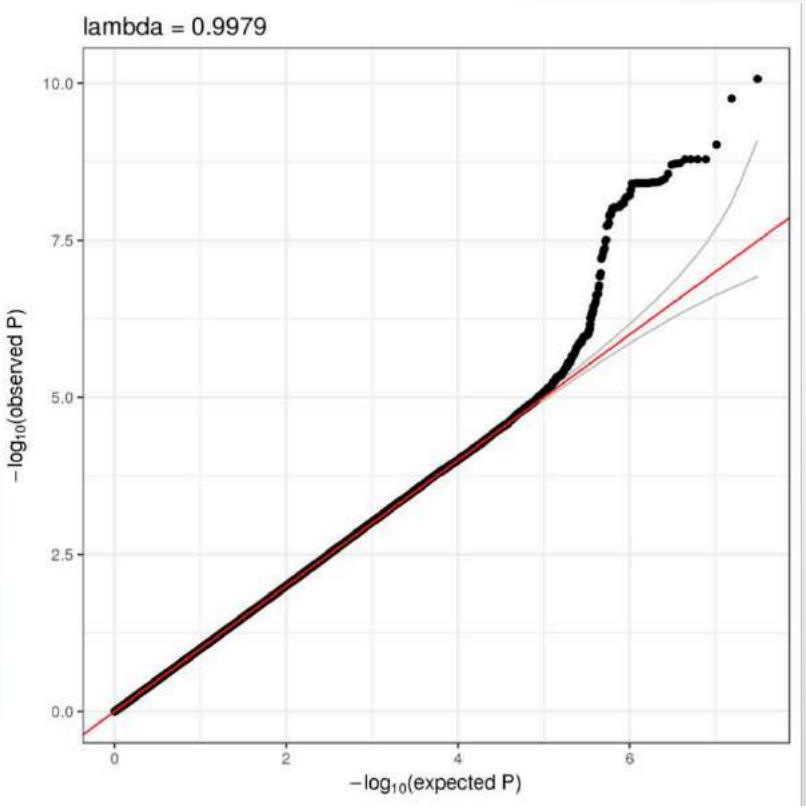
Heterogenous Residual Variance by Race



Hemoglobin Single Variant Tests: Mega-Analysis

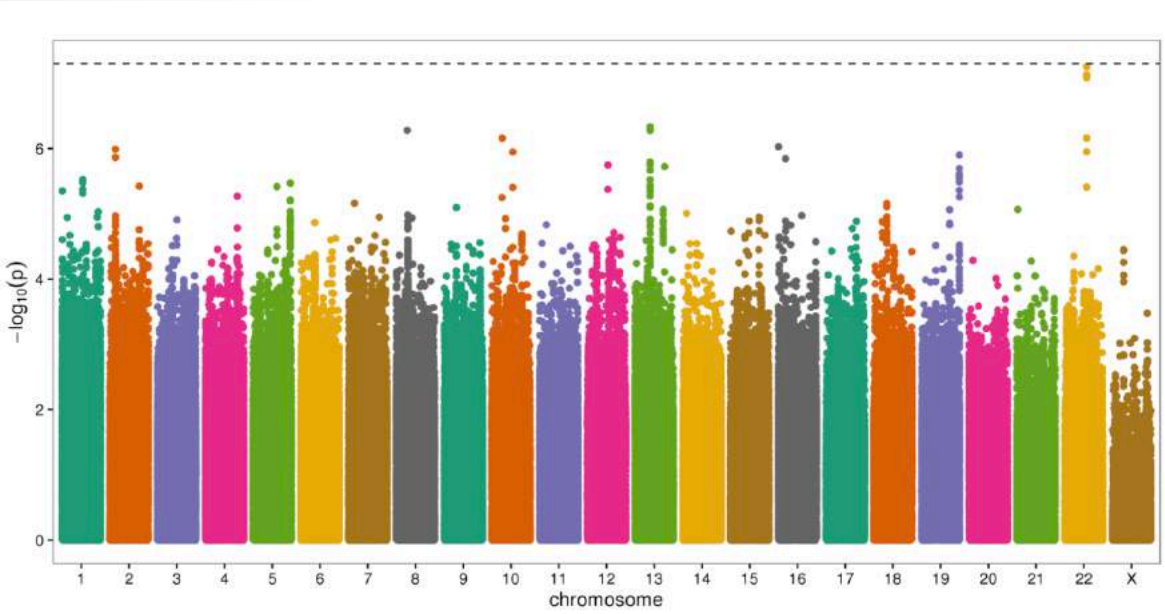
- Hemoglobin data results for three phase I TOPMed studies: Framingham Heart Study, Jackson Hearty Study, and the Amish
- Considered variants with minor allele count ≥ 10 in all studies combined: 29,342,569 variants tested for association

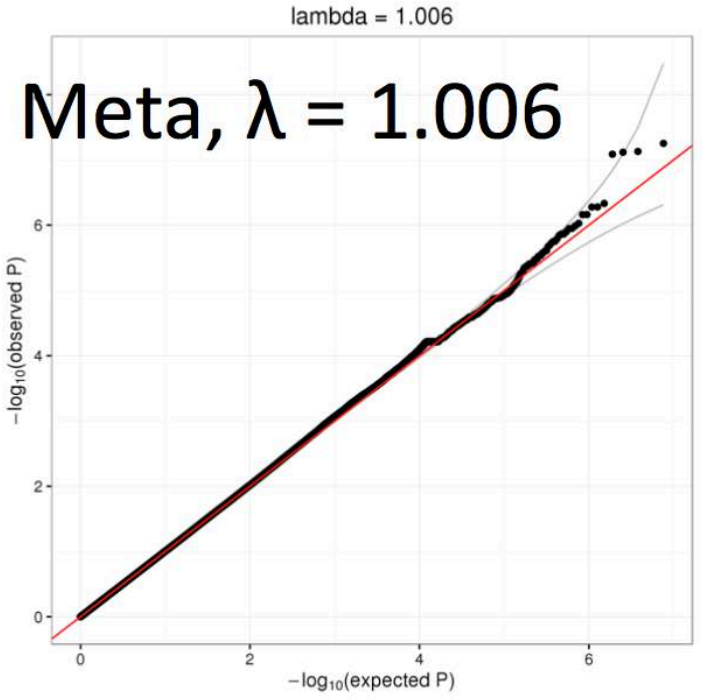




Hemoglobin Single Variant Tests: Meta-Analysis

- Association analysis conducted within each study separately
- Meta-analysis of the association results across all studies
- 7,570,518 variants tested for association
- No significant results





Association testing with rare variants in TOPMed

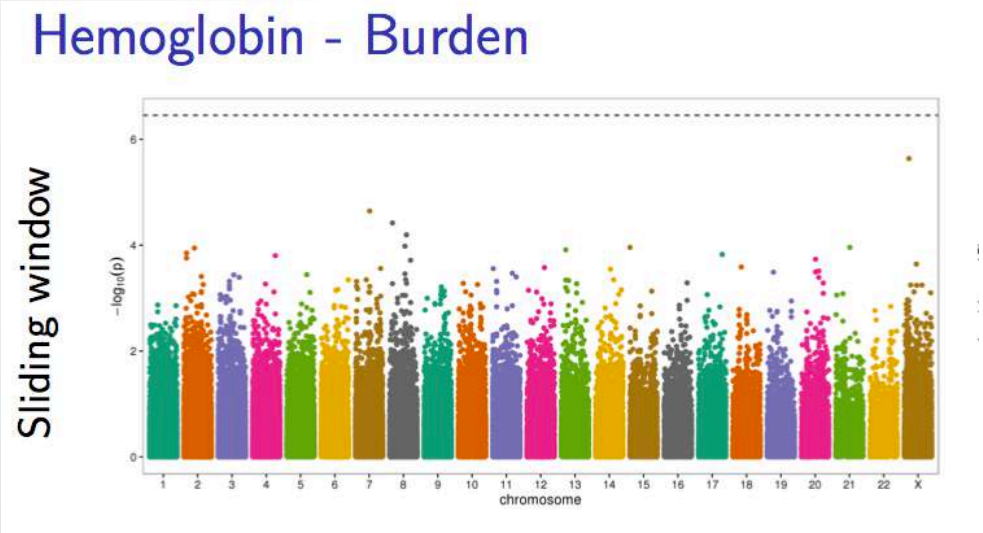
TOPMed Association Analysis with rare variants

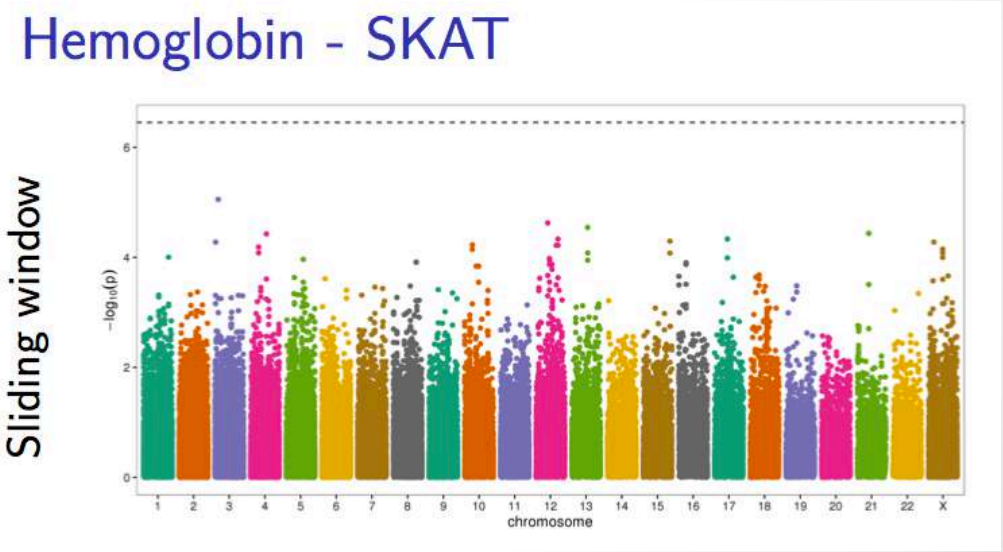
- There are an abundance of rare variants in the TOPMed WGS
- Phase 1 of TOPMed has 18,526 samples and 219,154,455 variants.
- 95,252,627 of the variants are singletons
- Extended widely used aggregate tests, such as SKAT and burden tests, for association testing of multiple rare variants in TOPMed.
- Allow for heterogenous variances

TOPMed Hemoglobin

Used a sliding window for the units:

- 50 kilobases
- 142,293 windows
- Considered rare variants only (MAF < 1%) in each window





SOFTWARE

- **GENESIS**: R software package is available from Bioconductor
- Installation in R:
  - **source("https://bioconductor.org/biocLite.R")**
  - **biocLite("GENESIS")**
- Current release of GENESIS:
  - **PC-AiR**
  - **PC-Relate**
- Recent release includes **LMM-OPS**

Acknowledgements

UW Genetic Analysis Center

Department of Biostatistics, University of Washington

Bruce Weir
Tim Thorton
Ken Rice
Sharon Browning
Brian Browning
Katie Kerr
Tamar Sofer
Adam Szpiro

Cathy Laurie
David Levine
Cecelia Laurie
Stephanie Gogarten
Adrienne Stilp
Caitlin McHugh
Quenna Wong
Xiuwen Zheng

TOPMed

Grant Funding:

NIGMS: P01 GM099568

NHLBI: R01 HL120393,
HHSN26820130005C

Special Acknowledgements

- Matthew Conomos, PhD; Arivale

- Ken Rice, PhD; UW Biostatistics

Special Acknowledgements

- Stephanie Gogarten, PhD; UW GAC

- Xiuwen Zheng, PhD; UW GAC

- Jen Brody; UW Cardiovascular Health Research Unit