Statistical Methods for Transcriptome-Wide Association Studies in Ancestrally Diverse Populations

Anna V Mikhaylova

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Timothy Thornton, Chair

Ali Shojaie

Michael Wu

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Methods for Transcriptome-Wide Association Studies in Ancestrally Diverse Populations

Anna V Mikhaylova

Chair of the Supervisory Committee:
Dr Timothy Thornton
Biostatistics

Transcriptome-wide association studies (TWAS) have become more commonly used in recent years. TWAS integrate genome-wide association studies (GWAS) with gene expression mapping studies in order to identify genes whose gene expression is associated with the phenotype. The main goals of TWAS are in providing insights into biological mechanisms underlying disease etiology and in helping interpret the results of GWAS. TWAS conducted in large-scale ancestrally diverse cohorts face multiple challenges, including the presence of population structure, known or cryptic relatedness and heterogeneity in phenotypic distributions across subgroups. There is a dearth of statistical methodology available to researchers that addresses the aforementioned issues. In this dissertation, we evaluate the performance of existing TWAS methods in ancestrally diverse populations and identify their limitations. We then develop new statistical methodology that addresses these limitations. We validate the performance of the novel methods in extensive series of simulations as well as in applications to large cohorts of ancestrally diverse populations from the Trans-Omics for Precision Medicine (TOPMed) program.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

ACKNOWLEDGMENTS

The completion of this dissertation would not be possible without the guidance and encouragement from my advisor Dr. Timothy Thornton. I would like to express my deepest gratitude for his mentorship and support all these years. I would like to acknowledge the other mentors I have had during my years at the University of Washington, Drs. Matthew Conomos, Amalia Magaret and Ellen Wijsman. Finally, I would like to thank my classmates for providing invaluable support throughout my graduate studies.

DEDICATION

This dissertation is dedicated to my parents. For their endless love, support and encouragement.

Chapter 1

INTRODUCTION

For more than a decade, genome-wide association studies (GWAS) have been widely used and have successfully identified tens of thousands of genetic variants associated with a variety of complex traits and diseases. Many of these associations have been replicated in independent studies which suggests that they are valid associations. However, the biological mechanisms underlying trait-variant associations generally remain unclear due to several factors [20]. First, an association between a locus and a phenotype does not indicate which variant is causal. In fact, the vast majority of GWAS-identified variants are non-causal but are in strong linkage disequilibrium (LD) with the causal loci, which are generally unknown. Causal and non-causal variants segregate on the same haplotype and have statistically indistinguishable associations with the trait of interest. Second, more than 90% of disease-associated variants are located in non-coding regions of the genome, so they do not specify which gene is affected by the causal variant. At the same time, these variants may affect gene expression by having effects on gene regulation, transcription, and splicing.

Several studies have shown that GWAS-identified variants are enriched for expression quantitative trait loci (eQTLs), the genetic variants that affect gene expression [47, 48, 2]. In particular, eQTLs may affect disease risk by altering genetic regulation of one or more genes and, therefore, help to prioritize likely causal variants among the plethora of variants within genomic regions identified by GWASs. However, eQTL studies have similar pitfalls as GWASs – identified variants are often not causal but are in LD with the true eQTLs.

Transcriptome repositories, such as the Genotype-Tissue Expression (GTEx) project [34], and discovery of eQTLs motivated the development of methods for transcriptome-wide association studies (TWAS) [53] which evaluate associations between expression levels of each gene and a phenotype of interest. Early TWA studies integrated data from eQTL and GWA studies by evaluating whether the same genetic variants affected gene expression and a trait of interest under a single causal variant model.

Because of the limited availability of samples with measured gene expression, genotype data and trait values, more recently researchers proposed a series of methods that leveraged predicted gene expression; thus, predicted gene expression can be used instead of measured gene expression to perform TWAS. Under this approach, prediction models are trained on a reference dataset for which both transcriptome and genotype data are available. For every gene there is a set of *cis* variants (that are within a certain window, usually 1Mb, of the transcription region) from which the best predictors are selected by one of the machine learning methods. Thus, predicted gene expression is a weighted linear combination of variants.

This approach offers multiple advantages. First of all, TWAS can increase power over a traditional GWAS in cases where the relationship between causal variants and a phenotype is mediated through gene expression, i.e. where genetic variants regulate gene expression and gene expression affects the trait of interest. Secondly, it reduces the multiple hypothesis burden because instead of testing millions of variants, we only have to test a maximum of around twenty thousand genes across the genome. Finally, TWAS can aid in understanding the mechanisms underlying trait-variant relationships, disease susceptibility, and drug response. Since these methods for TWAS can be applied to large biobanks with genotype data and used to identify genes whose expression levels are associated with disease risk, they can help in facilitating translational genomics medicine.

However, TWAS methods have some limitations. When the expression data are

not available in the right tissue or the disease risk is not mediated through expression, TWAS can be underpowered. Another often overlooked issue is the lack of genetic diversity in the datasets used for training predictive modeling. The datasets used in training prediction models for TWAS are highly biased towards subjects of European ancestry. Differences in genetic architecture and genotype frequencies across diverse populations present potential problems for applications of European ancestry-based models in non-European ancestry populations. Previous studies evaluating polygenic risk scores showed biases and poor prediction accuracy when the ancestries of the training and target populations were different [39]. However, the portability of TWA prediction models across diverse populations has not been thoroughly investigated.

While there exist multiple statistical methods for conducting GWAS in samples with population structure and relatedness, there is a lack of such methods for TWAS. Most TWA methods make simplifying assumptions of homogeneity and independence of the samples, so they are not suitable for samples with population structure and/or familial relatedness. Accurately modeling population and pedigree structure can prevent spurious association and increase statistical power to detect true association.

Finally, phenotypic distributions often vary across populations, resulting in different means as well as variances. In modern omics studies, it is common to pool participant-level data from multiple study centers and analyze them together in a single analysis. However, unaccounted for population variance structure can lead to decreased statistical power and increased false positives. Therefore, it is important to account for differential phenotype variances by study or ethnic group to correct the bias and improve power of TWAS.

In this dissertation, we address existing challenges of performing TWAS in diverse populations. The specific aims for this dissertation are as follows:

(1) To evaluate prediction quality of an existing method of predicting gene expression in diverse populations;

(2) To develop a novel method of predicting gene expression in diverse populations;

(3) To develop a method to conduct transcriptome-wide association studies in samples with population structure and familial relatedness; and

(4) To develop a method to conduct transcriptome-wide association studies in samples with phenotypic variance heterogeneity among groups.

1.1 PrediXcan performance in diverse populations

Using genetic data to predict gene expression has garnered significant attention in recent years. PrediXcan [21] has become one of the most widely used gene-based methods for testing associations between predicted gene expression values and a phenotype, which has facilitated novel insights into the relationship between complex traits and the component of gene expression that can be attributed to genetic variation. The gene expression prediction models for PrediXcan were developed using supervised machine learning methods and training data from the Depression Genes and Networks (DGN) study and the Genotype-Tissue Expression (GTEx) project. The majority of subjects in these studies were of European descent. However, many recent genetic studies include samples from multi-ethnic populations.

In Chapter 3 of this dissertation, we evaluate the accuracy of PrediXcan for predicting gene expression in diverse populations. Using transcriptomic data from the GEUVADIS (Genetic European Variation in Disease) RNA sequencing project and whole genome sequencing data from the 1000 Genomes project, we evaluate and compare the predictive performance of PrediXcan in an African population (Yoruban) and four European ancestry populations for thousands of genes. We evaluate a range of models from the PrediXcan weight databases and use Pearson's correlation coefficient to assess gene expression prediction accuracy with PrediXcan.

From our evaluation, we find that the predictive performance of PrediXcan varies substantially across populations from different continents (F-test p-value $< 2.2 \times 10^{-16}$).

Prediction accuracy is lower in the Yoruban population from West Africa compared to the European-ancestry populations. Moreover, not only did we find differences in predictive performance between populations from different continents, we also found highly significant differences in prediction accuracy among the four European ancestry populations considered (F-test p-value $< 2.2 \times 10^{-16}$). While there is variability in prediction accuracy across different PrediXcan weight databases, we also found consistency in the qualitative performance of PrediXcan for the five populations considered, with the African ancestry population having the lowest accuracy across databases.

1.2 Predicting gene expression in diverse populations

Several TWA methods have been developed in the last few years. However, the majority of training data for TWA methods consists of European-ancestry populations, whereas many genetic studies include populations of diverse ancestry. Diverse populations pose additional challenges to developing accurate transcriptome prediction models. For example, populations of African descent have different genetic architecture and allele frequencies, greater genetic variation, and less linkage disequilibrium, compared to European ancestry populations. Additionally, a large proportion of gene expression phenotypes are differentially expressed between European ancestry populations and African Americans, as has been previously shown in the literature [63]. These factors may bias and lower accuracy of existing transcriptome prediction models when applied to diverse populations.

Prediction model training can be performed in pooled datasets, where multi-ethnic samples are combined into one training dataset and analyzed together. Alternatively, training sets can be split by ancestry and analyzed separately to produce ancestry-specific prediction models. Both of these approaches have drawbacks. The first approach has increased power to identify expression quantitative trait loci (eQTLs) that are common in all populations but may miss population-specific eQTLs. The

second approach has decreased power to detect variants that are common in multiple populations, since it is not leveraging information across samples. Thus, identifying and accounting for shared and population-specific eQTLs can lead to improved accuracy of gene expression imputation models.

In Chapter 4, we develop a novel method of predicting gene expression from genotype data for diverse populations that decomposes eQTL architecture into shared and population-specific components. We validate our method in a dataset of African American and European-ancestry individuals with whole genome sequencing data and peripheral blood mononuclear cells RNA-seq measurements from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort. We show that our method is more flexible and robust than existing methods that are trained on datasets that do not take into account population labels of each participant.

1.3 TWAS in related samples with population structure

Linear mixed models (LMMs) have become a popular method of choice for GWAS in samples with population structure and relatedness. However, most TWA methods assume independence and population homogeneity of samples which can result in increased false positive associations and reduced power to detect true associations. To address this issue, we extend the linear mixed model methodology for GWAS with complex sample structure to transcriptome-wide association studies.

In Chapter 5, we propose a new LMM approach for transcriptome association testing in samples with population structure and relatedness. We account for population structure by including ancestry representative principal components as fixed effect covariates and we account for relatedness by modeling the trait covariance structure with a genetic relationship matrix (GRM). Additionally, we include a random effect that accounts for predicted transcriptome correlation, a transcriptome relationship matrix (TRM). Variance components of the random effects are calculated via average information restricted maximum likelihood (AI-REML) with two empirical relationship

matrices, one for polygenic background and one for the transcriptome.

In simulation studies, we demonstrate that, by inclusion of an additional random effect for the TRM, our method can provide an improvement, in terms of type I error and power, over an LMM method that includes a single variance component with a GRM. We apply our method to ancestrally diverse sample of individuals from the TOPMed program with whole genome sequencing data. We perform a TWAS using a white blood cell (WBC) phenotype. We use PrediXcan models to impute whole blood gene expression and then regress the WBC trait onto imputed gene expression values. Our method detects previously identified genes associated with the WBC and, in some regions of the genome, helps prioritize genes that are potentially causal.

1.4 TWAS in samples with heterogeneous variance groups

In the past several years, transcriptome-wide association studies have garnered a lot of interest. Many of them have been conducted in large-scale cohorts that include participants of diverse ancestral backgrounds from multiple studies. In such studies, participant data is often pooled and analyzed together, which increases the power to detect associations. However, phenotypic variances often vary across different subgroups – a phenomenon we refer to as population variance structure. Unaccounted for, it can results in decreased power and increased false positives rate.

In Chapter 6, we describe an approach to accounts for population variance structure across subgroups in TWAS. We extend the approach proposed in Chapter 5 by allowing for heterogeneous residual variances among subgroups in the linear mixed model. In simulation studies, we compare the performance of the LMM method allowing for heterogeneous variances vs the LMM method not allowing for them. We demonstrate that accounting for population variance structure improves statistical properties of tests. It is properly calibrated in terms of type I error rate. Additionally, it yields improved statistical power to detect gene-trait associations over an LMM method that does not allow for heterogeneous variances among subgroups. Finally, we validate our

method in real data by performing a TWAS of white blood cell phenotype measured on the participants of the TOPMed program.

Chapter 2

TWAS BACKGROUND

In this chapter, we will describe general TWAS framework and give an overview of TWA methods and their differences. We will discuss TWA methods' modeling assumptions and choice of reference panels. In particular, we will focus on key features of PrediXcan, a TWA method that is widely used for predicting gene expression from genotype data, and describe PredictDB, a database of PrediXcan prediction weights. We will conclude with addressing some challenges and limitations of interpreting TWAS results.

2.1 General TWAS framework

Since the majority of GWAS datasets do not include transcriptomic data, we can leverage reference panels (cohorts with expression and genotype data) to identify functional gene-trait associations. The general TWAS framework consists of the following three steps (Figure 2.1):

1) training predictive models of expression from genotype on a reference panel;

2) using these models to predict expression for individuals in the GWAS cohort; and

3) correlating this predicted expression with the phenotype of interest to identify gene-trait associations.

Consider a reference panel with sample size $\mathcal{N}$. In each tissue separately, prediction models are constructed from expression levels for a given gene and a set of common

variants, denoted $\mathcal{S}$, that are within a predefined window of the transcription region. Genotypes are assumed to have additive genetic effects on gene expression levels:

$$\mathbf{y_g} = \sum_{s \in \mathcal{S}} \mathbf{x_s} w_s + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{y_g}$ is a vector of expression values for a gene g, $\mathcal{S}$ is a set of variants in the vicinity of the gene, $\mathbf{x_s}$ is the vector of numbers of reference alleles for s markers, w_s is the effect size of variant s on gene expression, and $\boldsymbol{\epsilon}$ is the contribution of environmental effects and other factors.

The set of variants $\mathcal{S}$ typically may contain a thousand or more variants, so $|\mathcal{S}| > |\mathcal{N}|$, and thus multivariate regression can not be directly fit. Instead, the effect sizes are estimated using penalized regression methods, such as lasso and elastic net. As a result, s' variants are selected as good predictors of the gene expression, such that $s' \in S'$ and $S' \subset S$, and their corresponding estimated weights $\hat{w}_{s'}$ are stored in a database of prediction models.

Weights stored in PredictDB can be used to predict gene expression in a dataset where only genotype data is available. Gene expression is predicted as a linear combination of predictor variants and their corresponding weights:

$$\hat{\mathbf{y}}_\mathbf{g} = \sum_{s' \in \mathcal{S}'} \mathbf{x_{s'}} \hat{w}_{s'} \tag{2.2}$$

The predicted gene expression values together with other relevant covariates, can be used to perform TWAS testing genes one at a time:

$$\mathbf{y_t} = \beta_0 + \beta_1 \hat{\mathbf{y}}_\mathbf{g} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{y_t}$ is the vector of phenotypic values, $\hat{\mathbf{y}}_\mathbf{g}$ is the vector of imputed expression values for gene g, $\mathbf{Z}$ is a matrix of covariates, and $\boldsymbol{\epsilon}$ is the contribution of other factors.

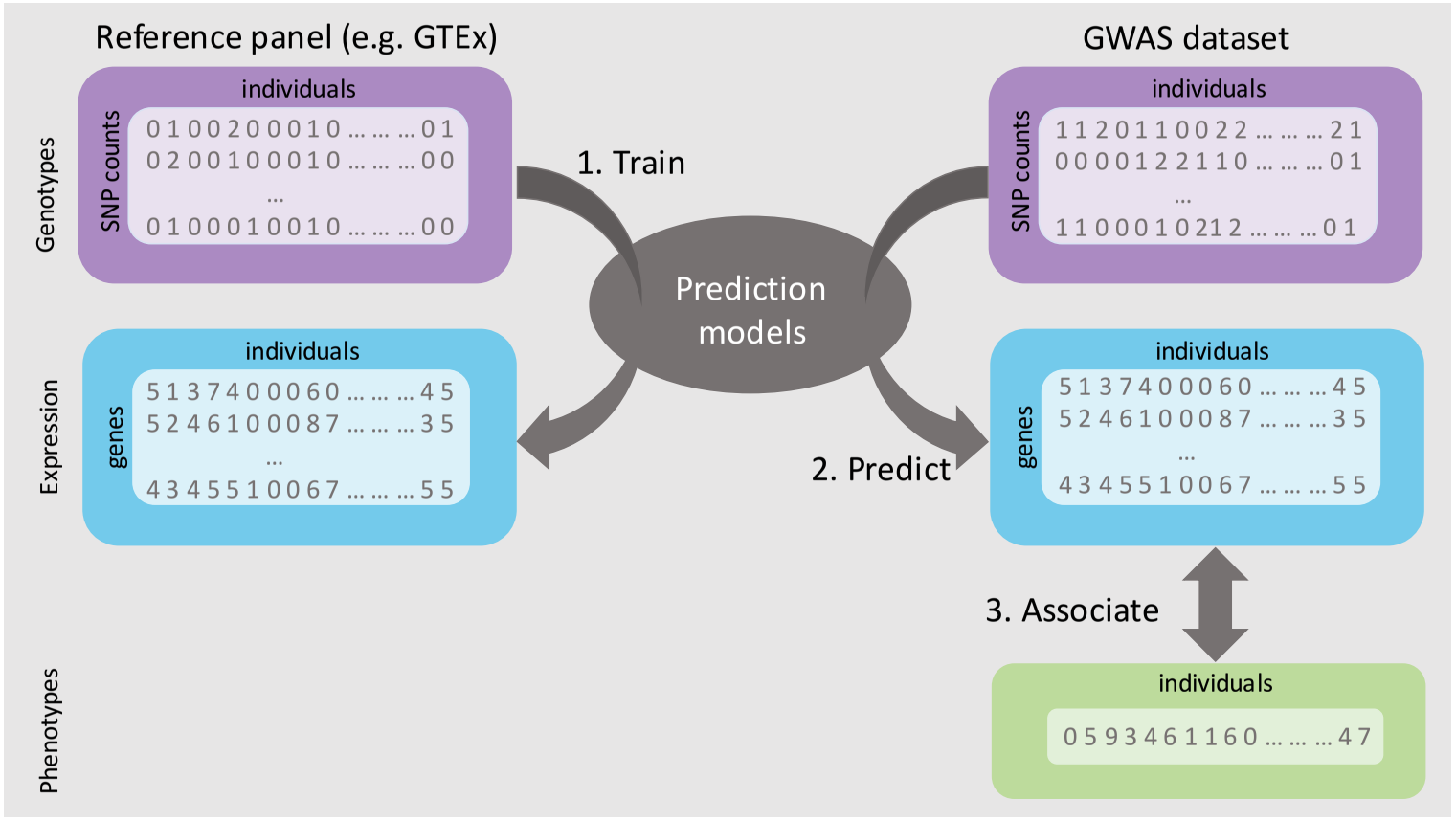


Figure 2.1: **An overview of TWAS.** TWAS procedure includes: (1) training a transcriptome prediction models from genotype or whole-genome sequencing data on a reference panel; (2) applying transcriptome prediction models to individuals in an independent cohort; and (3) testing for an association between the predicted transcriptome and a trait trait.

2.2 TWA approaches

Depending on the choice of frequentist or Bayesian approach, TWA methods use different modeling assumptions on the estimation of weights w_s from Equation 3.1. The first TWA method PrediXcan, introduced by Gamazon et al [21], uses a frequentist framework. The model uses elastic net which imposes a lasso and ridge penalties on w_s in Equation 3.1. PrediXcan selects a sparse set of *cis*-variants (within a 1Mb window of the transcription site) with non-zero effects on gene expression levels. The derived estimates are stored in a database PredictDB and used to predict the gene expression in equation 3.2 and perform the association analysis in Equation 3.3 to obtain the effects of gene expression on the phenotype.

PredictDB includes weights derived from various transcriptome data, tissues and sets of variants. The original PrediXcan paper used the DGN study ($n = 922$, all European ancestry) whole blood RNA-seq data and two sets of imputed variants, from 1000 Genomes Project and only the HapMap subset, as training data. Later releases included weights for multiple tissues trained on the GTEx v6p (about 85% European, depending on tissue) and GTEx v7p (all European ancestry) and had models trained on both 1000 Genomes and Hapmap variant sets. The most recent set of weights was trained on GTEx v8 (all European ancestry) using fine-mapped variants and borrowing information across tissues. Using multivariate adaptive shrinkage in R (*mashr*), Barbeira et al. estimated the models' effect sizes by leveraging cross-tissue variations [4]. We refer to these models as *mashr* models.

Bayesian models tend to be more flexible but more computationally intensive, compared to sparse models, such as elastic net and lasso. FUSION [25] is a TWA method that uses a Bayesian sparse linear mixed model (BSLMM) to estimate the weights w_s. It assumes that variant effect sizes follow a mixture of two normal distributions:

$$w_s \sim \pi\mathcal{N}(0, \sigma_a^2 + \sigma_b^2) + (1 - \pi)\mathcal{N}(0, \sigma_b^2)$$

Thus, BSLMM splits the variants into two groups: a small group of variants π with larger effect sizes and larger variance $\sigma_a^2 + \sigma_b^2$ and a large group of variants $(1 - \pi)$ with smaller effect sizes and smaller variance σ_b^2.

An obstacle one might encounter in training TWA models is that individual level GWAS data are often not available because of consent and privacy concerns. For many large studies, only summary statistics are very easy to access. GWAS summary statistics are often stored in datasets with open access for hundreds of thousands of individuals, which can boost statistical power and improve prediction accuracy. Moreover, summary-based TWA methods tend to be faster and require less computing memory storage, compared to individual level data methods. PrediXcan was extended to a summary statistics version (S-PrediXcan [5]), while other methods, such as UTMOST [27], the Factored QTL (fQTL) [52] and FOCUS [36] proposed to directly use summary statistics without an initial individual-level data model.

Another extension of early TWA methods stems from leveraging gene expression data from multiple tissues at the same time. Univariate TWA methods model one gene at a time from a single tissue. Such methods include PrediXcan [21], FUSION [25], DPR [76], TIGAR [44], CoMM [73], and PMR [75]. Multivariate TWA methods leverage the co-regulation of *cis*-variants on multiple tissues. Such methods include TisCoMM [60] (an extension of CoMM), UTMOST [27], MultiXcan [6], and fQTL [52]. Multivariate TWA methods offer advantages due to sharing of eQTLs across tissues and correlation in transcriptome between different tissues. However, in certain cases, single-tissue methods may be more suitable when one is investigating a specific genetic condition and tissue-specific effects are of interest.

2.3 TWAS challenges and limitations

TWAS offer multiple advantages in prioritizing potentially causal genes and reducing multiple testing burden, in comparison with GWAS. However, TWAS are still prone to spurious associations. There is a number of issues that make distinguishing between

the true and false positive TWAS associations a difficult task. We highlight some of the challenges of TWAS and discuss possible solutions to them.

The first choice that needs to be made before training transcriptome prediction models is selecting appropriate transcriptome reference panel in terms of tissue and ancestral composition of the training sample. A trait's most relevant tissue may not be known or unavailable in sufficiently large reference panels. In that case, a researcher might decide to choose the closest related tissue or aggregate information across multiple tissues [71]. Additionally, heavily relying on European ancestry-based reference panels has been shown to lower the accuracy of transcriptome predictions [29, 41] in diverse populations. Development of ancestry-specific TWA methods and gathering more diverse transcriptome datasets remain open areas of research.

Similarly to GWAS that identify blocks of variants in LD instead of a single variant, TWAS frequently identify multiple genes in proximity with each other. In such cases, interpretation of TWAS findings is especially difficult. Wainberg et al. [71] investigated how gene co-regulation can lead to multiple hit loci. In TWAS, it is important to distinguish total expression correlation among among individuals and predicted expression correlations due to sharing of predicted variants. Predicted gene expression of a pair of genes can be due to the same eQTLs or two different eQLTs in LD regulating each one of the genes. Both types of correlation can lead to spurious results and cannot be distinguished by TWAS.

In recent years, there have been attempts to address co-regulation and fine mapping in TWAS. Fine-mapping of causal gene sets (FOCUS) models multiple genes at a time by calculating predicted expression correlations and using them to assign genes posterior probabilities of causality [36]. Even though fine mapping can help identify a credible set of genes in a locus, it does not necessarily single out the most biologically relevant gene.

Chapter 3

GENE EXPRESSION PREDICTION IN DIVERSE POPULATIONS

3.1 Introduction

In the past decade, genome-wide association studies (GWAS) have identified thousands of genetic variants significantly associated with a wide range of human phenotypes [64, 70, 46, 35]. The vast majority of these studies, however, were conducted in samples from European ancestry populations [45, 11, 54, 55, 26, 8]. Differences in allele frequencies, genetic architecture, and linkage disequilibrium (LD) patterns across ancestries suggest that GWAS discoveries can fail to generalize across populations, and recent publications have provided compelling evidence that GWAS findings often do not transfer from European populations to other ethnic groups [32, 1]. For example, Carlson et al. analyzed multi-ethnic data from the PAGE Consortium and concluded that some variants identified in GWAS in European ancestry populations had different magnitude and direction of allelic effects in non-European populations and the differential effects were more persistent in African Americans [13]. Moreover, genetic risk prediction models derived from European GWAS were found to be unreliable when applied to other ethnic groups [13]. Martin et al. examined the impact of population history on polygenic risk scores and demonstrated that they can be biased and confounded by population structure. [39]. Since genetic risk prediction accuracy depends on genetic similarity between the target and discovery cohorts, Martin et al. advised against interpreting the scores across populations and recommended computing them in genetically similar cohorts.

Associations between genetic variation and molecular traits, such as gene expres-

sion, have advanced our understanding of the mechanisms underlying trait-variant associations [2, 47, 68]. Prior studies have shown that a large proportion of GWAS variants identified for complex traits are expression quantitative trait loci (eQTLs), i.e., they play a role in regulating gene expression [48]. Thus, eQTLs can aid in prioritizing likely causal variants among the ones identified by GWAS, especially if they are found in non-coding regions, and can help uncover the mechanisms by which genotypes influence phenotypes [2]. As a result, having three types of data – genotype, phenotype and gene expression – on the same set of subjects can be advantageous for improved understanding of the relationships between complex traits, the genetic backgrounds of study subjects, and the underlying biological processes. However, collecting all of these different types of data on the same study subjects is often not feasible due to cost and tissue availability. Additionally, eQTL studies have the same pitfalls as GWASs – the majority of the detected eQTLs are not causal, but may be in LD with causal variants. Similar to variants identified through GWAS, eQTL findings might fail to replicate in diverse populations due to differential LD patterns across populations [28].

Recently, there has been increased interest in integrating eQTL studies and GWASs for improved complex trait mapping. PrediXcan [21] is one of the most widely used integrative methods for testing associations between a phenotype and gene expression values predicted from SNP genotyping or sequencing data. PrediXcan can have increased power over traditional GWAS methods, particularly when differential changes in gene expression is an intermediary stage of the causal pathway from genetic variation to the outcome of interest. A useful feature of PrediXcan (and other similar methods) is the ability to obtain predicted gene expression values on study subjects when tissue types relevant to phenotypes are not available. We now give a very brief overview of the PrediXcan method. PrediXcan uses machine learning methods and large reference datasets consisting of both genotype and trascriptome data for supervised training to construct prediction models for expression of each gene. With PrediXcan, genetic

training data is restricted to common *cis*-variants that are within 1Mb upstream and downstream from the transcription region [21]. Gene-specific derived SNP weights from the prediction models are then stored in databases, with separate sets of weights for different tissue types. Using these weights, PrediXcan allows for the prediction of gene expression values for study subjects with available genotype data, where predicted expression values are computed as a weighted linear combination of SNP dosages. Finally, the predicted expression values can then be used to test for associations with a phenotype of interest. By conducting tests on gene expression obtained from an aggregation of variants, PrediXcan dramatically reduces multiple testing burden as compared to single variant association testing.

Previous studies have reported differences in gene expression levels across diverse populations from the HapMap3 project, noting that 77% of eQTLs are population specific and only 23% are shared between two or more populations [63, 19]. More distantly related populations have more differentially expressed genes than closely related populations, although this can often be explained by the expression of different gene transcripts across populations [30]. One potential limitation of PrediXcan, however, is that the method may not perform well in diverse populations, as the supervised learning for PrediXcan was conducted using data from the Depression Genes and Networks (DGN) and the Genotype-Tissue Expression (GTEx) Project – both of which consist primarily of European-ancestry subjects [7, 34]. Many genetic studies include samples from multi-ethnic populations, and understanding the accuracy of gene expression prediction with PrediXcan across populations is of interest to many genetic researchers .

Recent works have evaluated the performance of PrediXcan in diverse populations [31, 23]. Li et al. evaluated PrediXcan whole-blood prediction models and investigated the factors that influence prediction accuracy using the Yoruban (YRI) and European (CEU) samples from the Genetic European Variation in Health and Disease (GEUVADIS) [30] cohort.They reported PrediXcan performance to be unsatisfactory

for most genes due to predicted gene expression values not correlating well with the observed values[31]. Differences in prediction accuracy with PrediXcan between the YRI and CEU, however, were not directly compared. Gottlieb et al. investigated the performance of PrediXcan for a small subset of 116 genes that are in the warfarin-response pathway in European and African American samples where they concluded that PrediXcan performed poorly in African Americans [23].

Here, we evaluate the predictive performance of PrediXcan both across and within continental populations using thousands of genes across the genome. Using the GEUVADIS transcriptome data and whole genome sequencing data from the 1000 Genomes Project [30, 3], we consider four closely related European ancestry populations and one African population. In our analysis, we test the null hypotheses of (1) no difference in prediction accuracy with PrediXcan across European and African continental populations; and (2) no difference in predictive performance among the four European derived populations. We obtain predicted gene expression levels using seven PrediXcan weight databases derived from whole blood and lymphoblastoid cell lines (LCL) transcriptome data for each individual. To evaluate differences in prediction accuracy among the populations, we use a linear mixed effects model framework where Pearson's correlation coefficients for observed and predicted gene expression levels are included as the outcome and the populations are included as categorical predictors. In addition, we evaluate the utility of whole-blood-based models when making predictions for LCL expression data. We find from our analyses that accuracy of PrediXcan for gene expression prediction not only differs between European and African continental populations, but also among closely related populations of European ancestry. Furthermore, prediction accuracy with PrediXcan is the lowest in Africans across all seven weight databases considered, which further illustrates the need to develop new predictive models using training data composed of individuals who have similar ancestry to the target sample for which gene expression is to be predicted [41].

3.2 Materials and methods

3.2.1 Datasets

We obtained gene expression data from the GEUVADIS Consortium and whole genome sequencing data from the 1000 Genomes Project. The gene expression data consisted of RNA sequencing on lymphoblastoid cell line (LCL) samples for 464 individuals from five populations. Of these, 445 subjects were in the 1000 Genomes Phase 3 dataset, including 358 subjects of European descent and 87 subjects of African descent. European samples included: Utah residents with Northern and Western European ancestry (CEU, $n = 89$), British individuals in England and Scotland (GBR, $n = 86$), Finnish in Finland (FIN, $n = 92$), and Toscani in Italy (TSI, $n = 91$). African samples included individuals of African descent from Yoruba in Ibadan, Nigeria (YRI, $n = 87$). Gene expression measurements were available for 23,722 genes.

We used seven PrediXcan weight databases: DGN whole-blood (further referred to as DGN), GTEx v6 1KG whole blood, GTEx v6 1KG LCL, GTEx v6 HapMap whole blood, GTEx v6 HapMap LCL, GTEx v7 HapMap whole blood (GTEx WB), and GTEx v7 HapMap LCL (GTEx LCL). The databases were downloaded from http://predictdb.org/.

3.2.2 Filtering procedure for poorly predicted genes

Linear regression models were used to identify genes whose predicted values were not associated with the observed values at significance level of 0.05 in order to filter out genes that have poor prediction accuracy across all subjects. For each gene, we fit a linear regression model with observed gene expression as the outcome and predicted gene expression as the predictor of interest. A Wald test was used to assess significance of the coefficient for each gene in the linear model. Genes with corresponding p-values that were higher than a nominal significance level of 0.05 were identified and labeled as "poorly predicted".

We then calculated Pearson's correlation coefficient, r, between observed and predicted expression values for every gene, in each population separately. A few genes had the same predicted gene expression levels across all subjects. Since we could not calculate the correlation if one of the variables was constant, we excluded those genes. Thus, for every gene considered there were five Pearson's correlation coefficients, one for each population. Note that we used r instead of the square of Pearson correlation, r^2, in order to take directionality of correlation into account when assessing predictive performance. We found that using r^2 as a measure of predictive accuracy can be misleading as there were genes for which predicted and observed expression values had a significant negative correlation.

It should be noted that we also performed an evaluation of the performance of PrediXcan without doing any filtering of genes in order to assess the impact on the analysis when poorly predicted genes are excluded, as discussed below.

3.2.3 Assessing prediction accuracy differences across populations and across tissues

In the analyses described below to assess differences in prediction accuracy with PrediXcan across populations, two sets of genes were considered – all genes without any filtering and a subset of genes using the filtering process previously described.

We first compared prediction performance between the two continental groups – European and African. For each gene, we calculated two Pearson's correlation coefficients between observed and predicted gene expression levels – one based on all European samples and the other one based on the African samples. We then used a paired t-test to assess differences in mean prediction accuracy between the correlation coefficients for European samples vs correlation coefficients for African samples.

To assess differences in prediction accuracy across the five populations, we used a linear mixed effects model approach where we fit the following model:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{FIN,i} + \beta_2 \mathbb{I}_{GBR,i} + \beta_3 \mathbb{I}_{TSI,i} + \beta_4 \mathbb{I}_{YRI,i} + \epsilon_{ij}, \tag{3.1}$$

where r_{ij} is the correlation coefficient for gene i in population j; and $\mathbb{I}_{FIN,i}$, $\mathbb{I}_{GBR,i}$, $\mathbb{I}_{TSI,i}$, and $\mathbb{I}_{YRI,i}$ are indicator variables that are equal to 1 if the gene correlation was calculated on the population indicated in the subscript, and otherwise are equal to 0. Thus, we modeled population as a categorical predictor, with the CEU population as a reference. To account for variation between genes, we included a random intercept γ_i for each gene and we assumed that $\gamma_i \sim \mathcal{N}(0, \sigma^2_\gamma)$. We also included an error term ϵ_{ij}, such that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. We used repeated measures ANOVA to test the null hypothesis of $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ for no difference in mean Pearson's correlation coefficients among the populations. A Wald test was used to assess significance of differences in mean Pearson's correlation coefficients between CEU, the reference population, and each of the other four populations.

We also ran a similar analysis where we excluded the CEU population due to potentially lower quality of the CEU cell lines, as reported in the literature [74, 12]. We fit a model identical to (3.1), excluding the CEU and using the FIN population as a reference:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{GBR,i} + \beta_2 \mathbb{I}_{TSI,i} + \beta_3 \mathbb{I}_{YRI,i} + \epsilon_{ij}, \tag{3.2}$$

where the notation is the same as above.

Additionally, we tested for differences in prediction accuracy across the four European populations. For this analysis, we included only individuals of European ancestry and fit the following linear mixed effects model:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{FIN,i} + \beta_2 \mathbb{I}_{GBR,i} + \beta_3 \mathbb{I}_{TSI,i} + \epsilon_{ij}, \tag{3.3}$$

where CEU is included as the reference population in the model. As in the previously described analyses, a repeated measures ANOVA was used to test for differences in prediction accuracy across the four European populations.

To evaluate how the PrediXcan performance with whole-blood (WB) databases differed from LCL databases, we restricted the set of genes to only those that were present in both the WB and LCL databases. First, we presented scatter plots of

correlation coefficients comparing WB and LCL databases in the five populations separately. Then we recalculated Pearson's correlation coefficients between observed and predicted expression values with all the five populations combined but separately for every database, i.e. as a result, we had two correlation coefficients per gene, one that corresponded to a GTEx WB database and one to a GTEx LCL database. We compared each pair of GTEx WB and GTEx LCL databases using a paired t-test between LCL-based correlation coefficients and WB-based correlation coefficients. All the statistical analyses described above were performed in `R` version 3.3.3 [56]. All plots were generated with `ggplot2` [72].

3.3 Results

3.3.1 Overview of PrediXcan weight databases

In Table 3.1, we summarize the main features of the PrediXcan weight databases that we used in the analyses. Compared to DGN database, GTEx databases have fewer gene models and smaller training sample sizes. HapMap and 1KG-based models differ in the number of variants used for training: GTEx Hapmap models were trained on the HapMap SNP set while GTEx 1KG were trained on the 1000 Genomes SNP set, so the latter utilize more SNPs when predicting expression. While GTEx LCL databases are based on relatively small training sets, they are derived from the same tissue as the GEUVADIS RNA-seq data we analyzed. Lastly, DGN and GTEx v7 sets of weights were trained only on Europeans samples, while GTEx v6 databases had a small fraction of non-Europeans.

To avoid repetition, results using the DGN, GTEx v7 WB and GTEx v7 LCL databases are included in the main text, while the results for other four databases are provided in Appendix A.

Table 3.1: Summary of PrediXcan databases used in analyses.

PrediXcan Database	Training set size	Number of models	Number of SNPs used
DGN whole blood	922	13,171	249,696
GTEx v6 1KG whole blood	338	6,759	185,786
GTEx v6 1KG LCL	114	3,759	125,045
GTEx v6 HapMap whole blood	338	6,588	136,941
GTEx v6 HapMap LCL	114	3,441	91,237
GTEx v7 HapMap whole blood	315	6,297	140,931
GTEx v7 HapMap LCL	96	3,045	88,143

3.3.2 PrediXcan prediction accuracy differs across diverse populations

Using DGN, GTEx WB and GTEx LCL models and sequence data, gene expression was predicted for 10387, 5432 and 2777 genes, respectively (see Table 3.2). The number of genes with available predictions varied by population, where the four European populations had a similar number of gene predictions while the counts for YRI were slightly lower. We excluded 33 genes, 13 genes, and 10 genes from DGN, GTEx WB, and GTEx LCL, respectively, due to there being no variation in predicted gene expression values for at least one of the populations. For the remaining genes, we identified those that had poor prediction accuracy based on associations between observed and predicted values, as described in the Materials and Methods section on filtering poorly predicted genes. From the genes predicted with DGN database, two-thirds were labeled by this criterion as "poorly predicted", while slightly less than a half were labeled as such from gene sets predicted using the GTEx databases. As previously mentioned, we also considered the performance of PrediXcan without doing any filtering of the genes. For every weight database, we had two sets of genes – before and after filtering – where the latter set is a much smaller subset of the former. Both

versions were used and evaluated in downstream analyses.

Table 3.2: Number of genes for which Pearson correlation coefficients are available by population and by PrediXcan weight database.

PrediXcan database	DGN	GTEx v7 WB	GTEx v7 LCL
Genes with observed and predicted expression values	10,387	5,432	2,777
By population:			
CEU	10,385	5,432	2,777
FIN	10,385	5,432	2,777
GBR	10,385	5,432	2,777
TSI	10,385	5,432	2,776
YRI	10,354	5,419	2,767
Genes before filtering	10,354	5,419	2,767
Genes after filtering	3,493	2,288	1,699

We first evaluated performance of PrediXcan for the two continental populations, European and African. We compared Pearson's correlation of predicted and observed gene expression values for the combined sample consisting of all individuals from the four European-ancestry populations to Pearson's correlation calculated for the YRI African population sample. As only two groups were being compared in this analysis, a paired t-test was used to assess differences in prediction accuracy, where the pairing was based on the gene. With or without the filtering of genes, we find the mean difference in gene correlation coefficients between the European and African samples to be highly significantly different from zero, regardless of the weight database used (all p-values $< 2.2 \times 10^{-16}$), with the African population having lower prediction accuracy than the European samples.

Next, we computed gene correlation coefficients, separately in each of the five populations. Violin plots display the correlation coefficients by population across genes before and after filtering (see Figures 3.1A and 3.1B, respectively). Figure 3.1A shows correlation coefficients for the genes before any filtering was done and we observe that LCL-derived models perform better than WB-derived: i.e., DGN and GTEx v7 WB correlation distributions are centered at values close to 0, whereas GTEx LCL correlation distributions are centered at higher values, especially for the four European populations. We also note that prediction accuracy is slightly lower for the African populations than for any of the European populations across the three weight databases. This trend is even more obvious after the filtering process. As we can see in Figure 3.1B, the overall performance accuracy improved after filtering in all the populations, as expected. However, the difference in prediction performance in Europeans vs Africans is even more apparent. The four European populations have similar prediction accuracy, whereas it is lower for the African population. Similarly to panel A, LCL-derived prediction models perform better than WB-derived in filtered genes in Figure 3.1B.

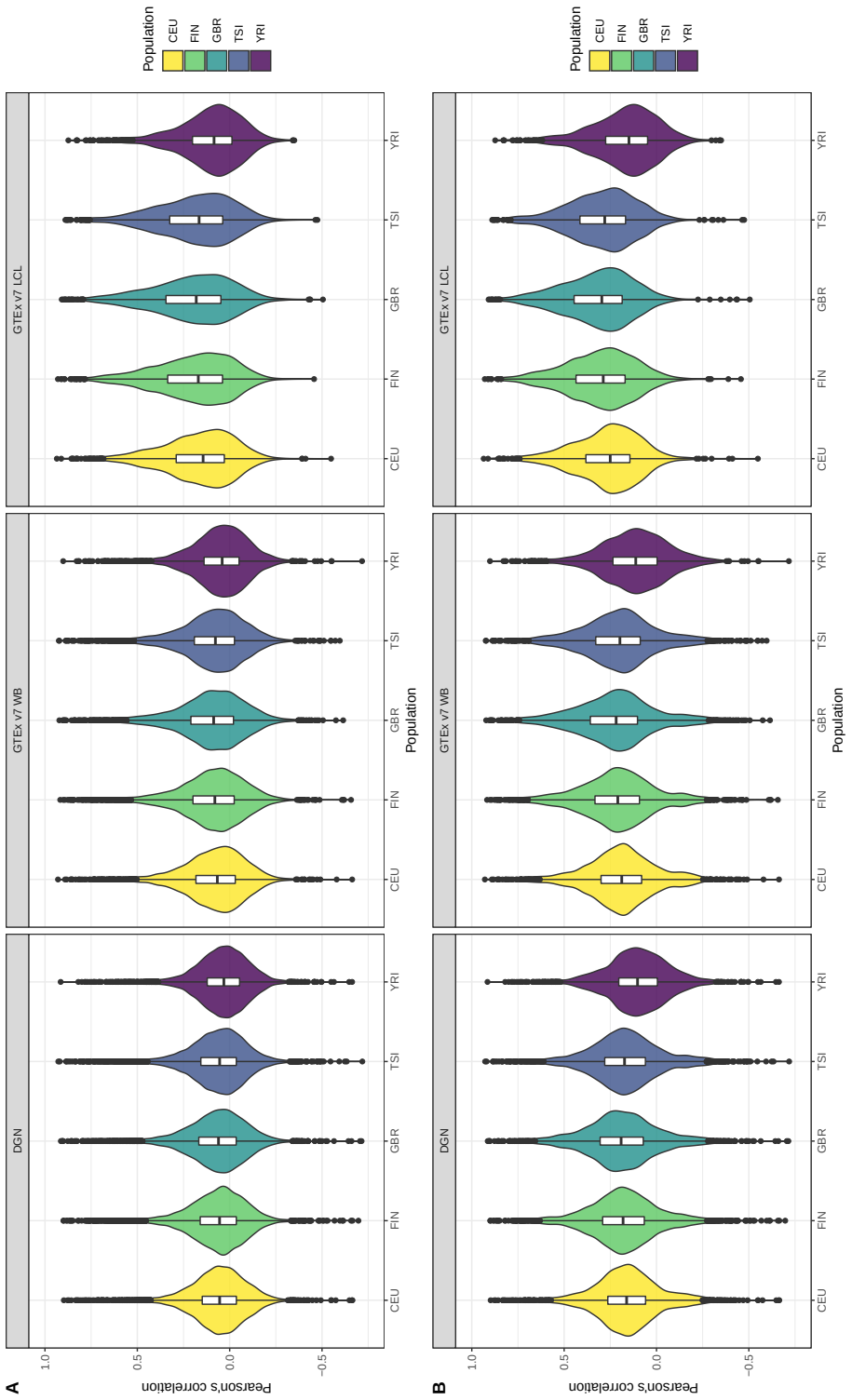


Figure 3.1: **Violin plots of gene expression correlation coefficients by five populations using DGN, GTEx v7 WB and GTEx v7 LCL weight databases;** (A) before and (B) after filtering out poorly predicted genes.

Afterwards, we binned the genes into six categories based on the gene correlation coefficients (see Table 3.3). The majority of genes have very poor prediction accuracy – of the genes predicted with whole-blood databases, a third have negative correlations and a half have correlations between 0 and 0.2. Of the genes predicted with LCL, a fifth have negative correlations and over a third have correlations between 0 and 0.2. The distribution of gene correlation coefficients is fairly similar across the four European populations, although predictive accuracy seems worse in CEU compared to FIN, GBR, and TSI. The predictive accuracy is the lowest in the African sample. Across all populations, only a small number of genes were predicted with high accuracy (with $r > 0.6$). Furthermore, all European populations have a greater number of well-predicted genes than the African population, regardless of the weight database used.

Next, we assessed the association between the prediction accuracy (as gene correlation coefficients) and population category via repeated measures ANOVA and linear mixed models using both sets of genes, all and filtered. The results for unfiltered and filtered genes were comparable and led to equivalent conclusions. Based on the repeated measures ANOVA, we find that prediction accuracy differs across populations for filtered and unfiltered sets of genes, regardless of the weight database used (p-values for all databases were $< 2.2 \times 10^{-16}$). Below, we focus our attention on filtered genes and present the parameter estimates and their 95% confidence intervals calculated using model-based standard errors for the model 3.1 in Table 3.4. From the linear mixed model 3.1, we find that the prediction accuracy is significantly higher in FIN, GBR and TSI and significantly lower in YRI, compared to CEU. This suggests that predictive performance varies not only among distant populations, but also among closely related populations. When we performed the analysis on a full set of genes, without any filtering, regression coefficients were slightly attenuated towards zero; however, the conclusions from hypothesis testing remained the same.

We repeated the analysis described above, this time excluding the CEU population.

Table 3.3: Gene counts per population, per database, per correlation category for the five populations using DGN, GTEx WB and GTEx LCL weight databases.

	Unfiltered					Filtered				
	CEU	FIN	GBR	TSI	YRI	CEU	FIN	GBR	TSI	YRI
	DGN database									
$r < 0$	3,583	3,491	3,480	3,587	4,156	561	547	554	585	911
$0 < r < 0.2$	5,107	4,976	4,812	4,954	5,001	1,533	1,379	1,258	1,409	1,674
$0.2 < r < 0.4$	1,359	1,480	1,589	1,434	1,016	1,097	1,162	1,209	1,121	728
$0.4 < r < 0.6$	239	302	354	290	147	236	300	353	289	146
$0.6 < r < 0.8$	56	93	105	75	31	56	93	105	75	31
$0.8 < r < 1$	10	12	14	14	3	10	12	14	14	3
	GTEx v7 WB database									
$r < 0$	1,756	1,621	1,622	1,684	2,101	336	309	314	335	590
$0 < r < 0.2$	2,471	2,450	2,366	2,456	2,491	877	786	732	820	993
$0.2 < r < 0.4$	902	958	981	901	668	788	804	793	758	546
$0.4 < r < 0.6$	210	282	329	278	117	207	281	328	275	117
$0.6 < r < 0.8$	69	93	100	85	38	69	93	100	85	38
$0.8 < r < 1$	11	15	21	15	4	11	15	21	15	4
	GTEx v7 LCL database									
$r < 0$	546	488	484	509	774	80	69	55	69	274
$0 < r < 0.2$	1,119	1,031	996	1,050	1,296	560	443	426	477	777
$0.2 < r < 0.4$	718	742	761	736	510	675	681	692	681	461
$0.4 < r < 0.6$	293	361	369	360	145	293	361	369	360	145
$0.6 < r < 0.8$	80	126	137	96	38	80	126	137	96	38
$0.8 < r < 1$	11	19	20	16	4	11	19	20	16	4

We present the parameter estimates and the corresponding 95% confidence intervals in Table 3.5. From the repeated measures ANOVA, we find that prediction accuracy differs across the four populations (p-values for all databases were $< 2.2 \times 10^{-16}$). Moreover, based on the coefficients and the corresponding p-values from the linear mixed model 3.2, we estimate the prediction accuracy to be significantly higher in GBR and significantly lower in TSI and YRI, compared to the FIN population (see corresponding p-values in Table 3.5). This difference in prediction accuracy is the greatest between YRI and FIN when GTEx v7 LCL weight database was used. Like in the analysis above, we notice that predictive performance differs across populations, including European populations.

Finally, we evaluated PrediXcan prediction accuracy on a subset of subjects with European ancestry. Based on the repeated measures ANOVA test, prediction performance differs across the four European populations in genes before and after filtering, regardless of the weight database used (p-values for all databases were $< 2.2 \times 10^{-16}$).

Table 3.4: Results from linear mixed models for population category (with CEU as a reference) and change in gene correlation coefficient among filtered genes.

	DGN			GTEx v7 WB			GTEx v7 LCL		
	Estimate	95% CI	p-value	Estimate	95% CI	p-value	Estimate	95% CI	p-value
FIN	0.019	(0.014, 0.025)	1.3×10^{-11}	0.021	(0.015, 0.028)	1.3×10^{-9}	0.038	(0.030, 0.046)	$< 10^{-16}$
GBR	0.029	(0.023, 0.034)	$< 10^{-16}$	0.032	(0.025, 0.039)	$< 10^{-16}$	0.051	(0.043, 0.059)	$< 10^{-16}$
TSI	0.010	(0.004, 0.016)	3.9×10^{-4}	0.013	(0.007, 0.020)	4.6×10^{-5}	0.027	(0.019, 0.035)	2.9×10^{-11}
YRI	-0.054	(-0.059, -0.048)	$< 10^{-16}$	-0.070	(-0.077, -0.063)	$< 10^{-16}$	-0.097	(-0.105 -0.089)	$< 10^{-16}$

Table 3.5: Results from linear mixed models for population category (excluding CEU, with FIN as a reference) and change in gene correlation coefficient among filtered genes.

	DGN			GTEx v7 WB			GTEx v7 LCL		
	Estimate	95% CI	p-value	Estimate	95% CI	p-value	Estimate	95% CI	p-value
GBR	0.010	(0.004, 0.015)	9.2×10^{-4}	0.011	(0.004, 0.018)	3.1×10^{-3}	0.013	(0.005, 0.021)	$2.0 \times \times 10^{-3}$
TSI	-0.009	(-0.015, -0.003)	1.8×10^{-3}	-0.008	(-0.015, -0.001)	2.8×10^{-2}	-0.011	(-0.019, -0.003)	8.9×10^{-3}
YRI	-0.073	(-0.079, -0.067)	$< 10^{-16}$	-0.091	(-0.098, -0.084)	$< 10^{-16}$	-0.134	(-0.143, -0.126)	$< 10^{-16}$

3.3.3 PrediXcan prediction accuracy differs between tissues

As can be seen in the violin plots in Figure 3.1, both databases based on whole blood perform similarly, and LCL-based database displays improved prediction accuracy. In order to compare pairwise gene correlations, we restricted our analyses to the 1,595 genes common for both GTEx v7 WB and GTEx v7 LCL.

Scatter plots presented in Figure 3.2 suggest that the majority of genes have very similar correlation coefficients when using WB and LCL databases across all populations. However, we see more genes in the upper left corner, above the dotted line, indicating that using the LCL database results in more genes with better prediction accuracy. This result is not surprising since the expression data we used were derived from LCL. The results of the paired t-test are consistent with the visual examination of the data: the mean difference between gene correlations based on the GTEx v7 LCL models and based on the GTEx v7 WB models is 0.03 (p-value $< 2.2 \times 10^{-16}$), with predictions based on the LCL model having higher performance.

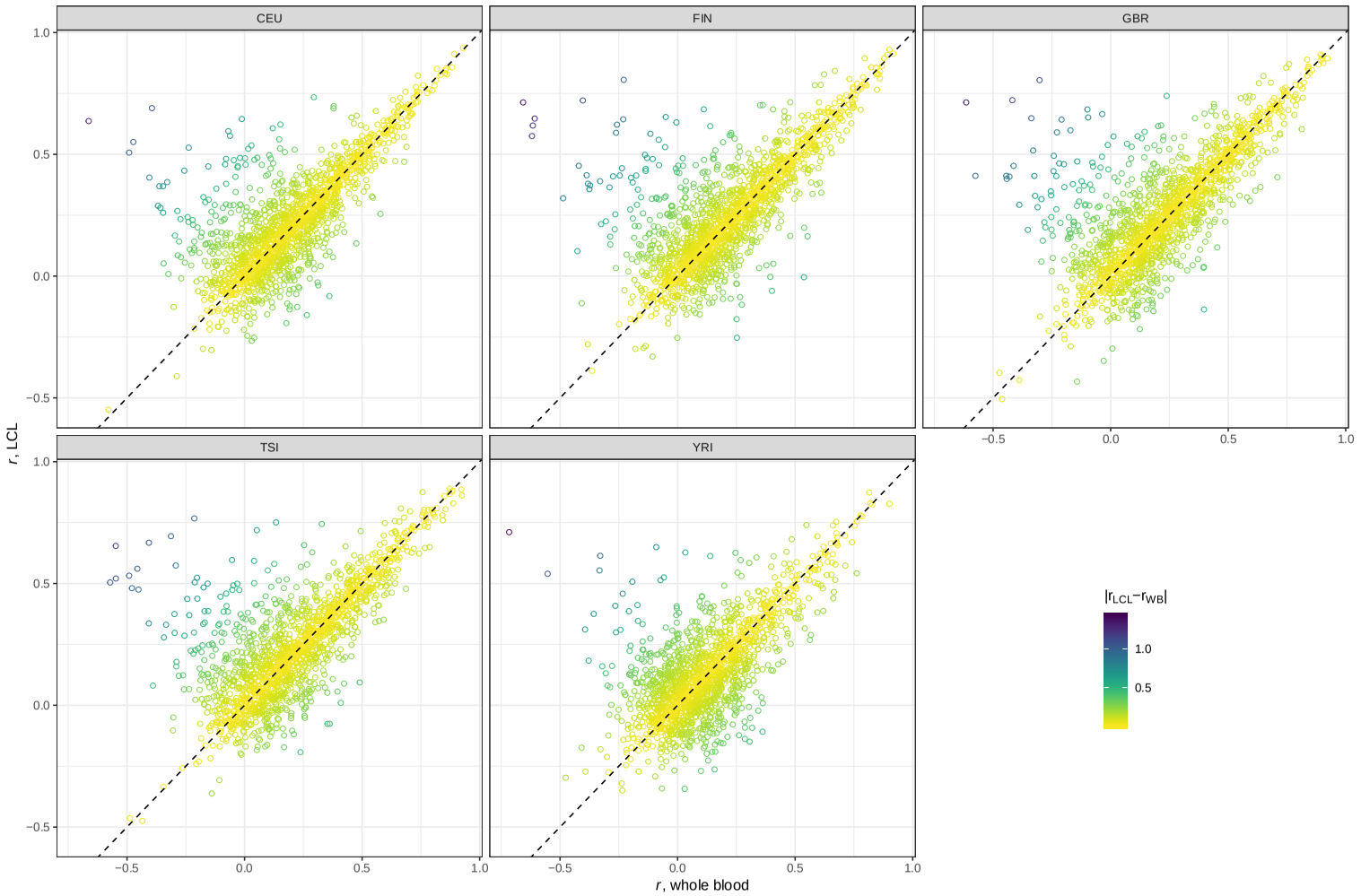


Figure 3.2: **Scatter plots comparing gene correlation coefficients by population using GTEx v7 LCL vs GTEx v7 WB databases.**

3.4 Discussion

In this chapter, we evaluated the performance of PrediXcan and compared the prediction accuracy of the method across five geographically diverse populations from two continents for seven weight databases. Models from all weight databases considered were trained on subjects primarily of European ancestry; three of the databases were derived from LCL and the remaining four from whole blood. As a measure of prediction accuracy, we computed correlation coefficients for each gene in all populations and used both paired t-tests and linear mixed effects models to assess evidence of significant differences in prediction performance across populations. We also investigated whether whole blood models could be used for predicting gene expression levels in LCL.

We find highly significant differences in prediction accuracy with PrediXcan in the European ancestry populations as compared to the YRI African population, with the prediction accuracy being lower in YRI. The lower accuracy with PrediXcan in the African population is expected since the PrediXcan models were largely trained using European ancestry samples, and this result is consistent with recent works showing that prediction accuracy is expected to be higher when the training and testing cohorts are of similar ancestry [41, 31, 23]. Surprisingly, we also find highly significant differences in prediction accuracy with PrediXcan among the closely related European ancestry populations, with the Finnish, British, and Italian populations having significantly higher prediction accuracy than the CEU. These results are consistent across all seven PrediXcan weight databases we considered. Lastly, we also find that LCL-trained models outperformed whole-blood-trained models across populations, although the prediction accuracy was similar for many of the genes.

Among the European populations, we find that prediction accuracy for the CEU population was the lowest. LCLs are derived from B cells found in whole blood, and they provide a continuous supply of genetic material for GWAS and gene expression studies. However, they do undergo a transformation to become immortal that can change

their biology and they do not have the same properties as native tissue [28]. Storage conditions, freeze-thaw cycles, and maturity of cell lines can also affect gene expression patterns [12, 74]. The CEU cell lines were collected much earlier than the other cell lines and LCL age can have a confounding effect and bias downstream analyses[74]. This factor could have contributed to the differences in prediction accuracy among European populations. We did, however, perform a sensitivity analysis that excluded the CEU population, and the highly significant differences in prediction accuracy with PrediXcan among the three European populations and the YRI African population remained.

Overall, PrediXcan accurately predicted gene expression for some genes; however, the majority of genes had very poor correlation between measured and predicted expression levels. For almost half the genes, for example, the correlation was negative. There are some important caveats and limitations to point out with the PrediXcan method. First, the prediction models of PrediXcan are based on common *cis*-variants and they do not take rare *cis*- and *trans*-regulatory elements into account. Common *cis*-eQTLs only account for 9%-12% of genetic variance in gene expression, according to a large twin study [24]. Another recent study demonstrates that *trans*-acting variants largely contribute to gene expression variation, with estimates of genetic variance in expression due to *trans*-acting variation ranging from 60% to 90% [33]. However, individual effects of each *trans*-variant are very weak and difficult to map because they require well-powered studies.

We conclude this chapter by highlighting that the lack of genomic data from diverse populations limits the ability to effectively interpret and translate genomic results into clinical applications for individuals from diverse populations, and particularly non-European ancestry populations. The results presented in this paper illustrate that gene expression prediction models are, in general, not transferable across diverse populations from different continents, and further corroborate the importance of including more ancestrally diverse individuals in medical genomics to ensure that everyone gets the

benefits of precision medicine and to avoid further exacerbating healthcare inequality [38, 51, 50]. We also demonstrate that there can be differences in prediction accuracy among closely related European populations, suggesting that prediction models that take into account fine-scale ancestry differences among individuals may be important for improved prediction of gene expression from genetic data. Lastly, our study had only modest sample sizes and evaluated gene expression prediction accuracy with PrediXcan in European and African populations. Future transcriptomic studies with much larger samples sizes are needed for the development of improved gene expression prediction models for multi-ethnic populations, including admixed populations such as African Americans and Hispanic/Latino populations, who have recent ancestry derived from multiple continents.

Chapter 4

FUSED LASSO IMPROVES TRANSCRIPTOME PREDICTION IN DIVERSE POPULATIONS

4.1 Introduction

In recent years, transcriptome-wide association studies (TWAS) have become a popular approach of detecting effects of gene expression on complex traits. Unlike a genome-wide association study (GWAS), where phenotypes are regressed onto genotypes, in a TWAS the phenotype is regressed onto transcriptome data, which makes a TWAS a type of a gene-based test. These methods help elucidate molecular mechanisms underlying common diseases and interpret GWAS discoveries by identifying gene-trait associations and prioritizing candidate genes. Due to the cost of collecting gene expression data and tissue unavailability, transcriptome-wide association studies often use imputed gene expression values, instead of measured values. Imputed gene expression values are calculated from predictive models of expression variation. Predicted transcriptome allows researchers to conduct analyses in very large cohorts (samples of tens or hundreds of thousands of individuals) instead of typical gene expression studies with measured gene expression (samples of hundreds to thousands of individuals).

Several methods, such as PrediXcan [21], have been developed in recent years to conduct TWAS. These methods use various regularized regression techniques for transcriptome prediction. These techniques first identify a set of genetic variants that influence gene expression, called expression quantitative trait loci (eQTLs), from reference datasets. Then, they build per-gene prediction models that are linear combinations of precomputed weights and eQTLS, identified in the previous step.

Subsequently, these models can be downloaded and applied to independent datasets with genotype or whole-genome sequencing data. Finally, the phenotype of interest and predicted transcriptome values can be tested for associations.

Unfortunately, most transcriptome prediction models are constructed from training data with predominantly subjects of European descent, whereas TWAS are often conducted in multi-ethnic cohorts. The European ancestry overrepresentation in genomic studies has been well documented and puts limitations not only on gene discovery and fine mapping, but, more importantly, on applications in personalized medicine [11, 55, 8]. The training data used for PrediXcan models are highly biased towards European ancestry. For example, GTEx version v6p subjects are over 85% European, while the GTEx v7 and DGN subjects are entirely of European descent [34, 7]. The lack of suitable reference datasets in non-European individuals leads to scenarios in which models trained in Europeans are used to predict into non-European populations. It has been shown that gene expression models trained on subjects of European descent show low cross-population generalizability to other populations [41, 29].

Genetic architectures, linkage disequilibrium, and allele frequencies vary across populations and, thus, pose substantial challenges in constructing accurate prediction models for diverse populations. Comparative studies on the genetic regulation of gene expression across populations have shown that, while the effects sizes and effect directions for some eQTLs are shared across populations, a substantial number of eQTLs is population-specific [63, 30, 59]. Therefore, failure to identify and account for shared and population-specific eQTLs can lead to a reduction in accuracy of gene expression imputation models.

Here we present a new flexible and robust method for predicting gene expression from genotype data that deconvolutes eQTL architecture into shared and population-specific components. We start with a brief overview of existing methods used for gene expression prediction, such as standard lasso [66] and elastic net [77]. We then outline

a novel method that uses fused lasso regression [67] to incorporate subjects' ancestry information when building transcriptome prediction models from genotype data from multiple populations. We also demonstrate how to build such prediction models using an ethnically diverse TOPMed Multi-Ethnic Study of Atherosclerosis (MESA) cohort [9].

4.2 Methods

4.2.1 Regularized regression framework for eQTL mapping and transcriptome prediction models

We begin with an overview of lasso usage for transcriptome prediction and how it can be extended to jointly analyze samples from multiple populations. Hereafter, we assume that the population label for each individual is either known or has been inferred from the genotyping array or sequencing data.

Let $\mathbf{X}$ denote an $n \times s$ matrix of genotype data for a homogeneous populations for a given gene, where n represents the number of subjects in a sample and s is a number of genetic *cis*-variants, from a set $\mathcal{S}$, that are within some window of the transcription region for the gene. Let $\mathbf{y}$ be the vector of length n with gene expression measurements. We assume that genotypes have additive genetic effects on gene expression levels:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is a vector of s effect sizes of genetic variants on gene expression, and $\boldsymbol{\epsilon}$ is a zero-mean Gaussian error term for environmental effects and other factors. We assume that the $\mathbf{y}$ and each column of $\mathbf{X}$ are centered, so that they have zero mean and we do not have to model the intercept term. When the number of genetic variants s is small and the number of subjects n is relatively large, the regression coefficients $\boldsymbol{\beta}$ can be estimated by minimizing the sum of squared residuals:

$$\arg\min_{\beta} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}$$

Since s is often larger than n, the regression coefficients $\boldsymbol{\beta}$ can be estimated using penalized regression methods, such as lasso and elastic net. In the case of standard lasso, the estimates of $\boldsymbol{\beta}$ are obtained by minimizing:

$$\arg\min_{\beta} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1\}$$

where the last term $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{s} |\beta_j|$ is an l_1 penalty that encourages sparsity in the coefficients, so that only a few variants have non-zero coefficients. The tuning parameter λ_1 controls the amount of sparsity, i.e. larger values of λ_1 correspond to higher penalization and, thus, more variants with zero regression coefficients.

When using elastic net, we seek to minimize:

$$\arg\min_{\beta} \left\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\right\}$$

where the last two terms are an l_1 sparsity penalty term $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{s} |\beta_j|$ and an l_2 ridge penalty term $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{s} \beta_j^2$, with corresponding tuning parameters λ_1 and λ_2, respectively, that are usually chosen via cross-validation.

4.2.2 Fused lasso identifies group-specific predictors and weights

When dealing with samples that include subjects from diverse or admixed populations, we can apply the aforementioned regularized regression techniques to either pooled datasets (datasets that combine subjects from different populations) or to each population group separately. In the former case, we can detect eQTLs that have common effects on all populations but we may miss the variants that have different effects on different populations or effects that are specific to a population. In the latter case, we may lose power to detect variants that are common in multiple population due to not leveraging information across samples.

Consider a collection $\mathcal{K}$ of group labels, e.g. population, ancestry, race or ethnicity, either self-reported or inferred. We propose to group the genotype and transcriptome

data according to the labels and introduce group-specific regression coefficients $\boldsymbol{\beta}^k$, i.e. we concatenate $\boldsymbol{\beta}$.

The fused lasso estimates minimize the criterion:

$$\arg\min_{\beta^k} \left\{ \sum_{k\in\mathcal{K}} \left[(\mathbf{y^k} - \mathbf{X^k}\boldsymbol{\beta}^k)'(\mathbf{y^k} - \mathbf{X^k}\boldsymbol{\beta}^k) + \lambda_1^k \|\boldsymbol{\beta}^k\|_1\right] + \lambda_2^k \sum_{\substack{k,l\in\mathcal{K} \\ l\neq m \\ k<l}} \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^l\|_1 \right\} \quad (4.1)$$

where we have two penalty terms: standard lasso l_1 penalty, λ_1, and fusion penalty, λ_2. Larger λ_1 puts more priority on sparsity of the coefficients, while larger λ_2 encourages multiple populations to have similar features.

We concatenate $\boldsymbol{\beta}^k$s into a vector of length $s \cdot K$, such that $\boldsymbol{\beta}_c = ((\boldsymbol{\beta}^1)', ..., (\boldsymbol{\beta}^K)')'$ and concatenate $\mathbf{y^k}$s into a vector of length n, $\mathbf{y_c} = ((\mathbf{y^1})', ..., (\mathbf{y^K})')'$. We also form an $n \times (s \cdot K)$ block-diagonal matrix $\mathbf{X_c}$, where $\mathbf{X^k}$s are along the diagonal and all the other elements are set to 0. Then, the minimization problem 4.1 can be expressed in the following form:

$$\arg\min_{\beta_c} \left\{(\mathbf{y_c} - \mathbf{X_c}\boldsymbol{\beta}_c)'(\mathbf{y_c} - \mathbf{X_c}\boldsymbol{\beta}_c) + \lambda_1\|\boldsymbol{\beta}_c\|_1 + \lambda_2\|\mathbf{D}\boldsymbol{\beta}_c\|_1\right\}, \quad (4.2)$$

where λ_1 is a sparsity penalty, λ_2 is fusion penalty, and $\mathbf{D}$ is an $(s \cdot K) \times (s \cdot K)$ penalty matrix that fuses coefficients corresponding to the same genetic variant in each population.

The fused lasso technique allows us to combine information and identify predictor variants jointly across populations, thus, maximizing the power to identify eQTLs that are common across populations. Its main strength is in its flexibility of allowing effect sizes to differ across different ancestries and allowing us to detect predictors that are population-specific or predictors that have effect sizes with different directions. Additionally, it allows us to detect with greater power predictors that have weak signals in each population separately but common to all the populations combined. At the same time, it limits influence of predictor variants that have weak associations in only one of the populations.

4.2.3 Simulation studies

We performed simulation studies in order to: (1) assess prediction accuracy of the proposed method in various settings of eQTL sharing across populations, and (2) to compare the proposed method with existing methods, such as elastic net and standard lasso. We explored genetic architecture factors that play a role in transcriptome predictive modeling and how the choice of training sets affects the accuracy of predictions when the models are applied to datasets with diverse ancestral backgrounds. We tested multiple parameter configurations for simulating gene expression phenotypes. We varied the number of causal variants (k = 1, 2, and 10), the proportion of location sharing of eQTLs across populations (p = 0, 0.5, and 1), and the magnitude and direction of effect sizes of the causal variants. We set *cis*-heritability for all simulated gene expression phenotypes to a realistic value of 0.15.

To explore how the choice of the training set affects prediction performance, we considered the following scenarios:

1) ancestry-stratified training sets that result in ancestry-specific predictors and weights;

2) multi-ancestry training set where models have identical predictors and weights for all ancestral groups;

3) multi-ancestry training set where each subject has an ancestry label and the models have group-specific weights and predictors.

We compared our proposed fused lasso method in scenario 3 to elastic net and standard lasso in scenarios 1 and 2. For each setting, we used the same train-test-validate scheme to derive transcriptome prediction models. We then applied prediction models to the test sets in concordance with the ancestry label of each individual. We evaluated prediction models by calculating per-gene Pearson's correlation coefficients between predicted and simulated gene expression levels.

Gene expression phenotype models

We consider a random gene m from a set of autosomal protein-coding genes $\mathcal{M}$ and a corresponding set of *cis*-variants $\mathcal{S}$ that are within a 500kb window of the gene transcription region. We randomly choose causal variants from an LD-pruned set of variants from S with the estimated minor allele frequency (MAF) of at least 0.05 in each population considered. We refer to the set of causal variants as $\mathcal{S}_c$. In each population, we simulate gene expression phenotypes assuming additive genotype effects with:

$$\mathbf{Y^k} = \sum_{s \in \mathcal{S}_c} \mathbf{x_s^k} \beta_s^k + \boldsymbol{\epsilon}$$

where $\mathcal{S}_c$ is a set of causal *cis*-variants, and $\epsilon_i \sim \mathcal{N}(0, 1)$ is individual random noise. We draw genotype effect sizes, β_s^k, from $\mathcal{N}(\mu_k, \sigma^2)$.

4.2.4 Training of prediction models in the TOPMed MESA dataset

To assess the performance of our method, we used 556 African American and European ancestry participants from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort, which is a part of the Trans-omics for Precision Medicine (TOPMed) program. We randomly split participants into training ($n = 360$) and test ($n = 198$) sets, so that the ratio of African American to European ancestry individuals in each set was 1:1. We compared three approaches: elastic net, standard lasso, and fused lasso. Using the three methods, we trained models on ancestry-stratified (training AA or training EA sets, each $n = 180$) or multi-population (training AA+EA set, $n = 360$) training sets. We used a total of 12,691 protein-coding genes for training models. To adjust for confounders, we fit a multivariate linear model for gene expression values controlling for sex, age, and the first 3 PC-AiR principal components (PCs) reflecting genetic ancestry [16] and used the obtained residuals as the outcome. To assess the performance of the prediction models, we applied them to MESA test sets and compared measured

and predicted gene expression values by calculating per-gene R^2s. Additionally, we calculated a mean squared error (MSE) for each gene and performed a one-sided Wilcoxon signed-rank test testing a null hypothesis of no difference in MSEs of fused lasso vs existing methods.

4.3 Results

We developed a method that utilizes subjects' ancestry information when training transcriptome prediction models from genotype data. Before applying our method to real data, we performed an extensive set of simulation, under various genetic architecture scenarios. In our simulations, we used 1000 Genomes whole-genome sequencing data and simulated phenotypes. We compared the performance of existing methods, such as standard lasso and elastic net, to the proposed fused lasso method using ancestry-stratified and pooled training sets. Afterwards, we evaluated performance of our method using the Multi-Ethnic Study of Atherosclerosis (MESA) study with whole-genome sequencing data and peripheral blood mononuclear cells gene expression data. We found that for the simulation settings where genetic architectures are similar among ancestral populations, jointly analyzing all samples together can result in improved prediction performance. In contrast, when genetic architectures differ among populations, it is advantageous to use ancestry-stratified training sets. Since the genetic architecture of a particular gene or set of genes is usually unknown, our method allows flexibility determining both population-specific and shared eQTLs.

4.3.1 Simulated data with ancestral groups as population labels

We used real whole-genome sequencing (WGS) data of 372 unrelated individuals from the 1000 Genomes dataset. The dataset included two European-ancestry populations: Utah residents with ancestry from northern and western Europe (CEU) and British from England and Scotland (GBR). It also included two African-ancestry populations: Yoruba (YRI) individuals from Ibadan, Nigeria and Nigerian (ESN) individuals from

Esan, Nigeria. Ancestry-stratified training sets included CEU ($n = 98$) and YRI ($n = 98$) whereas combined training set included both CEU and YRI ($n = 196$) populations. We validated our models on test sets consisting of participants from GBR ($n = 88$) and ESN ($n = 88$) populations.

For each simulation scenario, we randomly selected 1,000 protein-coding genes and variants corresponding to the region within 500kb of the genes' start and end sites. We simulated eQTL architectures under an additive model of size k causal alleles ($k = 1$, 2, and 10) and an expression phenotype with *cis*-heritability of $h^2 = 0.15$. We allowed the two populations to share eQTLs in various proportions (0, 0.5, and 1.0) and varied the proportions of eQTLs with the same direction of effect sizes (0, 0.5, and 1.0). We then applied the same train-test scheme using fused lasso, standard lasso and elastic net. Prediction models using standard lasso and elastic net were trained on CEU and YRI separately, resulting in population-specific models, as well as on CEU+YRI, resulting in population-shared model. We only applied ancestry-concordant prediction models to each test set, i.e. models trained on CEU or CEU+YRI were applied to GBR and models trained on YRI or CEU+YRI were applied to ESN. To assess prediction accuracy, we calculated Pearson's correlation between predicted and simulated expression values. We report simulation results for fused lasso and elastic net in this chapter and include results comparing fused lasso to standard lasso in Appendix B.

Similar genetic architecture between ancestral populations

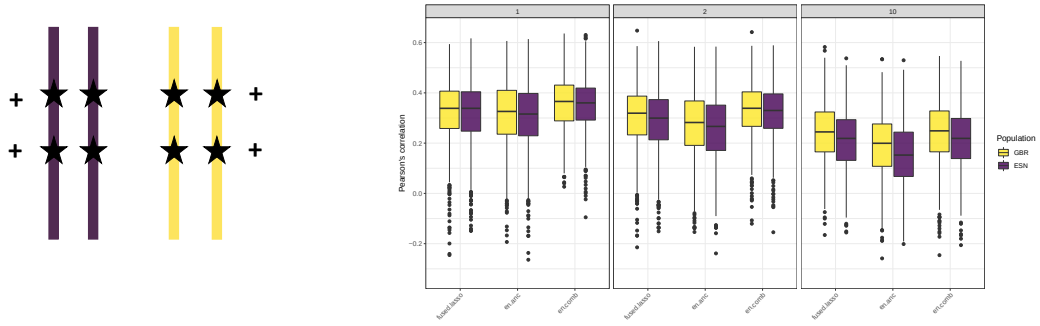
Figure 4.1 shows the results from a simulation scenario where two populations have similar genetic architecture. In this simulation setting, eQTL positions and their corresponding effect sizes are exactly the same between the two populations. In boxplots, we present Pearson's correlations between between predicted and simulated phenotypes for GBR and ESN partitioned by the number of causal eQTLs, comparing fused lasso, ancestry-specific elastic net and ancestry-pooled elastic net, trained on a

combined dataset. These trends are very similar for $k = 1$ and $k = 2$ causal eQTLs, although fused lasso shows slightly diminished performance, compared to elastic net trained on a combined dataset. This is also the case when we compare fused lasso to standard lasso models (Figure B.1 and Table B.1). This suggests that when eQTLs are causal in all populations, we can reliably impute gene expression by using pooled datasets, while training models on ancestry-specific datasets might lead to loss in power to detect eQTLs and, thus, less accurate predictions.

Overall, elastic net trained on a combined set (EN comb) and fused lasso performs with the highest accuracy, whereas elastic net trained on an ancestry-specific set (EN anc) performs with the lowest. For example, for $k = 10$ causal eQTLs, the median correlation value is 0.219 in ESN and 0.249 in GBR for EN comb; 0.219 in ESN and 0.249 in GBR for fused lasso; and 0.153 in ESN and 0.199 in GBR for EN anc (Table B.1). Prediction quality is also slightly lower for ESN than GBR, which is more noticeable when the number of causal variants is larger.

Different genetic architecture between ancestral populations

For cases where eQTL architecture is not fully shared across populations, the prediction quality of each method varies across simulation scenarios. In the scenario where populations share 50% of eQTLs and eQTLs have the same effect sizes, all three methods perform similarly, as shown in Figure 4.2 (top panel). However, the more differences in eQTL architecture between the populations, the less cross-population generalizability there is. EN comb has the least accurate performance when the positions of causal eQTLs are completely different in the two populations, even though the effect sizes are the same. For example, with $k = 2$ causal eQTLs that have the same effect sizes but different positions in the two populations, the median correlation value is 0.178 in ESN and 0.170 in GBR for EN comb; 0.283 in ESN and 0.278 in GBR for EN anc; and 0.225 in ESN and 0.220 in GBR for fused lasso (Table B.2 and Figure 4.2). In this case, fused lasso outperforms ancestry-combined models EN comb,



(a) genetic architecture

(b) simulation results

Figure 4.1: **Simulation results comparing fused lasso vs elastic net: similar genetic architecture.** (a) Each vertical bar depicts one copy of a chromosome, and colors represent the ancestral origin. Stars refer to the locations of eQTL on the chromosome and the signs (+/-) refer to the direction of the corresponding effect sizes. (b) Boxplots of correlations between predicted and simulated phenotypes comparing the three methods: fused lasso, -specific elastic net and combined set elastic net. On panels from left to right, the number of causal variants is varied: 1, 2, and 10. Test set populations are color-coded: yellow refers to European-ancestry population (GBR) and purple refers to African-ancestry population (ESN).

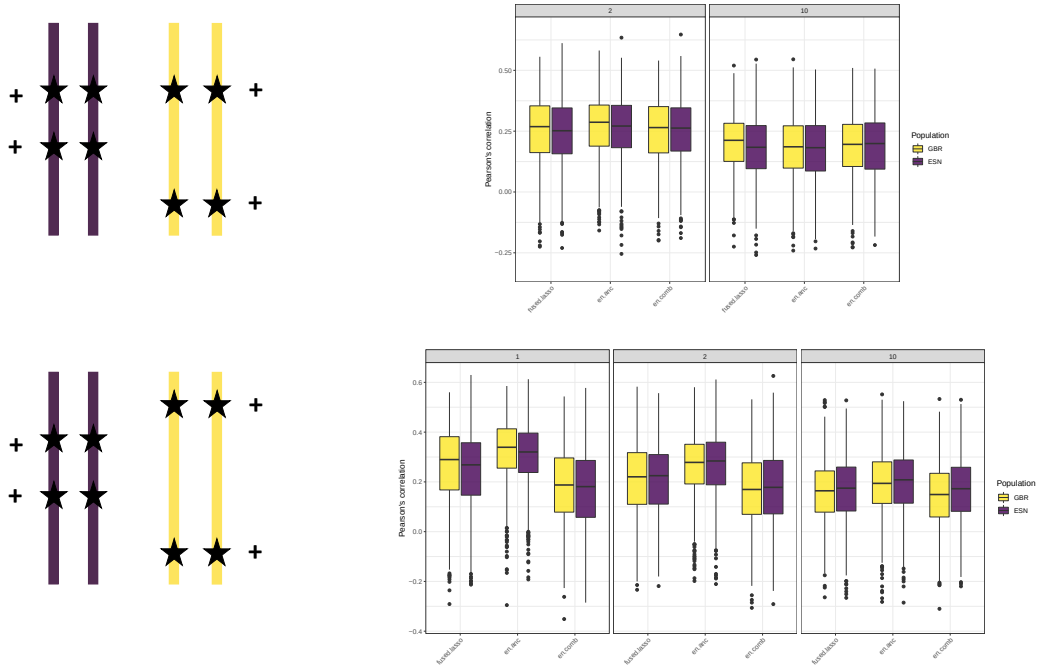
but its performance is not as accurate as the performance of ancestry-specific models EN anc.

The lack of cross-population generalizability gets worse when the effect sizes of the causal eQTLs are in the opposite directions in the two populations. Combining samples from different populations in the same training set and using EN comb leads to the worst performance in simulation scenarios where causal eQTLs have opposite effect sizes in the two populations (Figures 4.3, 4.4, B.3, and B.4). Prediction accuracy of EN comb is lower when the proportion of causal eQTLs with the opposite direction effect sizes is higher in the two populations. Again, fused lasso's prediction accuracy is

slightly lower than the accuracy of the ancestry-specific methods (EN anc, lasso anc) but higher than the accuracy of ancestry-combined methods(EN comb, lasso comb) (Figure 4.3 and Figure B.3).

Finally, in the setting where causal eQTLs are located in different positions and the direction of the effects is different in the two populations, ancestry-specific methods (EN anc and lasso anc) perform the best (Figure 4.4). As expected, when eQTL architectures are different between the populations, combining samples into one training set leads to poor prediction performance. On the other hand, separating the two populations into two training sets helps identify population-specific predictors.

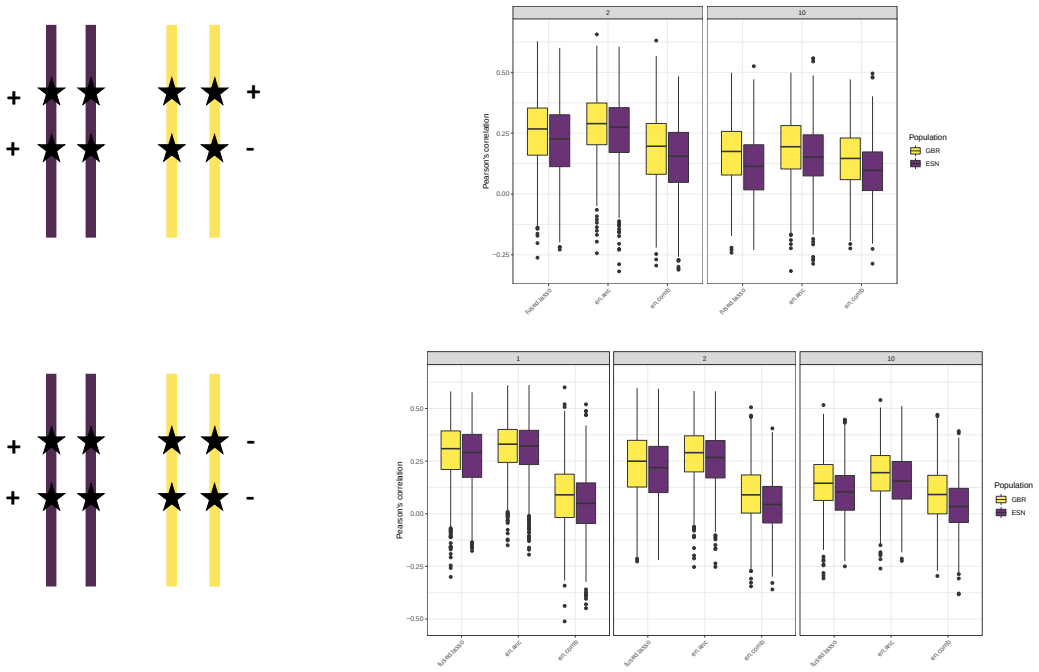
Our investigation into the architecture of gene expression indicates that the accuracy of a prediction method is mostly determined by the degree of shared eQTLs across populations. Under the 100% shared eQTL scenarios, combining multiple populations into one training set leads to the highest accuracy predictions. However, the opposite is true when eQTL architectures differ across populations and it is best to use ancestry-specific training sets. In situations when true eQTL architecture is unknown, fused lasso offers a 'hybrid' solution that has more flexibility. It enables some degree of cross-population prediction, while allowing for population-specific eQTLs.



(a) genetic architecture

(b) simulation results

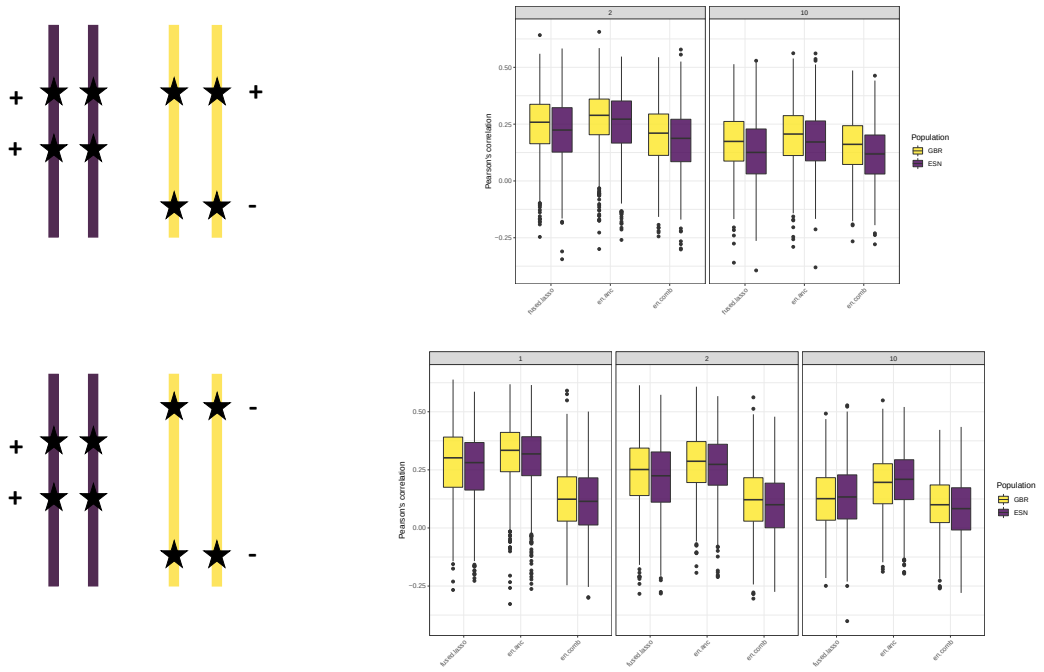
Figure 4.2: **Simulation results comparing fused lasso vs elastic net: different location of eQTLs, same direction of effects.** The top panel corresponds to the genetic architecture where two ancestral populations have 50% shared eQTLs and the same effect sizes; the bottom panel corresponds to a scenario where populations share 0% eQTLs, but the eQTLs have the same effect sizes. (a) Each vertical bar depicts one copy of a chromosome, and colors represents the ancestral origin. Stars refer to the locations of eQTL on the chromosome and signs (+/-) refer to the direction of the corresponding effect sizes. (b) Boxplots of correlations between predicted and simulated phenotypes comparing the three methods: fused lasso, ancestry-specific elastic net and combined set elastic net. On panels from left to right, the number of causal variants is varied. Test set populations are color-coded: yellow refers to European-ancestry population (GBR) and purple refers to African-ancestry population (ESN).



(a) genetic architecture

(b) simulation results

Figure 4.3: **Simulation results comparing fused lasso vs elastic net: same location of eQTLs, different direction of effects.** The top panel corresponds to the genetic architecture where two ancestral populations have the same eQTLs but 50% of the eQTLs have different direction effect sizes; the bottom panel corresponds to a scenario where populations share eQTLs, but 100% of the eQTLs have the opposite direction effect sizes. (a) Each vertical bar depicts one copy of a chromosome, and colors represent the ancestral origin. Stars refer to the locations of eQTL on the chromosome and signs (+/-) refer to the direction of the corresponding effect sizes. (b) Boxplots of correlations between predicted and simulated phenotypes comparing the three methods: fused lasso, ancestry-specific elastic net and combined set elastic net. On panels from left to right, the number of causal variants is varied. Test set populations are color-coded: yellow refers to European-ancestry population (GBR) and purple refers to African-ancestry population (ESN).



(a) genetic architecture

(b) simulation results

Figure 4.4: **Simulation results comparing fused lasso vs standard lasso: different location of eQTLs, different direction of effects.** The top panel corresponds to the genetic architecture where two ancestral populations have 50% shared eQTLs and 50% of the same effect sizes; bottom panel corresponds to a scenario where populations share 0% eQTLs and the eQTLs have the opposite direction effect sizes. (a) Each vertical bar depicts one copy of a chromosome, and colors represent the ancestral origin. Stars refer to the locations of eQTLs on the chromosome and signs (+/-) refer to the direction of the corresponding effect sizes. (b) Boxplots of correlations between predicted and simulated phenotypes comparing the three methods: fused lasso, ancestry-specific elastic net and combined set elastic net. On panels from left to right, the number of causal variants is varied. Test set populations are color-coded: yellow refers to European-ancestry population (GBR) and purple refers to African-ancestry population (ESN).

4.3.2 Application to the TOPMed MESA cohort

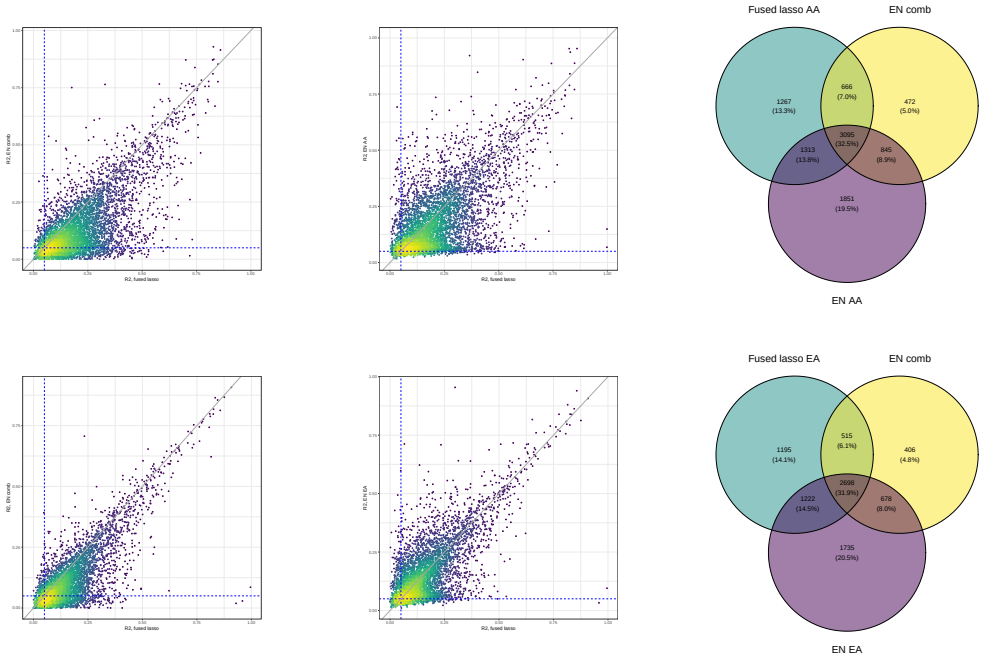
Data description We used 556 unrelated individuals from the MESA study. The dataset included 278 African American (AA) individuals and 278 individuals of European ancestry (EA) with whole-genome sequencing and RNA-sequencing measurements from peripheral blood mononuclear cells (PBMCs) from Exam 1. We used whole-genome sequencing data from TOPMed Freeze 8 dataset where reads were aligned to human-genome build GRCh38. Variant quality control (QC) included removing variants based on Mendelian discordance, a support vector machine (SVM) quality filter and excess heterozygosity filter. Additionally, we checked for concordance between annotated and inferred genetic sex, concordance between prior array genotype data and TOPMed WGS data, and pedigree checks. Details regarding the genotype freezes, laboratory methods, data processing, and quality control are described on the TOPMed website and in a common document accompanying each studys dbGaP accession (https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8). All participants provided informed consent and the protocols of MESA were approved by the IRBs of collaborating institutions and the National Heart, Lung and Blood Institute.

Model training We used AA, EA and AA+EA (combined) training sets to train gene expression prediction models using fused lasso, standard lasso and elastic net (EN). For each gene, we selected variants located within 500kb-window of the gene's start and end sites and used them in model training. Tuning parameters were chosen via 3-fold cross-validation. The model training scheme resulted in eight sets of prediction weights: elastic net AA, elastic net EA, elastic net combined, lasso AA, lasso EA, lasso combined, fused lasso EA, and fused lasso AA. Finally, we calculated R^2 for each model and each method and compared performances of fused lasso vs elastic net models and fused lasso vs standard lasso models, in EA and AA populations separately.

We attempted to train transcriptome prediction models for a total of 12,691 genes.

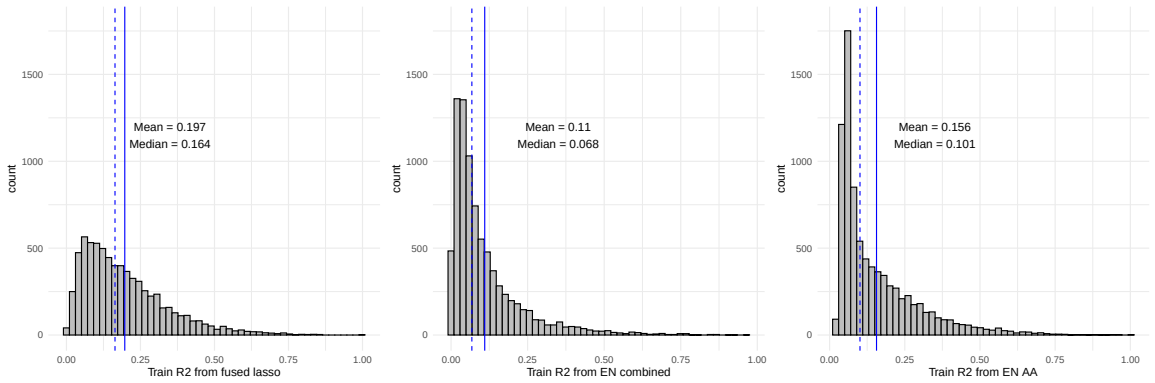
In MESA AA, we obtained 7,090 models with fused lasso AA; 8,323 with EN combined 8,309 with lasso combined; 8,390 using EN AA and 8,398 using lasso AA. Comparably, training in MESA EA resulted in 6,864 models using fused EA; 8,323 models using EN combined and 8,309 using lasso combined; 8,300 using EN EA and 8,221 using lasso EA. Based on model $R^2 > 0.05$ from the training results, 3,095 genes were well-predicted in MESA AA training set using fused lasso and two EN methods (Figure 4.5c (top panel)). However, 1,267 genes were only well-predicted with fused lasso AA, 1,851 with EN AA, and 472 with EN combined. Similarly, there was a total of 2,698 genes that were well-predicted in MESA EA training set using all three methods, but only 1,195 genes were well-predicted in fused lasso EA; 1,735 genes in EN EA; and 406 genes in EN comb (Figure 4.5c).

Comparing the distributions of model R^2 from training sets in all genes, fused lasso performs slightly better than ancestry-specific EN or EN trained on a combined set (Figure 4.6), both in AA and EA. We see similar trends when comparing training R^2 from standard lasso models to fused lasso (Figure B.6). Based on mean and median model R^2 values, fused lasso outperforms elastic net and standard lasso models both in EA and AA (Figure 4.6 and B.6). However, we observe striking differences in model R^2 for some genes. For example, the gene *CSNK1D* is well predicted by fused lasso and EN AA, but very poorly predicted by EN comb (R^2 values of 0.72, 0.34, and 0.02, respectively, in AA; 0.49, 0.08, and 0.08, respectively, in EA). In contrast, gene *MTCL1* is well predicted by EN comb ($R^2 = 0.34$) in AA, but has only $R^2 = 0.03$ with fused lasso and has no model available with EN AA.

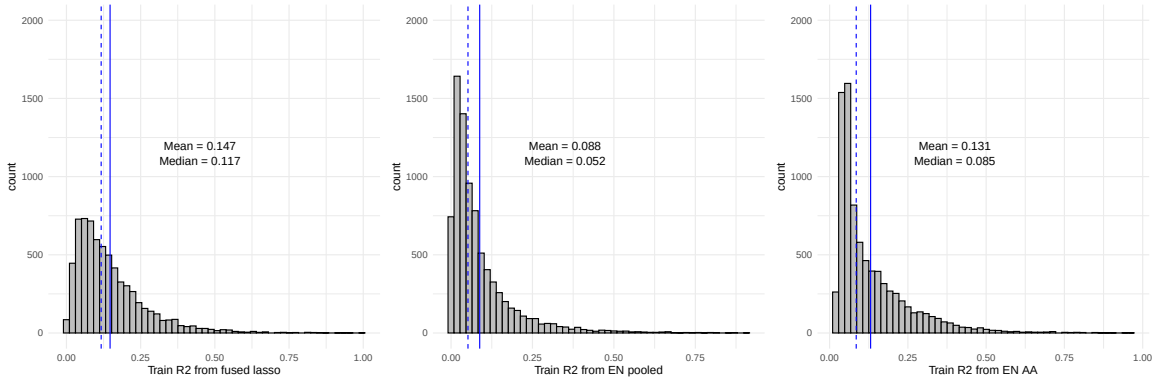


(a) (b) (c)

Figure 4.5: **Performance of fused lasso and elastic net models in training sets, based on model R^2 of all genes.** Smooth scatter plots of R^2 values between measured gene expression and predicted transcriptome values in MESA AA (top panel) and MESA EA (bottom panel) training sets. (a) Comparison of model R^2 values using EN comb vs fused lasso models; (b) Comparison of model R^2 values using ancestry-specific EN vs fused lasso models; (c) Venn diagram showing the overlap of well-predicted genes (with $R^2 > 0.05$) by fused lasso and elastic net models in training sets.



(a) MESA AA training set



(b) MESA EA training set

Figure 4.6: **Histograms of the model R^2 distribution of all genes comparing fused lasso and elastic net.** Histograms showing model R^2 for three methods in (a) African American and (b) European ancestry training sets. The blue solid line denotes the mean of model R^2 values; the blue dashed line denotes the median of model R^2 values. All genes, including those that fall below the $R^2 = 0.05$ cut-off, are displayed.

Assessment of prediction models To assess the performance of the prediction models, we applied each set of prediction weights to participants in MESA AA and MESA EA test sets and predicted gene expression values. We compared the performance of different methods by calculating R^2 between measured and predicted

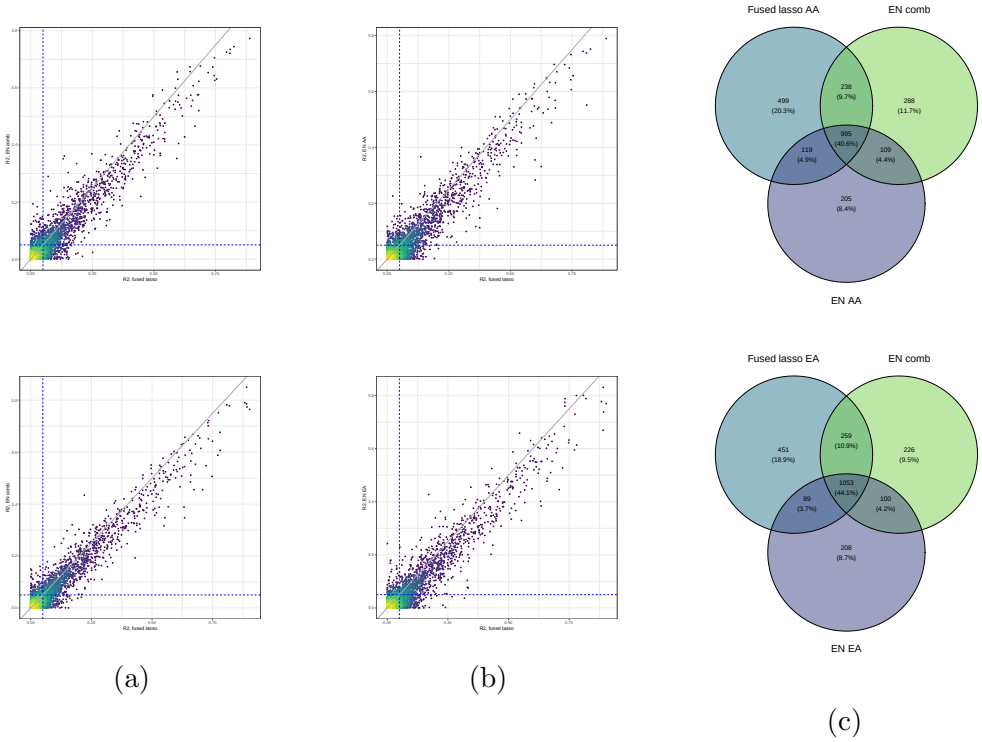
gene expression values for each gene. Additionally, we calculated mean square error (MSE) for each gene and performed hypothesis testing of pairwise differences in MSEs of fused lasso and other methods.

Based on the true R^2 calculated from measured and predicted transcriptome values in test data, fused lasso performs better than elastic net and standard lasso. In MESA AA test set, the median R^2 values are 0.016 for fused lasso AA; 0.010 and 0.009 for EN comb and EN AA, respectively; and 0.010 and 0.009 for lasso comb and lasso AA, respectively. Similarly, in MESA EA test set, the median R^2 values are 0.016 for fused lasso EA; 0.010 and 0.009 for EN comb and EN EA, respectively; and 0.010 and 0.008 for lasso comb and lasso EA, respectively. When comparing genes found only in fused lasso and EN comb or fused lasso and EN ancestry-specific panels, we found that fused lasso performs better than elastic net in prediction of gene expression in MESA test set, based on true R^2 values (Figure 4.7a, b). We observed the same trend and came to the same conclusion when comparing the performance of fused lasso to standard lasso models (Figure B.7a, b).

As expected, the number of well-predicted (with true $R^2 > 0.05$) genes is lower in the MESA test set, compared to the MESA training set (Figure 4.5c, Figure 4.7c, Figure B.7c). Overall, there is a lot more consistency in prediction accuracy across different methods in the test set, i.e. genes that predicted well with one method are more likely to be predicted well with another method. In contrast, there was variability in how accurate prediction was for a given gene across different methods in training data (Figure 4.5a, b and Figure B.5a, b).

Finally, we calculated the MSEs for each gene across all methods in each testing set and made pairwise comparisons of fused lasso to existing methods, based on the MSE values. In particular, we tested the null hypothesis of no difference in MSE of fused lasso vs MSE of each elastic net and standard lasso methods. In each pair of methods (fused lasso vs existing methods), we rejected the null hypothesis of no difference in MSEs in fused lasso and other method and concluded that fused lasso had lower MSEs

than existing methods (p-value $< 2.2 \times 10^{-16}$ for each pair of methods) both in MESA EA and MESA AA testing sets, based on the one-sided Wilcoxon signed-rank test.



(a) (b) (c)

Figure 4.7: **Performance of fused lasso and elastic models in test sets, based on true R^2 of all genes.** Smooth scatter plots of R^2 values between measured gene expression and predicted gene expression values in MESA AA (top panel) and MESA EA (bottom panel) test sets. (a) Comparison of true R^2 values using EN comb vs fused lasso models; (b) Comparison of true R^2 values using ancestry-specific EN vs fused lasso models; (c) Venn diagram of well-predicted ($R^2 > 0.05$) genes in test sets.

4.4 Discussion

Despite recent advances in methodology for TWAS, some limitations remain when it comes to gene expression prediction models for non-European populations. Here, we

proposed a transcriptome prediction method for ancestrally diverse populations. In particular, we developed a flexible fused lasso framework that incorporates participants' ancestry labels to accommodate ancestry-specific effects sizes while borrowing information across ancestral groups. Fused lasso includes both a sparsity (l_1) penalty and a fusion penalty. The sparsity penalty performs predictor selection while the fusion penalty encourages similarity of ancestry-specific weights for the same predictor.

Although there is evidence that effect sizes of true causal eQTLs are similar across ancestries in most cases, that is not true for effect sizes of variants tagging causal eQTLs [14, 49]. Their effect sizes are expected to differ across populations due to differences in ancestral allele frequencies and the linkage disequilibrium patterns. The fused lasso approach estimates ancestry-specific effect sizes by jointly modeling a set of variants together with their ancestry label. It allows to detect with greater power predictors that have weak associations in each population separately but common to all the populations considered. At the same time, it can detect eQTLs that are present and have strong associations in only one of the populations but are absent in other populations. Therefore, the flexibility of fused lasso positively impacts predictions for many more genes in the genome than only genes with different location and effect sizes of causal eQTLs across ancestry groups.

We performed extensive simulation studies in which we assessed the prediction accuracy of fused lasso and compared it to the prediction accuracy of existing methods trained on pooled-ancestry and ancestry-specific datasets. We showed that training on ancestry-specific datasets has an advantage over training on pooled datasets, primarily when genes have different eQTL architectures across ancestral populations, i.e. there are effect size differences across groups. An extreme case of this is when an eQTL is only present in one ancestry, but not others. Conversely, when there is no difference in the estimated effect size between ancestry groups, there are gains in prediction accuracy from pooling participants into one training set. However, since fused lasso incorporates ancestry information, it can accommodate multiple scenarios.

After testing our framework in simulations, we applied our fused lasso method to RNA-seq and whole genome sequencing data of African American and European-ancestry individuals from the TOPMed MESA cohort. We found that fused lasso improved prediction accuracy, measured by R^2 and mean squared error, over traditional methods in both the MESA AA and MESA EA test sets. Our method demonstrated advantages over traditional methods, particularly when eQTL effect sizes are heterogeneous across populations.

Our method has a number of limitations. First, current prediction models would be best suited for gene expression predictions in PBMCs or relatively similar blood cell tissues. In other tissues, we would expect them to underperform, depending on eQTL sharing across tissues. Second, participants were classified as African American or European-ancestry based on self-report and such self-identified groupings may include a variety of genetic ancestry backgrounds. Instead, we could extend our method to accommodate admixture. In place of population labels, we could incorporate local ancestry information, inferred with RFMix or similar software, since separate models are built for each gene. The statistical framework can also accommodate three or more populations. However, it will require reference panels with transcriptome data for all the groups included. Sample sizes for transcriptome datasets from non-European populations still remain limited. In order to improve the quality of transcriptome prediction and TWAS discoveries, in general, it is important to expand RNA-sequencing efforts in diverse populations. Future work will be needed to evaluate prediction accuracy and optimize performance of fused lasso models for multi-population datasets.

In summary, using a flexible fused lasso framework, we were able to construct ancestry-informed transcriptome prediction models. This approach provides a number of benefits over traditional approaches, elastic net and lasso, that use either trans-ethnic or ancestry-specific training sets. This prediction framework is designed to be flexible; it can be easily extended to accommodate more than two populations and

include local ancestry information for admixed populations, once more transcriptomic datasets for diverse populations become available.

Chapter 5

MIXED MODELS FOR TRANSCRIPTOME-WIDE ASSOCIATION STUDIES IN STRUCTURED SAMPLES

5.1 Introduction

Genome-wide association studies (GWAS) have interrogated millions of genetic variants across the genomes of millions of study participants to identify genetic associations with numerous phenotypes. Since 2005, more than 50,000 associations of genome-wide significance have been reported between genetic variants and a large number of complex traits [65]. These findings have led to a better understanding of disease etiology, advances in clinical care and personalized medicine. However, despite these successes, GWA studies face multiple challenges.

Firstly, GWAS are penalized by the multiple testing burden. Small-scale GWAS are underpowered to detect small or moderate effect sizes, since association signals have to reach a significance threshold level of $p < 5 \times 10^{-8}$. This limitation can be overcome by increasing sample size, but this approach is often not possible. Studies on isolated populations, cost and difficulty of measuring the phenotype are some examples that lead to smaller sample sizes. Additionally, genome-wide significant associations that have been identified account for only a small proportion of estimated heritability of most complex traits. Finally, most GWAS do not identify causal variants and genes and the vast majority of association signals lie in non-coding, intronic or intergenic areas of the genome, thus functional relevance of these loci remains unclear. Because of this, additional steps, such as fine-mapping and functional follow-up, are required to elucidate biological mechanisms underlying the variant-trait associations.

To overcome the aforementioned limitations of GWAS, a family of gene-based

approaches that leverage transcriptomic data have been developed. These gene-based approaches are built on the assumption that gene expression is regulated by genetic variation, thus, may be useful in understanding biological mechanisms underlying disease. Moreover, GWAS-identified genetic variants and variants that alter gene expression levels often overlap. PrediXcan, proposed by Gamazon et al [21], leverages genotypic and transcriptomic data from the Genotype-Tissue Expression (GTEx) Project [34] to aggregate variants into gene-based units. PrediXcan uses machine learning methods to train prediction models that include a set of genetic variants that influence gene expression in a given tissue. Later, these models are applied to genetic variants extracted from independent datasets to predict transcriptome values. Finally, predicted transcriptome values are tested for an association with the phenotype, which we refer to as transcriptome-wide association study. Since these methods use the gene as a relevant unit of analysis, they have multiple advantages over GWAS. They reduce the multiple testing burden, help prioritize candidate genes underlying GWAS and can be conducted at multiple phenotype-relevant tissues, that may not have been collected on study subjects.

The majority of recent work in TWAS has focused on various approaches of imputing gene expression from genotype data or summary statistics. However, the majority of the methods focus on using European ancestry-based models and conducting studies on European-ancestry subjects. Far less attention has been given to the challenges of conducting TWAS in large-scale multi-ethnic datasets. We aim to adapt a linear mixed model (LMM) framework, which has become a very popular approach for association mapping in GWAS, for transcriptome-wide association testing in related samples with population structure. LMMs may include principal components as fixed effects to correct for population structure and a genetic relationship matrix (GRM) as a random effect to account for relatedness among sample individuals.

Here, we propose a novel method for conducting TWAS that includes random effects accounting for both background polygenic effects and transcriptomics, predicted

or measured. Variance components of the random effects are calculated via average information-restricted maximum likelihood (AI-REML) with two empirical relationship matrices – one for the genotypic data and one for the transcriptomic data. Through simulation studies, we demonstrate that inclusion of an additional random effect that accounts for transcriptome can provide an improvement, in terms of both type I error and power, over a LMM with a single random effect and a GRM. Finally, we apply our method to a cohort of 29,696 ancestrally diverse individuals from TOPMed study with PrediXcan-predicted transcriptome to identify genes associated with white blood cell count.

5.2 Methods

We first give an overview of existing methods for conducting GWAS in samples with population structure and familial relatedness. We then show how this framework can be extended for transcriptome-wide association testing. We propose a method that has two random effects. We outline the methods used to estimate the proposed random effects and how to apply them in a transcriptome-wide association analysis.

5.2.1 Linear mixed model framework for GWAS

Consider a set of sampled individuals $\mathcal{N}$ that have been sequenced or genotyped at a set of $\mathcal{S}$ variants and let $\mathbf{y}$ be a vector of length N of values for some quantitative phenotype. Let bold uppercase letters denote matrices, while bold lowercase letters indicate vectors. When performing a GWAS, each variant is often tested individually, assuming an additive, polygenic background of the trait. To test an association of $\mathbf{y}$ with a particular variant $s \in \mathcal{S}$ in a homogeneous sample of unrelated subjects, we fit a linear regression model:

$$\mathbf{y} = \beta\mathbf{g} + \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}),$$

where $\mathbf{g}$ is a vector of length N of genotypes for variant s with the corresponding effect size β on the trait value, $\mathbf{W}$ is a matrix of covariates including an intercept with a corresponding vector of effect sizes $\boldsymbol{\alpha}$, and $\boldsymbol{\epsilon}$ is a random variable that captures independent residual effects, σ^2_ϵ is a parameter that measures residual variance, and $\mathbf{I}$ is an identity matrix. In this case, these random environmental effects are i.i.d. Gaussian.

More recent studies include diverse samples with cryptic or known familial relatedness. Linear mixed models have become a popular approach for conducting GWAS in such samples. Population structure in the sample can act as a confounder for the genetic variant - phenotype relationship and result in spurious associations. To account for population structure, it is common to include a matrix of genetic principal components as fixed effect covariates. Additionally, we can include random effects in the model to account for relatedness.

Using mixed effects linear models, we can fit the model testing for an association of $\mathbf{y}$ with a particular variant $s \in \mathcal{S}$:

$$\mathbf{y} = \beta\mathbf{g} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\epsilon},$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_G\boldsymbol{\Phi}),$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\epsilon\mathbf{I}),$$

where $\mathbf{g}$, β, $\mathbf{W}$, $\boldsymbol{\alpha}$, $\boldsymbol{\epsilon}$ are defined above, $\mathbf{V}$ is the matrix of ancestry representative vectors with corresponding effect sizes $\boldsymbol{\gamma}$ to adjust for population structure, $\mathbf{u}$ is a random effect that captures the polygenic effect of genotypes, $\boldsymbol{\Phi}$ is the genetic relationship matrix, σ^2_G corresponds to the total additive genetic variance of the trait.

Methods, such as KING-robust [37] and PC-Relate [17] can be used to estimate an empirical genetic relationship matrix (GRM), $\hat{\boldsymbol{\Phi}}$, from observed genotypes or whole-genome sequencing data. A standard estimator for the $[i,j]^{th}$ element of $\hat{\boldsymbol{\Phi}}$ is:

$$\hat{\phi}_{ij} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \frac{(x_{is} - 2\bar{p}_s)(x_{js} - 2\bar{p}_s)}{2\bar{p}_s(1-\bar{p}_s)}, \tag{5.1}$$

where $|\mathcal{S}|$ is the number of variants, x_{is} is the genotype value for individual i at variant s, and $\bar{p}_s = \frac{1}{2|\mathcal{N}|}\sum_{i\in\mathcal{N}} x_{is}$.

5.2.2 Linear mixed model framework for TWAS

We propose to extend the LMM framework from GWAS to TWAS in samples with population structure and familial relatedness.

Consider a set of sampled individuals $\mathcal{N}$ with predicted transcriptome values for $\mathcal{M}$ genes and let $\mathbf{y}$ be a vector of length N of values for some quantitative phenotype. When performing a TWAS, we test each gene individually, like we would test each genetic variant in GWAS. We can fit the following mixed-effects model for testing an association of $\mathbf{y}$ with transcriptome levels $\mathbf{t}$ for gene $m \in \mathcal{M}$:

$$\mathbf{y} = \beta\mathbf{t} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\epsilon}, \tag{5.2}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_G^2\boldsymbol{\Phi}),$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}),$$

where $\mathbf{t}$ is a vector of length N of predicted transcriptome values for gene m for all individuals with the corresponding effect size β. Note that polygenic effects of genotypes, $\mathbf{u}$, are still included in the model.

We refer to the model in Equation 5.2 as the 'GRM model.'

5.2.3 Linear mixed model approach with GRM and TRM

We propose a new linear mixed model approach for TWAS, an LMM which includes an additional random effect $\mathbf{v}$ to capture the background effects of transcriptome on the phenotype:

$$\mathbf{y} = \beta\mathbf{t} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\gamma} + \mathbf{u} + \mathbf{v} + \boldsymbol{\epsilon}, \tag{5.3}$$

where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\boldsymbol{\Psi})$ and $\boldsymbol{\Psi}$ is the transcriptomic relationship matrix (TRM). The total additive variance of the trait across predicted transcriptome corresponds to σ_T^2. We refer to model in Equation 5.3 as the 'GRM+TRM model.'

5.2.4 Variance of y under the null hypothesis

Generalized least squares (GLS) can be used to fit the LMM in Equation 5.4 and test H_0: $\beta = 0$ vs H_A: $\beta \neq 0$. In order to do that, we need to take into the account the correlation among all individuals. The covariance matrix of the phenotype $\mathbf{y}$ under the null hypothesis of no association between the phenotype and the gene is $\mathbf{\Sigma} = \sigma_G^2\mathbf{\Phi} + \sigma_T^2\mathbf{\Psi} + \sigma_\epsilon^2\mathbf{I}$. However, $\mathbf{\Sigma}$ is typically unknown and must be estimated.

5.2.5 Estimation of variance components

We propose to estimate an empirical transcriptome relationship matrix (TRM) from actual or predicted gene expression values. Consider a set of sampled individuals $\mathcal{N}$ with predicted transcriptome values for $\mathcal{M}$ genes. Then the $[i, j]^{th}$ element of $\hat{\mathbf{\Psi}}$ is:

$$\hat{\psi}_{ij} = \frac{\sum_{m\in\mathcal{M}}(t_{im} - \bar{t}_m)(t_{jm} - \bar{t}_m)}{\left[\sum_{m\in\mathcal{M}}(t_{im} - \bar{t}_m)^2\right]^{1/2}\left[\sum_{m\in\mathcal{M}}(t_{jm} - \bar{t}_m)^2\right]^{1/2}}, \tag{5.4}$$

where $\mathcal{M}$ is a set of genes for which transcriptome values are available, $\hat{t}_{im}$ is the transcriptome value for individual i at gene m, and $\bar{t}_m = \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} t_{im}$ is the sample average transcriptome value for gene m, as $|\mathcal{N}|$ is the number of sampled individuals.

Using the empirical GRM $\hat{\mathbf{\Phi}}$, defined in Equation 5.1, and the empirical TRM $\hat{\mathbf{\Psi}}$, defined in Equation 5.4, in the null model, we can estimate the variance components σ_G^2, σ_T^2 and σ_ϵ^2 with restricted maximum likelihood (REML). Once we have the estimated variance components, we can perform GLS using the estimate of the overall phenotypic covariance structure, $\hat{\mathbf{\Sigma}} = \hat{\sigma}_G^2\hat{\mathbf{\Phi}} + \hat{\sigma}_T^2\hat{\mathbf{\Psi}} + \hat{\sigma}_\epsilon^2\mathbf{I}$.

5.2.6 Simulation studies

Throughout our simulation studies, we use real whole-genome sequencing data from a related sample of 10,000 individuals of diverse ancestral backgrounds from the TOPMed freeze8 dataset. We impute whole blood gene expression values using PrediXcan whole blood models with effect sizes computed with a multivariate adaptive shrinkage

(*mashr*) method [21, 5, 4] built from Gene-Tissue Expression Project (GTEx) version 8. We then simulate causal effects and phenotypes from the predicted gene expression values.

For each simulation study, we generate an initial phenotype vector, $\mathbf{y_0}$, from a random vector of $\mathcal{N}(0,1)$ values and empirical GRM and TRM, such that:

$$\text{Var}(\mathbf{y}_0) = \mathbf{\Sigma}_0 = \sigma^2_G\mathbf{\Phi} + \sigma^2_T\mathbf{\Psi} + \sigma^2_\epsilon\mathbf{I}$$

We then simulate a phenotype that is correlated with predicted gene expression values of one randomly selected gene according to the model:

$$\mathbf{y} = \mathbf{y_0} + \beta\mathbf{t},$$

where $\mathbf{t}$ is a vector of predicted transcriptome values for one randomly selected gene. Gene causal effect, β, is chosen such that each gene explains 0.05% of the variability of the trait:

$$\beta = \left[\frac{h^2\text{Var}(\mathbf{y_0})}{(1-h^2)\text{Var}(\mathbf{t})}\right]^{1/2},$$

where $\text{Var}(\mathbf{y_0})$ is the variance of the initial phenotype, and $\text{Var}(\mathbf{t})$ is the variance of the predicted expression values for the selected gene.

Genes are tested for an association with each simulated phenotype using three methods:

1) linear mixed model with GRM and TRM,

2) linear mixed model with GRM only,

3) and linear mixed model with TRM only.

5.2.7 Application to WBC trait in TOPMed samples

Transcriptome-wide association tests were performed using three linear mixed models approaches. First, we fit a model, referred to as the 'null model', that assumed

no association between the outcome and any predicted transcriptome values. In each null model, we included sex, age at trait measurement, a variable indicating TOPMed study and phase of genotyping (study_phase), and the first 11 PC-Air [16] principal components (PCs) of genetic ancestry as fixed effects. To account for genetic relatedness, we included a 4th-degree sparse empirical kinship matrix (KM) computed with PC-Relate [17], a TRM or both. Details on the computation of the PCs, the TRM and the sparse KM are provided in Appendix C.

In order to improve power and appropriately control Type I error when a trait has a non-Normal distribution, we used a fully-adjusted two-stage approach for fitting the null model [61]. In stage 1, we fit an LMM with untransformed phenotype values as an outcome, the fixed effects covariates, and random effects (KM, TRM or both). We applied a rank-based inverse-normal transformation to the residuals and then rescaled them by the original variance. In stage 2, we fit another LMM using the residuals obtained in stage 1 as an outcome and the same covariates and random effects, as in Stage 1. Finally, we used the output from stage 2 to perform TWA tests. All analyses were performed using the R/Bioconductor package `GENESIS` [22].

5.3 Results

5.3.1 Data description

The analyzed cohort consisted of 29,696 participants from 6 TOPMed studies: Coronary Artery Risk Development in Young Adults (CARDIA, $n = 3,042$), Framingham Heart Study (FHS, $n = 3,136$), Hispanic Community Health Study - Study of Latinos (HCHS-SOL, n=7, 225), Jackson Heart Study (JHS, $n = 2,841$), Multi-Ethnic Study of Atherosclerosis (MESA, n=2, 517), and Women's Health Initiative (WHI, $n = 10,935$). The composition of the 29,696 participants by race/ethnicity is 50% White, 21% Black, 28% Hispanic/Latino, and 1% Asian. Further descriptions of the design of the participating TOPMed cohorts and the sampling of individuals within each cohort for

TOPMed are provided in the section 'Participating studies' in Appendix C. Whole genome sequencing (WGS) was performed as part of the NHLBI TOPMed program, at an average depth of 38x using Illumina X10 technology and DNA from blood. UW DCC performed variant quality control (QC) using standard measures, such as concordance between annotated and inferred genetic sex, concordance between prior array genotype data and TOPMed WGS data, and pedigree checks. Details regarding laboratory methods, data processing, and quality control are described on the TOPMed website and in a common document accompanying each study's dbGaP accession (https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8).

We used PrediXcan whole blood *mashr* models from PredictDB to predict genetically regulated levels of expression, further referred to as predicted transcriptome, for each individual in the dataset. Among the 12,623 genes in the PrediXcan database, 12,563 genes had predicted transcriptome values in our dataset and were used to compute an empirical TRM as described in Equation 5.4.

The total white blood cell (WBC) count trait considered for analyses was measured from freshly collected whole blood samples at local clinical laboratories using automated hematology analyzers calibrated to manufacturer recommendations according to clinical laboratory standards. WBC is defined as the count of white blood cells in the blood, by number concentration in thousands per microliter. In studies where multiple white blood cell measurements per participant were available, we selected a single measurement for each participant and excluded outliers to minimize the possibility of measurement error.

5.3.2 Simulation study

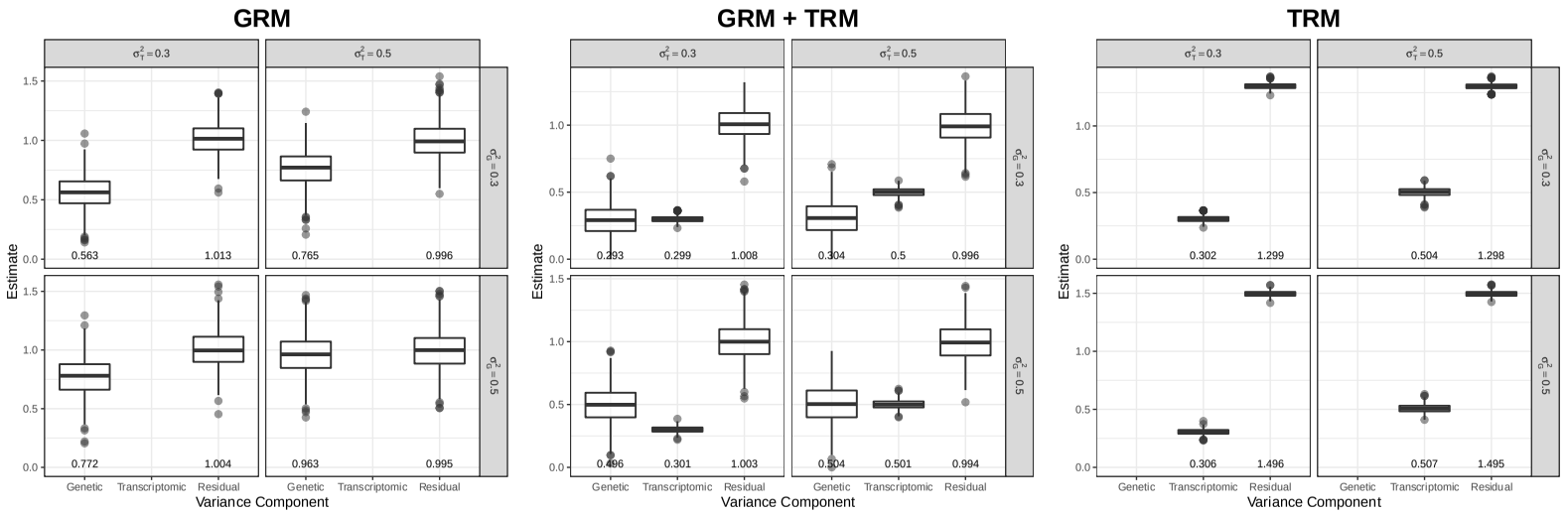
We performed simulation studies to validate the proposed method that includes two empirical relationship matrices, GRM and TRM, and to compare its performance to the methods that included only one of the two random effects. We demonstrate that using LMM that only utilizes a GRM for TWAS results in inflated type I error at null genes and loss of power to detect associations. We also demonstrate that including a TRM, in addition to modeling the trait covariance structure with a GRM, can correct this miscalibration via controlling Type-I error rate.

5.3.3 Estimation of variance components in simulated samples

We compared estimates of the variance components using three models: a model with GRM and TRM, a model including GRM only, and a model including TRM only random effects. We used the average information REML to estimate variance components for the three models. We varied σ^2_G, $\sigma^2_T \in \{0.3, 0.5\}$ but kept $\sigma^2_\epsilon = 1$ constant in all scenarios. We performed 1,000 simulations for each combination of parameters and fit three null models for each simulation.

Figure 5.1 shows boxplots of the estimates of the variance components from 1,000 simulations when fitting each of the three linear mixed models. Figure 5.1(b) presents the variance components estimates under the GRM+TRM model, which is the true underlying model in our simulations.For each simulated value of σ^2_G, σ^2_T and σ^2_ϵ, the estimates of variance components are centered around the true values. In contrast, when we ignore one of the random effects, the estimates of the variance components are inflated. As we can see from Figure 5.1 (a), when we do not account for the transcriptomic random effect, we overestimate the variance due to the genetic effect, σ^2_G. Similarly, when we ignore the genetic random effect and use only the transcriptomic random effect in the model, the estimate of the residual variance component σ^2_ϵ is overestimated, as we can see in Figure 5.1(c).

When we misspecify the correlation structure by not including all random effects, the estimates of variance components tend to be inflated. This happens because the estimated covariance structure attempts to account for unmodeled random effects. For example, when we only include GRM in the model, the estimate of the genetic variance component attempts to capture transcriptomic random effect, thus leading to inflated estimates of σ^2_G (Figure 5.1(a)). Now, when we only include TRM in the model, the estimate of the residual variance component is inflated, as it tries to scale up to incorporate the covariance structure explained by the genetic random effect (Figure 5.1(c)).



(a) (b) (c)

Figure 5.1: **Boxplot of estimates of variance components from the simulation study.** Estimates of variance components from 1,000 simulation iterations for σ_G^2, $\sigma_T^2 \in \{0.3, 0.5\}$ and $\sigma_\epsilon^2 = 1$. Plot (a) corresponds to the model with GRM only, plot (b) corresponds to the model with both GRM and TRM, and plot (c) corresponds to the model with TRM only. The mean estimate for each boxplot is presented on the bottom.

5.3.4 Evaluation of power and type I error

We assessed the type I error rate and power of the three models by performing transcriptome-wide association tests with simulated data. To examine how the three approaches control the type I error, we calculated the proportion of false positives, based on the association tests carried out with non-causal genes. We summarized the type I error rates for the three approaches for various significance thresholds and values of σ_G^2 and σ_T^2 in Table 5.1. From the simulation study, we observe multiple-folds more the amount of false positives when we do not include the TRM and only use the GRM in the model. At the same time, GRM+TRM and TRM approaches yield levels of type I error close to expected. Among the three methods, GRM+TRM and TRM approaches are properly calibrated in the simulation settings considered.

To assess the power of the three approaches, we calculated the false negative rate based on the association tests carried out with causal genes and the true positive rate based on the association tests carried out with null genes. We compared the false positive rate to the true positive rate for a sequence of significance levels ranging from 5×10^{-6} to 0.05. Power for the GRM+TRM model, the GRM only model and the TRM only model is shown in Figure 5.2. The GRM model has the lowest power in all simulation scenarios considered, while the GRM+TRM and the TRM models achieve higher power. Moreover, this difference in power between the GRM only model and models that include transcriptomic random effect (TRM and GRM+TRM) increases with higher value of the transcriptomic variance component (see lower right panel of Figure 5.2).

Table 5.1: Type I error rate for the three LMM approaches of 10,000 simulations of TWAS.

		$\alpha = 5e-05$			$\alpha = 5e-04$			$\alpha = 5e-02$		
σ_G^2	σ_T^2	GRM	GRM+TRM	TRM	GRM	GRM+TRM	TRM	GRM	GRM+TRM	TRM
0.3	0.3	$1.93e-03$	$3.17e-05$	$3.17e-05$	$4.46e-03$	$6.34e-04$	$6.42e-04$	$7.79e-02$	$5.09e-02$	$5.11e-02$
0.3	0.5	$3.58e-03$	$5.55e-05$	$4.75e-05$	$7.32e-03$	$4.36e-04$	$4.52e-04$	$8.77e-02$	$4.92e-02$	$4.96e-02$
0.5	0.3	$1.34e-03$	$5.55e-05$	$5.55e-05$	$3.56e-03$	$4.52e-04$	$4.36e-04$	$7.55e-02$	$5.03e-02$	$5.10e-02$
0.5	0.5	$3.07e-03$	$4.75e-05$	$4.75e-05$	$6.42e-03$	$4.67e-04$	$4.91e-04$	$8.37e-02$	$4.93-02$	$4.99e-02$

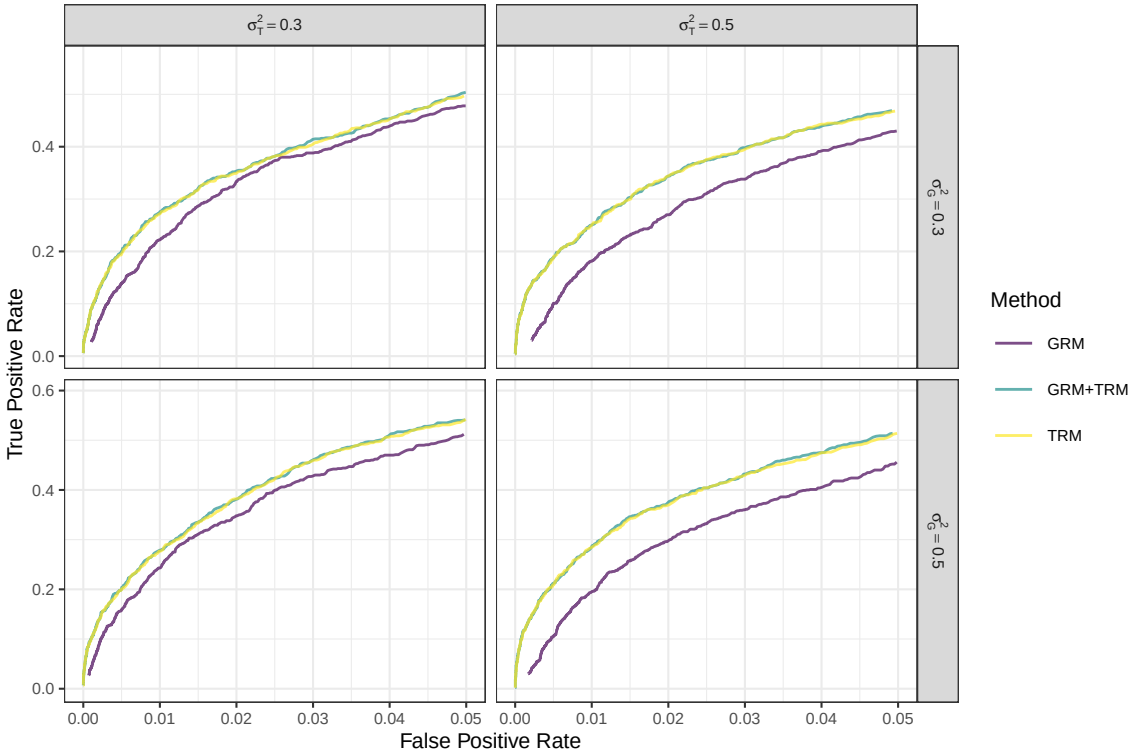


Figure 5.2: **Comparison of true and false positive rates.** The proportion of true positive associations identified (true positive rate) is compared to the proportion of null genes incorrectly identified as associations (false positive rate) by GRM only, GRM + TRM, and TRM only models in 1,000 simulations. A curve with a higher area under it indicates better performance.

5.3.5 TWAS of WBC trait in TOPMed samples

We applied PrediXcan whole blood *mashr* predictive models to predict gene expression in 29,696 participants with WGS data from the Trans-Omics for Precision Medicine (TOPMed) program. We then used the three LMM methods to identify expression-trait associations with the white blood cell (WBC) count. We tested each of the 12,563 genes individually with the WBC using the three null models. We used a

Bonferroni-corrected p-value threshold of 3.98×10^{-6} to determine significance.

The results of the WBC transcriptome-wide association study are presented in Manhattan plots in Figure 5.3 and Table 5.2. Overall, we identified 49 genes with statistically significant associations using at least one of the three methods. Among these, analysis with the GRM model had 48 statistically significant associations, whereas analyses with the GRM+TRM and the TRM models only had 32 statistically significant associations. Out of 32 genes, all but one gene ($DCAF8$ on chromosome 1) were significant under the GRM model. Among the significant genes, 37 hits are located on chromosome 1 in the region surrounding the Duffy antigen/chemokine receptor gene, $DARC/ACKR1$, on 1q23. Gene expression prediction for $ACKR1$ was not available in the set of PrediXcan *mashr* models, therefore it was not tested. Gene $BCAN$ that reached the lowest p-value among the genes tested (7.51×10^{-40}, 2.04×10^{-33}, and 1.11×10^{-33} using the GRM, GRM+TRM and TRM models, respectively) has not been previously associated with the WBC, but further studies would be needed to replicate this association.

Out of all significant genes on chromosome 1, six genes ($OR6K3$, $UBE2Q1$, $TTC24$, $FCRL3$, $FCER1A$, and $USF1$) have been previously identified as being close to one or more of GWAS sentinel variants associated with one of the white blood cell traits. Seven genes on chromosome 6 that are significant under the GRM model fall in the major histocompatibility complex (MHC) or HLA region on 6p11-p21. However, they do not reach the significance levels under the GRM+TRM and the TRM models. Out of four hits on chromosome 17, only $IKZF3$ $MED24$ remain significant under the GRM+TRM and the TRM models, while $GSDMB$ and $ORMDL3$ are not significant anymore. While all four genes have previously been annotated as the nearest genes for one or more GWAS sentinel variants that are associated with the WBC phenotypes, TWAS does not provide much additional information in prioritizing potentially causal genes.

It is interesting to note that two loci on chromosome 7, $CREB5$ and $JAZF1$ have

been previously annotated as being the nearest gene to a GWAS sentinel variant. However, in this TWAS $JAZF1$ does not reach the significance threshold (with p-value of 3.92×10^{-5}, 2.28×10^{-4}, and 3.23×10^{-4} for the GRM, GRM+TRM, and the TRM models, respectively). Moreover, according to Human Protein Atlas [69], $CREB5$ is highly expressed in blood and brain tissue, whereas $JAZF1$ is expressed in most tissues. Together these results might suggest that $CREB5$ should be prioritized as the more plausible gene for this locus.

From both the table and the QQ plot in Figure 5.4, we can see that p-values from the GRM model are much lower than the corresponding p-values from the GRM+TRM and the TRM models. The GRM+TRM and the TRM models do not show inflation and have a genomic control inflation factor, λ_{GC}, very close to 1. At the same time, the GRM model has a much higher inflation rate of 1.212 and shows that there is more early inflation in the QQ plot. Since most of the significant genes are on chromosome 1 with the signal driven mostly by the Duffy region, we looked at the inflation rates for the three methods on chromosomes 2-22. A QQ-plot including genes on chromosomes 2-22 is presented in Figure 5.5. The genomic control inflation factors are 1.153, 1.034, and 1.043 for the GRM, GRM+TRM and TRM models, respectively. The GRM model still appears to be more inflated than the GRM+TRM and the TRM models.

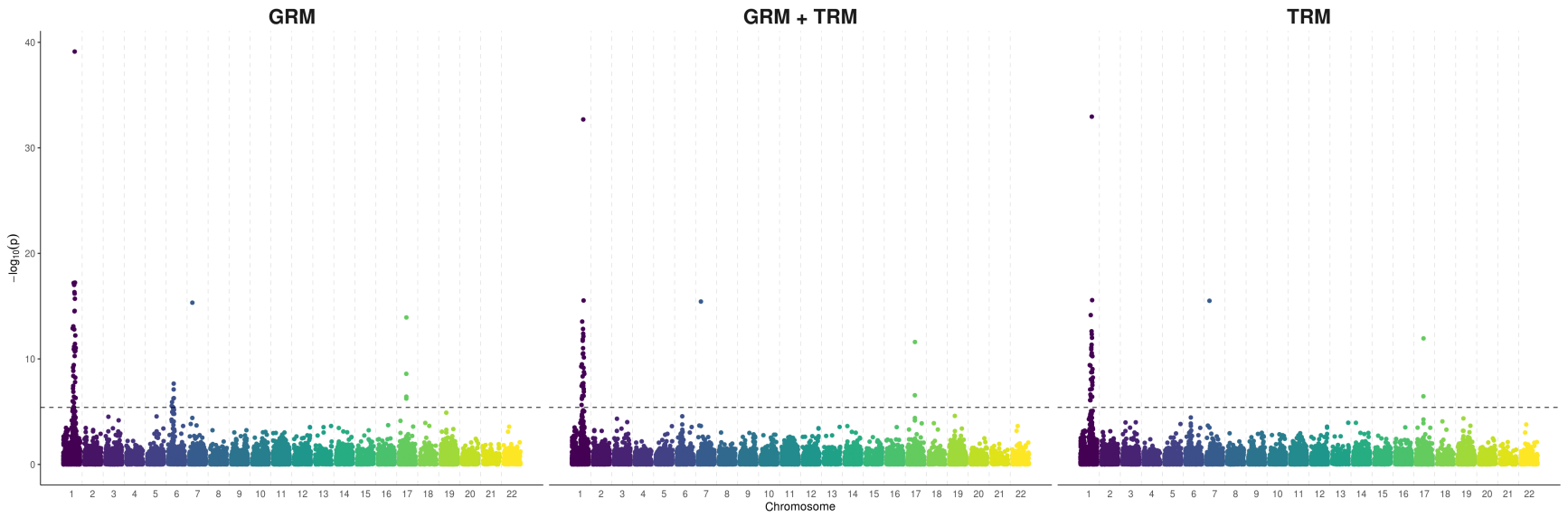


Figure 5.3: **Comparison of methods for WBC trait in 6 studies from TOPMed program.** Manhattan plots of $-log_{10}(p)$-values at 12,563 genes tested for an association with the WBC trait in TOPMEd samples using three methods. Methods left to right: model with GRM only, model with GRM and TRM, model with TRM only. The bottom dashed horizontal line represents the Bonferroni-corrected significance threshold ($P = 3.98 \times 10^{-6}$) and the top dashed horizontal line represents the breakpoint for the different scaling of the y axis. The dashed vertical lines separate the 22 chromosomes.

Table 5.2: TWAS results for white blood cell count in 6 studies from TOPMed program.

			GRM	GRM+TRM	TRM
λ_{GC}			1.212	1.078	1.092
ENSG Gene ID	Gene Name	Chr		p-value	
ENSG00000272031.2	ANKRD34A	1	1.27×10^{-13}	5.22×10^{-10}	3.89×10^{-10}
ENSG00000168509.19	HFE2	1	1.01×10^{-6}	6.22×10^{-5}	3.2×10^{-5}
ENSG00000203757.1	OR6K3	1	3.77×10^{-12}	7.02×10^{-10}	1.82×10^{-9}
ENSG00000136631.12	VPS45	1	6.27×10^{-8}	2.32×10^{-6}	8.16×10^{-7}
ENSG00000143401.14	ANP32E	1	6.5×10^{-10}	3.65×10^{-8}	7.91×10^{-8}
ENSG00000143437.20	ARNT	1	3.6×10^{-8}	2.68×10^{-5}	1.9×10^{-5}
ENSG00000143434.15	SEMA6C	1	4.11×10^{-9}	5.23×10^{-7}	8.23×10^{-7}
ENSG00000163154.5	TNFAIP8L2	1	1.31×10^{-7}	1.86×10^{-5}	1.59×10^{-5}
ENSG00000163155.11	LYSMD1	1	1.36×10^{-9}	6.08×10^{-7}	2.37×10^{-7}
ENSG00000163156.11	SCNM1	1	8.25×10^{-14}	3.27×10^{-10}	4.28×10^{-10}
ENSG00000163191.5	S100A11	1	3.58×10^{-6}	1.60×10^{-64}	2.32×10^{-4}
ENSG00000143520.6	FLG2	1	6.4×10^{-18}	2.9×10^{-14}	7.07×10^{-15}
ENSG00000188015.9	S100A3	1	1.21×10^{-11}	2.36×10^{-8}	8.64×10^{-9}
ENSG00000196154.11	S100A4	1	3.83×10^{-10}	3.26×10^{-7}	3.56×10^{-7}
ENSG00000143554.13	SLC27A3	1	3.77×10^{-7}	3.07×10^{-5}	7.28×10^{-5}
ENSG00000143570.17	SLC39A1	1	1.45×10^{-6}	9.91×10^{-6}	8.83×10^{-6}
ENSG00000143549.19	TPM3	1	3.84×10^{-6}	3.41×10^{-5}	2.66×10^{-5}
ENSG00000160714.9	UBE2Q1	1	6.98×10^{-12}	4.54×10^{-9}	1.83×10^{-9}
ENSG00000160691.18	SHC1	1	9.71×10^{-18}	1.89×10^{-12}	1.23×10^{-11}
ENSG00000160688.18	FLAD1	1	6.29×10^{-18}	1.4×10^{-12}	8.42×10^{-12}
ENSG00000169242.11	EFNA1	1	1.61×10^{-13}	3.37×10^{-11}	4.47×10^{-11}
ENSG00000132680.10	KIAA0907	1	4.73×10^{-17}	1.45×10^{-13}	2.42×10^{-13}

Table 5.2: TWAS results for white blood cell count in 6 studies from TOPMed program.

			GRM	GRM+TRM	TRM
λ_{GC}			1.212	1.078	1.092
ENSG Gene ID	Gene Name	Chr		p-value	
ENSG00000132698.14	RAB25	1	3.05×10^{-15}	9.67×10^{-12}	4.55×10^{-12}
ENSG00000116604.17	MEF2D	1	6.49×10^{-17}	4.07×10^{-13}	4.57×10^{-13}
ENSG00000187862.11	TTC24	1	2.81×10^{-15}	2.96×10^{-11}	2.88×10^{-11}
ENSG00000163382.11	NAXE	1	5.34×10^{-11}	3.59×10^{-8}	6.95×10^{-8}
ENSG00000132692.18	BCAN	1	7.51×10^{-40}	2.04×10^{-33}	1.11×10^{-33}
ENSG00000160856.20	FCRL3	1	1.95×10^{-11}	1.99×10^{-8}	1.1×10^{-8}
ENSG00000132704.15	FCRL2	1	2×10^{-16}	7.29×10^{-13}	9.96×10^{-13}
ENSG00000179639.10	FCER1A	1	5.91×10^{-18}	2.92×10^{-16}	2.71×10^{-16}
ENSG00000188004.9	C1orf204	1	1.59×10^{-8}	3.18×10^{-7}	3.88×10^{-7}
ENSG00000132716.18	DCAF8	1	9.21×10^{-6}	7.18×10^{-8}	1.67×10^{-8}
ENSG00000066294.14	CD84	1	5.86×10^{-9}	1.27×10^{-7}	3.1×10^{-8}
ENSG00000122223.12	CD244	1	6.05×10^{-13}	7.29×10^{-11}	5.55×10^{-11}
ENSG00000158773.14	USF1	1	1.74×10^{-11}	1.79×10^{-9}	9.29×10^{-10}
ENSG00000132185.16	FCRLA	1	4.98×10^{-7}	9.33×10^{-6}	8.37×10^{-6}
ENSG00000132196.13	HSD17B7	1	9.21×10^{-12}	2.63×10^{-9}	5.83×10^{-9}
ENSG00000187626.8	ZKSCAN4	6	2.87×10^{-6}	3.51×10^{-3}	6.75×10^{-3}
ENSG00000281831.1	HCP5B	6	1.23×10^{-6}	1.28×10^{-3}	1.01×10^{-3}
ENSG00000137310.11	TCF19	6	3.4×10^{-6}	1.09×10^{-3}	1.33×10^{-3}
ENSG00000213719.8	CLIC1	6	5.58×10^{-7}	2.26×10^{-3}	3.94×10^{-3}
ENSG00000204388.6	HSPA1B	6	7.8×10^{-8}	1.62×10^{-4}	1.34×10^{-4}
ENSG00000204385.10	SLC44A4	6	2.2×10^{-8}	2.74×10^{-5}	3.57×10^{-5}
ENSG00000244731.7	C4A	6	5.22×10^{-7}	5.08×10^{-3}	8.55×10^{-3}

Table 5.2: TWAS results for white blood cell count in 6 studies from TOPMed program.

			GRM	GRM+TRM	TRM
λ_{GC}			1.212	1.078	1.092
ENSG Gene ID	Gene Name	Chr		p-value	
ENSG00000146592.16	CREB5	7	4.71×10^{-16}	3.67×10^{-16}	3.12×10^{-16}
ENSG00000161405.16	IKZF3	17	2.58×10^{-9}	2.8×10^{-7}	3.49×10^{-7}
ENSG00000073605.18	GSDMB	17	5.76×10^{-7}	6.84×10^{-5}	1.24×10^{-4}
ENSG00000172057.9	ORMDL3	17	3.77×10^{-7}	3.99×10^{-5}	5.49×10^{-5}
ENSG00000008838.19	MED24	17	1.18×10^{-14}	2.5×10^{-12}	1.14×10^{-12}

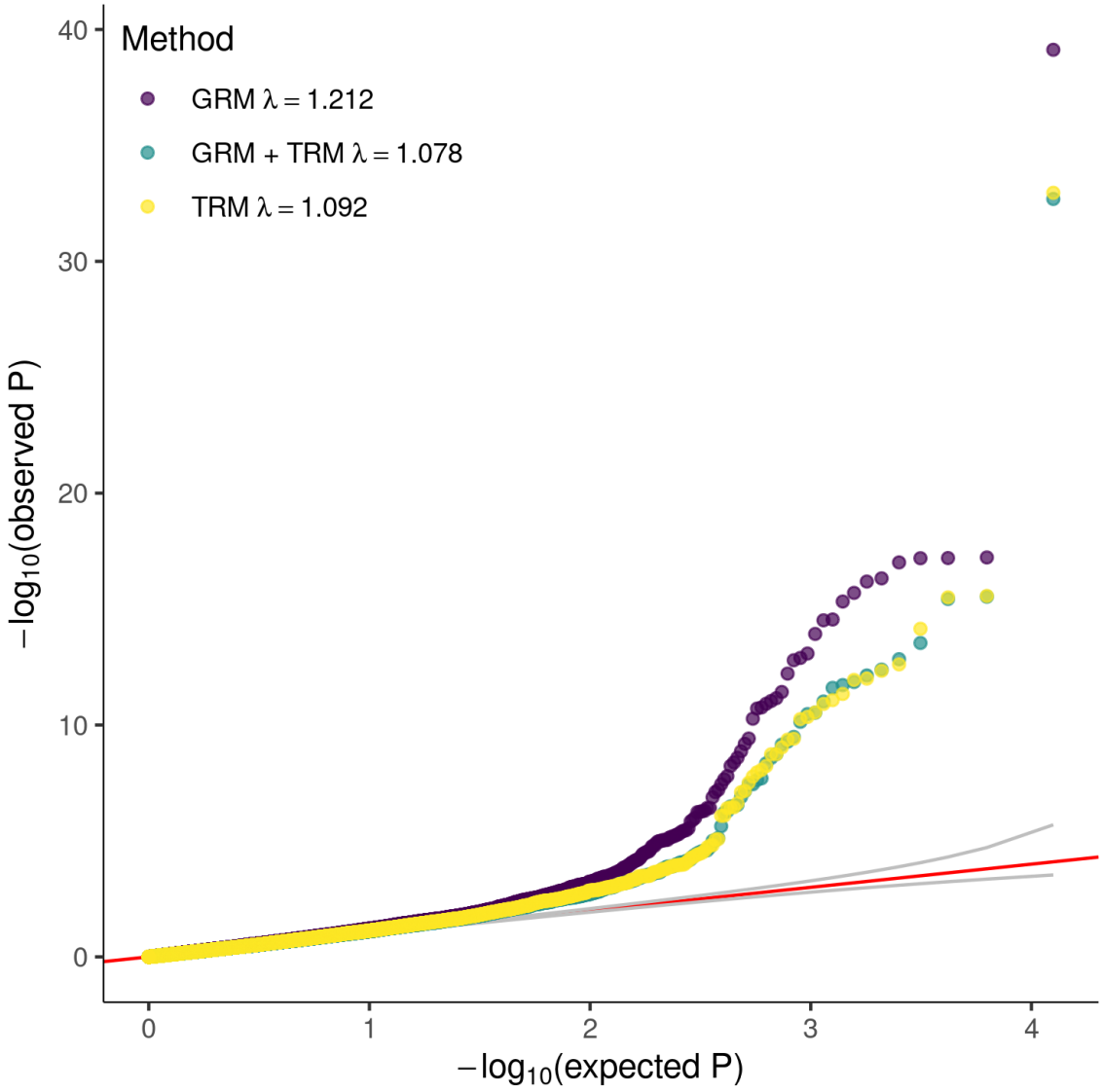


Figure 5.4: **QQ plot for the WBC trait in TOPMEd samples using three methods.** QQ-plot showing the distribution of $-log_{10}(p)$-values at 12,563 genes tested for an association from GRM, GRM+TRM, and TRM models for the white blood cell count in 6 studies from TOPMed freeze 8. The solid red line indicates the $x = y$ line.

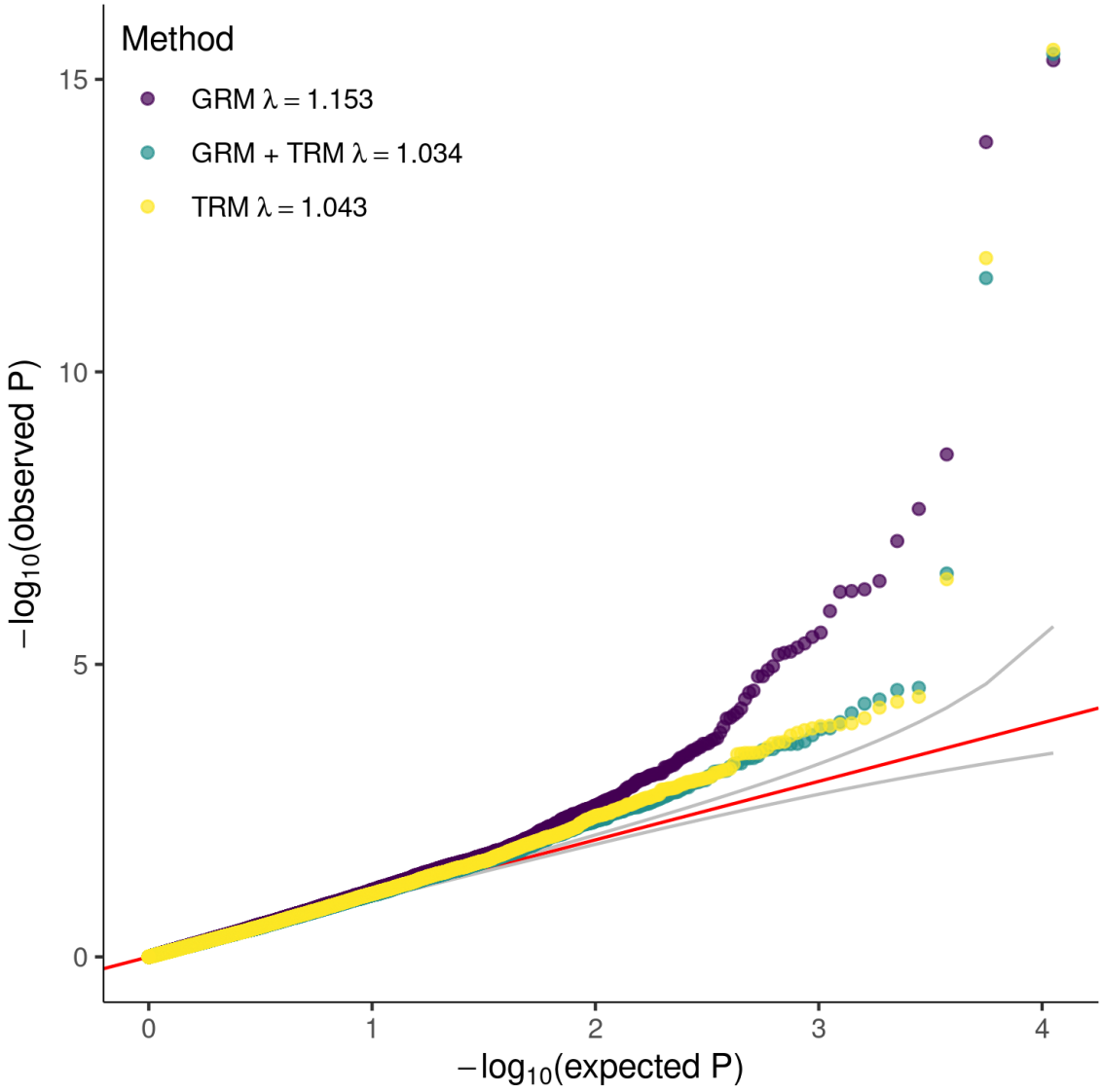


Figure 5.5: **QQ plot for the WBC trait in TOPMEd samples using three methods excluding chromosome 1.** QQ-plot showing the distribution of $-log_{10}(p)$-values at the genes tested for an association, excluding genes on chromosome 1, from GRM, GRM+TRM, and TRM models for the white blood cell count in 6 studies from TOPMed freeze 8. The solid red line indicates the $x = y$ line.

5.4 Discussion

Linear mixed models (LMMs) have become a widely used approach for conducting genome-wide association studies of quantitative and binary traits. Due to their ability to account for population structure and relatedness among samples, they are an attractive choice for large-scale studies with ancestrally diverse and admixed populations. Recently transcriptome-wide association studies (TWAS) have been gaining attention, as a way to integrate transcriptomic data into genetic association analyses and provide additional biological and functional insights over those produced by GWAS. The general TWAS framework requires first predicting gene expression levels from the genotypic data available on subjects and then testing phenotype associations to expression levels. However, TWA studies suffer from many of the same challenges as GWAS, i.e. the need to account for population structure and relatedness.

To address this issue, we have developed an approach that extends the LMM framework to be used for conducting transcriptome-wide association studies. The proposed method models both genetic and transcriptomic correlation by including two random effects in the model, a genetic relationship matrix (GRM) and a transcriptomic relationship matrix (TRM). Thus, it can accommodate population structure, admixture and relatedness, whether known or unknown. The sample structure does not need to be known *a priori*, as we estimate it empirically. Moreover, additional random effects, e.g. due to shared environmental variances, can be easily added to the model.

In simulation studies, we demonstrated that the proposed GRM+TRM method provides properly calibrated test statistics in terms of type I error, compared to the LMM model that only included the GRM. Furthermore, the GRM+TRM model achieves higher power than the GRM only model in all the simulation settings we tested. The GRM+TRM method has a better true positive to false positive ratio, whereas the GRM only method identifies too many false positives. Finally, the GRM+TRM method provides accurate estimates of variance components, whereas in settings when

we do not include one of the random effects, the estimates of variance components may be highly inflated.

We performed a large-scale TWAS on 29,696 ancestrally diverse individuals from the TOPMed program using the proposed GRM+TRM method and an LMM that includes the GRM only. We replicated the known signal – six genes in the Duffy region on chromosome 1, $CREB5$ gene on chromosome 7 and genes on chromosome 17. We found that the GRM+TRM model had the genomic control inflation factor a lot closer to 1, whereas the GRM model yielded more inflated results, with lower p-values, and, hence, a lot more statistically significant genes.

We demonstrated advantages of TWAS over GWAS in helping prioritize potentially causal genes on the $CREB5$ locus. Previously, two genes $CREB5$ and $JAZF1$ have been assigned to this locus, based on their distance to GWAS sentinel variants. However, $CREB5$ shows a stronger TWAS signal and is more enriched in blood tissue, especially neutrophils, based on Human Protein Atlas, than $JAZF1$. Together these results show evidence that $CREB5$ is more likely to be a biologically plausible candidate gene for causality.

We also want to highlight the challenges of TWAS when it comes to interpreting results where a cluster of genes is statistically significant. While some loci on chromosome 1 were not previously associated with the WBC phenotypes, it is impossible to conclude whether multiple genes contribute to trait regulation or some of the signal is spurious and is driven by shared eQTLs across genes or tissues. Importantly, one should not conclude that the most significant gene is more likely to be causal. For example, $ACKR1$ gene that is well studied and known to be associated with the WBC trait is not imputed with the PrediXcan *mashr* models, which were trained on subjects of European ancestry from GTEx. So, further replication studies would be needed to determine whether these associations, including the gene with the lowest p-value $BCAN$, are novel discoveries or spurious.

Finally, we want to address the computational burden of the proposed method.

Fitting mixed models in large samples requires significant memory and CPU time. In our analyses, we fit one 'null model' under the null hypothesis of no association and, subsequently, used score tests to assess each gene's association with the trait. However, even fitting one null model can be computationally demanding, when sample sizes are very large. To account for genetic relatedness, it has been proposed to use a sparse, block-diagonal empirical GRM, since pedigree-based relatedness captured by the GRM is sparse by nature [22]. An empirical TRM, estimated from transcriptomic levels, is dense, with no entries equal to zero. Thus, fitting a mixed model with a dense TRM as a random effect can be a computational hurdle. Our approach will need to be further optimized for computational speed and efficiency, as the datasets analyzed grow larger and larger in size. The computational burden limited the TOPMed analysis with GRM and TRM to $\sim 30K$ samples. Nonetheless, we perform a WBC TWAS on all TOPMed freeze 8 samples and present the results in the next chapter.

Chapter 6

MIXED-MODEL ASSOCIATION TESTING IN TRANSCRIPTOME-WIDE ASSOCIATION STUDIES WITH VARIANCE STRUCTURE

6.1 Introduction

Recent transcriptome-wide association studies (TWAS) include participants from multi-ethnic and multi-study programs, such as the Trans-Omics for Precision Medicine (TOPMed) program. Large-scale studies often include related individuals and this relatedness may be known or cryptic. They may also include individuals of diverse ethnic backgrounds with complex ancestry. It is often of interest for participant data from multiple studies or centers to be combined and analyzed together in a single analysis. When conducting association studies with transcriptomic data, it is essential to take into account all possible sources of confounding, such as participants' relatedness and the presence of population structure, for validity of the association results.

Statistical methods commonly used to correct for population structure, such as principal components and linear mixed models, primarily focus on controlling for differences in means across groups due to confounders. However, heterogeneity in phenotypic variance between different groups, in addition to phenotypic mean, has been observed for a variety of traits [15, 43, 62]. Phenotypic variances can differ between participants from different studies or different ethnic groups, and failure to account for heterogeneous variances between groups can result in false-positive associations and loss of power to detect associations.

Previous studies have identified some causes of heterogeneity in phenotypic vari-

ances. For example, gene-gene interactions, or epistasis, can increase the variance of a quantitative trait [58]. Gene-environment interactions can also affect phenotypic variance, when different population groups are exposed to different environmental factors [57, 42]. In addition, phenotypic variance can be affected by variance-controlling loci (vQTLs) [57] and natural selection has been shown to induce phenotype-variance structure [43].

Recent genome-wide association studies (GWAS) considered the implications of pooling groups with different phenotypic variances and analyzing them together. Conomos et al. performed a GWAS of a pooled Hispanic/Latino sample and showed that many biomedical traits have heterogeneous variances among Hispanic subgroups and that modeling this heteroscedasticity reduced genomic inflation [15]. Musharoff et al. illustrated that population variance structure is ubiquitous and developed a statistical framework to test for the association of genetic variants and phenotypic variability [43]. Sofer et al. considered the consequences of unaccounted for variance stratification and developed a procedure to address the variance stratification problem [62].

Here, we propose an approach for modeling heteroscedasticity among groups using a linear mixed model framework for TWAS. Our method is flexible and can accommodates heterogeneous variance groups defined by self-reported or inferred race or ethnicity, study or a combination of both. To properly account for known or cryptic relatedness, we use an empirical genetic relatedness matrix (GRM) and/or a transcriptomic relationship matrix (TRM). Moreover, our approach allows for inclusion of additional random effects in the model, e.g. environmental correlation.

To assess the utility of our approach, we first perform a series of simulations. We demonstrate that our method is properly calibrated in terms of type I error and results in higher power when testing gene transcriptome for associations. Finally, we apply our method to a large cohort of ancestrally diverse individuals from multiple studies that were included in the TOPMed program and perform a GWAS and a TWAS of the

white blood cell (WBC) count. We identify 31 genes that are statistically significant in the TWAS after Bonferroni correction.

6.2 Methods

We revisit the TWAS framework we introduced in Chapter 5 and extend it to incorporate samples with heterogeneity in phenotypic variance across subgroups.

6.2.1 Linear mixed model framework for TWAS

Consider a set of sampled individuals $\mathcal{N}$ with predicted transcriptome values for $\mathcal{M}$ genes and let $\mathbf{y}$ be a vector of values for some quantitative phenotype. We can fit the following mixed effects model for testing association of $\mathbf{y}$ with a gene $m \in \mathcal{M}$:

$$\mathbf{y} = \beta\mathbf{t} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\epsilon}, \tag{6.1}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_G^2\boldsymbol{\Phi}),$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}),$$

where $\mathbf{t}$ is a vector of predicted transcriptome values for gene m for all individuals with the corresponding effect size β; $\mathbf{W}$ is a matrix of covariates including an intercept with a corresponding vector of effect sizes $\boldsymbol{\alpha}$; $\mathbf{V}$ is the matrix of ancestry representative vectors with corresponding effect sizes $\boldsymbol{\gamma}$ to adjust for population structure; $\mathbf{u}$ is a random effect that captures the polygenic effect of genotypes, where $\boldsymbol{\Phi}$ is the genetic relationship matrix (GRM) and σ_G^2 corresponds to the total additive genetic variance of the trait; and $\boldsymbol{\epsilon}$ is a random variable that captures independent residual effects, where σ_ϵ^2 is a parameter that measures residual variance and $\mathbf{I}$ is an identity matrix. This model can include an additional random effect that accounts for transcriptomic data $\mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\boldsymbol{\Psi})$, $\boldsymbol{\Psi}$ is the transcriptomic relationship matrix (TRM) and the total additive variance of the trait across predicted transcriptome corresponds to σ_T^2. This model was described in more detail in Chapter 5.

We refer to the model in Equation 6.1 as the 'homogeneous variance model.'

6.2.2 Modeling heterogeneous variances across subgroups

In the previous subsection, trait variance homogeneity is assumed across subgroups. Now, we will allow for heterogeneity in phenotypic variance across the subgroups.

Consider $\mathcal{L}$ groups of individuals in the sample. We propose using an LMM that allows for heterogeneous variances among groups:

$$\mathbf{y_i} = \beta \mathbf{t_i} + \mathbf{W_i}\boldsymbol{\alpha} + \mathbf{V_i}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\epsilon}_i, \tag{6.2}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_G^2 \boldsymbol{\Phi}),$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma_i^2 \mathbf{I}),$$

where σ_i^2 corresponds to i-th group variance, $i \in \mathcal{L}$. Similarly to the model 6.1, this model can include a TRM as an additional random effect.

We refer to the model in Equation 6.2 as the 'heterogeneous variance model.'

6.2.3 Variance of y under the null hypothesis

Generalized least squares (GLS) can be used to fit the LMM in Equation 6.2 and test H_0: $\beta = 0$ vs H_A: $\beta \neq 0$. In samples with structure, we need to take into the account the correlation among all individuals. The covariance matrix of the phenotype $\mathbf{y}$ under the null hypothesis of no association between the phenotype and the gene is $\boldsymbol{\Sigma} = \sigma_G^2 \boldsymbol{\Phi} + \mathbf{D}$, where $\mathbf{D}$ is diagonal matrix with group-specific variances σ_i^2s on the diagonal. However, $\boldsymbol{\Sigma}$ is typically unknown and must be estimated.

6.2.4 Estimation of variance components

Using the empirical GRM defined above, $\hat{\boldsymbol{\Phi}}$, in the null model, we can estimate the variance components σ_G^2, and σ_i^2s with restricted maximum likelihood (REML). Afterwards, we can perform GLS using the estimate of the overall phenotypic covariance structure, $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_G^2 \hat{\boldsymbol{\Phi}} + \hat{\mathbf{D}}$.

6.2.5 Simulation study

Throughout our simulations, we used 10,000 participants with real whole-genome sequencing data from the TOPMed program. We included 5,000 participants from HCHS/SOL study (study 1) and 5,000 participants from WHI study (study 2). The sample included unrelated and related individuals of diverse ancestral backgrounds. We used imputed whole blood gene expression values, predicted with PrediXcan *mashr* models [21, 5, 4], and simulated causal effects and simulated phenotypes from the predicted gene expression values.

First, we generated an initial phenotype vector, $\mathbf{y_0}$, from an empirical GRM and a random vector of $\mathcal{N}(0, \sigma_i^2)$ values:

$$\text{Var}(\mathbf{y}_{0_i}) = \mathbf{\Sigma}_{0_i} = \sigma_G^2 \mathbf{\Phi} + \sigma_i^2 \mathbf{I}, \quad i = 1, 2$$

where group i had group-specific variance σ_i^2 and σ_G^2 was a common genetic variance parameter.

We then simulated a phenotype $\mathbf{y}_i$ that had a group-specific intercept for group i and was correlated with predicted gene expression values of a randomly selected gene according to the model:

$$\mathbf{y_i} = \mathbf{y_{0_i}} + \sum_i \alpha_i \mathbb{I}_{group_i} + \beta \mathbf{t_i}, \quad i = 1, 2 \tag{6.3}$$

where α_is were group-specific intercepts for group i and $\mathbf{t}$ was a vector of predicted transcriptome values for a randomly selected gene with the corresponding effect β.

We considered group-specific variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$, group-specific intercepts $\alpha_1 = 1$ and $\alpha_1 = 2$, while σ_G^2 was set to 1 in all simulation settings. Effect size, β, was chosen such that a randomly chosen causal gene explained 0.05% of the variability of the trait.

With the simulated phenotypes, we performed association tests with both the causal gene and null genes by fitting a linear mixed model (1) with homogeneous residual variances and (2) allowing for heterogeneous residual variances among groups.

Each model was adjusted for the study variable and included a 4-th degree sparse empirical kinship matrix computed with PC-Relate [15]. All analyses were conducted in the GENESIS `R` package [22].

6.2.6 TWAS and GWAS of WBC trait in TOPMed study

The analyses reported here included 61,802 participants from 13 TOPMed studies: Genetics of Cardiometabolic Health in the Amish (Amish, $n = 1,102$), Atherosclerosis Risk in Communities Study VTE cohort (ARIC, $n = 7,899$), Mount Sinai BioMe Biobank (BioMe, $n = 10,963$), Coronary Artery Risk Development in Young Adults (CARDIA, $n = 3,042$), Cardiovascular Health Study (CHS, $n = 3,488$), Genetic Epidemiology of COPD Study (COPDGene, $n = 5,792$), Framingham Heart Study (FHS, $n = 3,136$), Genetic Studies of Atherosclerosis Risk (GeneSTAR, $n = 1,711$), Hispanic Community Health Study/Study of Latinos (HCHS/SOL, $n = 7,225$), Jackson Heart Study (JHS, $n = 2,841$), Multi-Ethnic Study of Atherosclerosis (MESA, $n = 2,517$), Whole Genome Sequencing to Identify Causal Genetic Variants Influencing CVD Risk - San Antonio Family Studies (SAFS, $n = 1,151$) and Women's Health Initiative (WHI, $n = 10,935$). The composition of the 61,802 participants by race/ethnicity is 54% White, 23% Black, 22% Hispanic/Latino, and 1% Asian. All studies were approved by the appropriate institutional review boards (IRBs), and informed consent was obtained from all participants. White blood cell (WBC) count is the count of white blood cells in the blood, by number concentration in thousands per microliter. During QC, we excluded participants with WBC values $> 100\mu l$ ($n = 5$). Additionally, in cases where multiple measurements were available, we kept only one measurement for each individual.

First, we performed genome-wide single variant association tests using a two-step procedure. In the first step, we fit the null model under the null hypothesis of no genetic association, with no genetic variants in the model. We included sex, age, combined study by phase variable (e.g. WHI_2 refers to phase 2 of WHI study),

Duffy variant, and the first 11 PC-Air [16]) principal components (PCs) of genetic ancestry as fixed effects. To account for genetic relatedness, we included a 4th-degree sparse empirical kinship matrix (KM) computed with PC-Relate [15]). Additional details on the computation of the ancestry PCs and the sparse KM are provided in the Appendix C. We allowed for heteroscedasticity in the error variances by modeling separate residual variance components, one for each study by ancestry group (e.g., WHI_White). Details on estimating the ancestry group are in Appendix C.

In order to improve power and appropriately control Type I error in settings with a non-Normal phenotype distribution, we used a fully-adjusted two-stage approach for fitting the null model [61]. In stage 1, we fit an LMM with the observed phenotype values as the outcome, the fixed effects covariates, sparse KM and heterogeneous residual variances. We applied a rank-based inverse-normal transformation to the residuals from the results of stage 1 and then rescaled them by the original variance. In stage 2, we fit another LMM using the rescaled residuals obtained in stage 1 as the outcome and the same covariates, same KM and heterogeneous residual variance model as in stage 1. Finally, we used the output from stage 2 as the trait of interest to perform a score test of genetic association. In the GWAS analyses, we included variants with minor allele count (MAC) of at least 5 that passed the TOPMed IRC quality filters and had less than 10% of samples with sequencing read depth of less than 10. A threshold level of 5×10^{-8} was used to determine statistical significance.

In addition to single-variant association tests, we performed transcriptome-wide association tests. We used PrediXcan whole blood *mashr* models to impute transcriptome values for each individual in the dataset [4, 21, 5]. In our analyses, we used two null models: (1) a null model allowing for heterogeneous error variances and (2) null model with homoscedastic error variance. The null heterogeneous variances model for TWAS was identical to the model we fit for GWAS analyses. The homogeneous variance model did not model separate residual variance components, but, otherwise, was identical to the first null model. Statistical significance was determined using a

Bonferroni correction for the number of genes tested.

6.3 Results

6.3.1 Evaluation of type I error and power

We performed a series simulation studies in order to demonstrate that the heterogeneous variances model for TWAS is properly calibrated and outperforms the homogeneous residual variances model in the presence of heterogeneity. We randomly selected two sets of 5,000 samples from two different studies from the TOPMed program and simulated 1,000 phenotypes, such that the two groups had different phenotypic variances of $\sigma^2 = 1$ and $\sigma^2 = 2$. To assess type I error rate and power of the two methods, we performed association tests with simulated phenotypes. We calculated proportions of false and true positives, based on the association tests carried out with causal and non-causal genes. We compared the performance of the heterogeneous residual variances model to the performance of the homogeneous variances model.

To assess the type I error rate, we tested association of 10,000 null genes with each of the 1,000 simulated phenotypes using a homogeneous variances model and heterogeneous variances model. We then calculated the proportion of false positives for various significance levels (Figure 6.1). From our simulation studies, we observed very similar false positive rates for the two models. Both models produce reasonable type I error rates at the lower range of the nominal level α. However, there is a slight inflation in type I error rate for α closer to 0.05 when using the homogeneous variance model, whereas the heterogeneous variance model controls type I error rate very well. As seen in Figure 6.1, both models have very similar type I error rates.

To evaluate the power of the two methods, we simulated phenotypes where one gene has a non-zero effect size. We then calculated the proportion of false positives and true positives for various α significance levels. Figure 6.2 shows the false positive rate compared to the true positive rate for significance levels ranging from 5×10^{-6} to

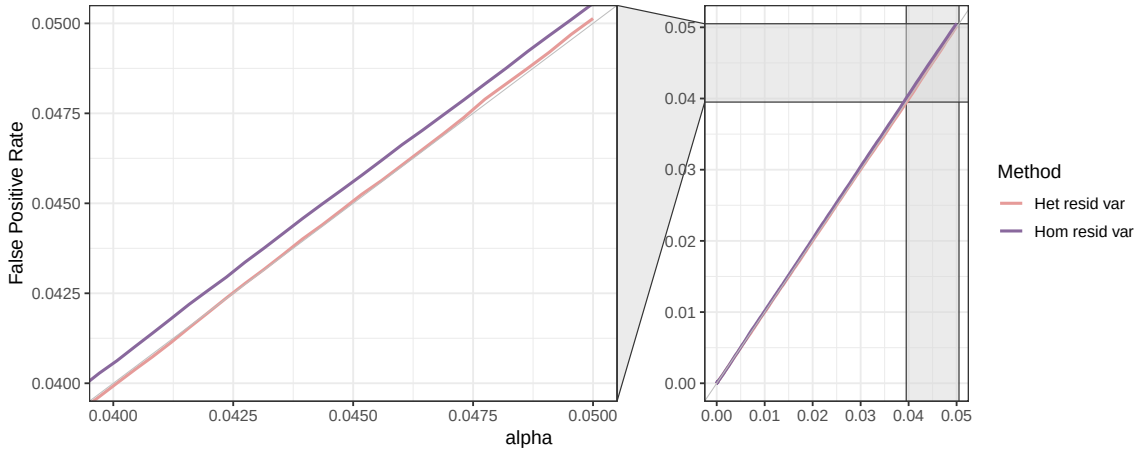


Figure 6.1: **Type I error rate for heterogeneous vs homogeneous variances models.** The proportion of true positive associations identified (true positive rate) vs significance level α for homogeneous and heterogeneous variance models. Gray line $y = x$ indicates the nominal type I error rate. Deviations from the gray line mean inflated (above) or deflated (below) type I error rates.

0.05 calculated from 1,000 simulation iterations. Both methods have similar power for the lower range of significance levels α. However, at higher significance levels, heterogeneous variances model achieves higher power. This means that not accounting for heterogeneity in phenotypic variance can lead to loss of power to detect gene-trait associations.

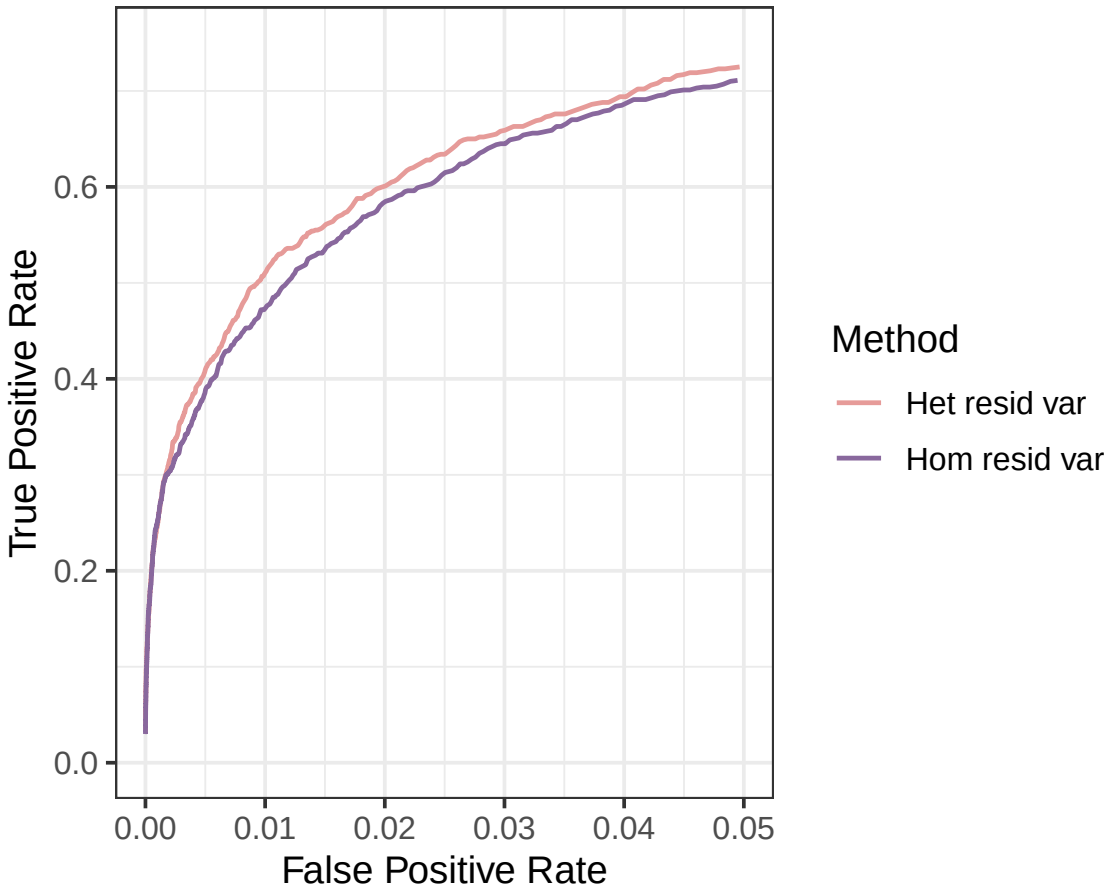


Figure 6.2: **Comparison of true and false positive rates.** The proportion of true positive associations identified (true positive rate) is compared to the proportion of null genes incorrectly identified as associations (false positive rate) by homogeneous and heterogeneous variance models in 1,000 simulations. A curve with a higher area under it indicates better performance.

6.3.2 TWAS and GWAS of WBC trait in TOPMed study

In 61,802 multi-ethnic individuals (33,285 European ancestry, 14,246 African ancestry, 13,585 Hispanic/Latino, and 686 Asian ancestry), we performed genome-wide association tests for each single-nucleotide variant (SNV) with a minor allele count (MAC)

> 5, for $\sim$105,091,710 association tests. We observed 2,846 statistically significant SNVs ($p < 5 \times 10^{-8}$).

We used PrediXcan *mashr* whole-blood prediction models from PredictDB to predict gene expression levels in the participants. Among the 12,623 genes in the PrediXcan *mashr* panel, 12,625 genes were predicted in the sample. We tested each of the 12,565 genes individually for an association with the WBC trait using the two heterogeneous and homogeneous variance null models. The significance threshold was calculated as the nominal significance level Bonferroni-adjusted for the number of genes tested for association ($0.05/12{,}625 = 4.0 \times 10^{-6}$).

Overall, we identified 31 statistically significant genes when using the heterogeneous variances null model. In comparison, 26 genes were identified as significant using the homogeneous variances model. Table 6.1 displays the genes that reach Bonferroni-adjusted significance in TWAS with either homogeneous variance or heterogeneous variances models. It can be seen from the p-values for the two models in Table 6.1, that heterogeneous variances model provides the stronger evidence of significance, possibly due to a slight increase in power to detect associations.

Figure 6.1 displays the Manhattan plot for the TWAS analysis with the heterogeneous variances model. Out of 31 statistically significant genes, 11 genes (bolded in Table 6.1) were previously identified in the NHGRI-EBI GWAS catalog [10]. The Manhattan mirror plot of TWAS and GWAS in Figure 6.4 shows signals on chromosomes 2, 5, 6, 7, 11, 12, and 17 that are present in both GWAS and TWAS.

We compared the two linear mixed models we fit for TWAS. In the homogeneous variance model we fit one residual variance component for all samples, and in the heterogeneous variances model we fit a separate residual variance component for each study by ancestry group. Figure 6.5 compares estimated residual variance components from the heterogeneous variances null model. Study/ancestry variance components estimates varied from 0.98 (95% CI: 0.81, 1.15) for the Amish group to 6.88 (95% CI: 6.40, 7.36) for the BioMe Puerto-Rican group. In comparison, the estimate for the

residual variance component from the homoscedastic null model was 2.14 (95% CI: 1.96, 2.33).

Table 6.1: TWAS results for white blood cell count in the TOPMed program.

Gene	Chr	Start pos	End pos	Het resid var p-value	Hom resid var p-value
NRBP1	2	27428187	27428508	8.76e-09*	4.00e-08*
KRTCAP3	2	27442494	27442494	3.31e-07*	1.88e-06*
ITGA4	2	181448739	181459459	7.35e-07*	2.19e-06*
CERKL	2	181459459	181459459	7.76e-07*	2.39e-06*
SLC22A5	5	132369605	132369896	2.73e-06*	4.99e-06
ZKSCAN4	6	28248649	28248649	1.88e-09*	1.63e-08*
HLA-G	6	29826540	30762132	1.61e-07*	5.77e-07*
HCP5B	6	29874768	29962980	2.81e-09*	3.79e-09*
TRIM10	6	30161172	30161172	2.46e-10*	1.68e-09*
TCF19	6	31158216	31168463	8.21e-08*	1.40e-07*
APOM	6	31652706	31652743	5.41e-10*	9.98e-10*
GPANK1	6	31665031	31668965	5.87e-07*	6.07e-06
DDAH2	6	31730180	31731796	1.68e-06*	2.68e-05
CLIC1	6	31740686	31740686	1.29e-11*	5.89e-11*
HSPA1B	6	31869050	31869050	9.46e-12*	2.26e-10*
NEU1	6	31863382	31864261	5.08e-10*	1.38e-09*
SLC44A4	6	31879419	31879419	2.17e-11*	1.60e-10*
C2	6	31902549	31902549	1.76e-11*	6.72e-11*
C4A	6	31973120	31979683	1.64e-12*	1.49e-11*
C4B	6	32010969	32061428	1.98e-10*	6.57e-10*
NOTCH4	6	31263056	32672132	1.48e-08*	1.43e-07*

Table 6.1: TWAS results for white blood cell count in the TOPMed program.

Gene	Chr	Start pos	End pos	Het resid var p-value	Hom resid var p-value
HLA-DRB1	6	32582949	32589837	1.69e-10*	2.40e-09*
HLA-DQA2	6	32555506	32707290	1.30e-06*	1.17e-05
CREB5	7	28684757	28684757	5.29e-24*	1.58e-21*
IFITM3	11	307726	327873	6.94e-07*	8.28e-06
ALDH2	12	111766623	111774029	4.58e-06	2.70e-06*
RABEP1	17	5282070	5381475	3.88e-07*	1.81e-06*
IKZF3	17	39863805	39863805	1.03e-16*	1.20e-15*
GSDMB	17	39895095	39926578	3.00e-12*	1.67e-11*
ORMDL3	17	39926554	39926578	4.99e-13*	2.38e-12*
MED24	17	40022374	40055793	2.59e-33*	2.56e-31*
RPS6KB1	17	59895240	59895240	2.45e-06*	2.14e-05

$*$ statistically significant after Bonferroni correction;

$**$ previously associated genes are in bold

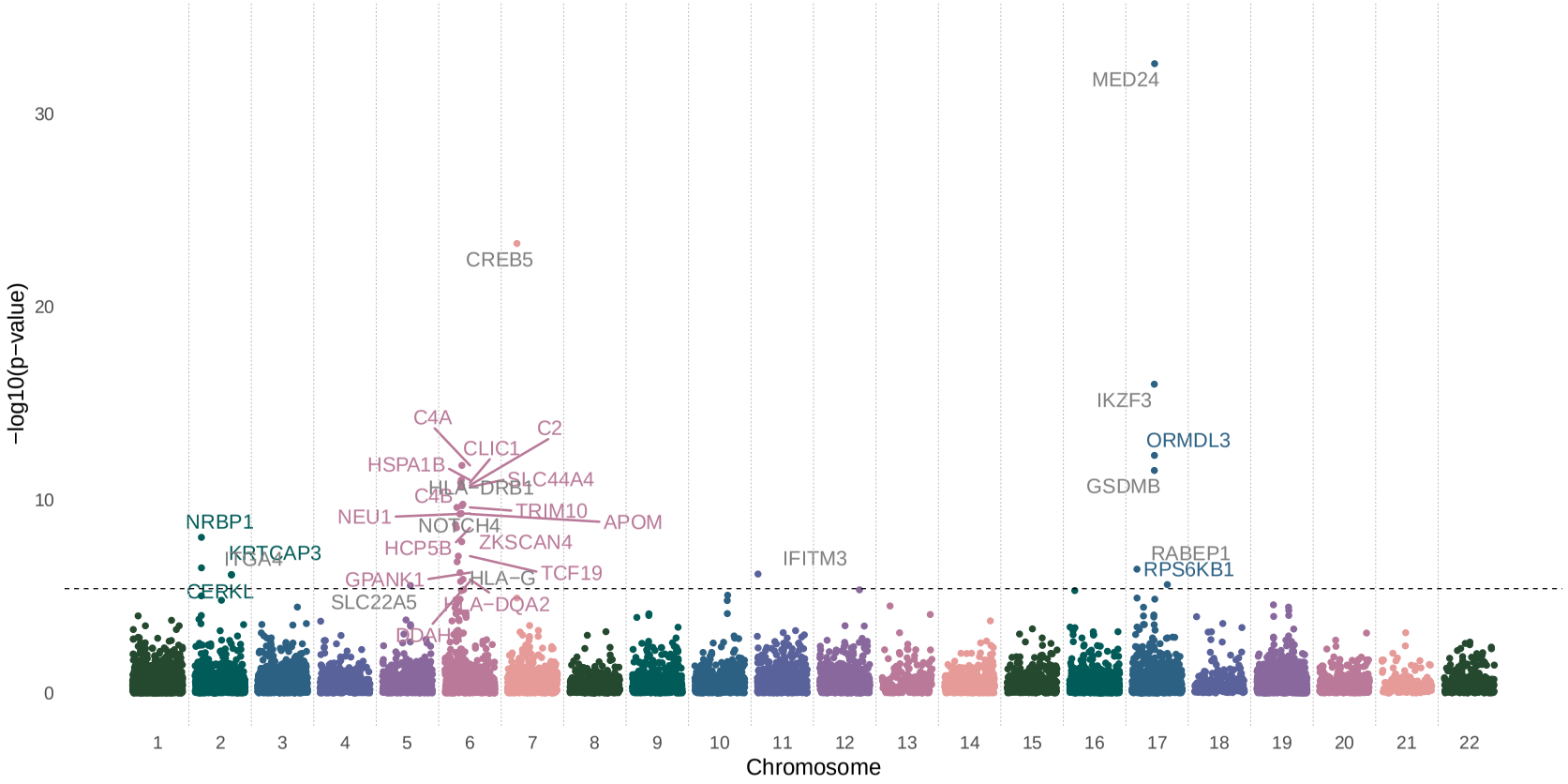


Figure 6.3: **Manhattan plot of TWAS (heterogeneous variances null model) for white blood cell count in the TOPMed program** . Manhattan plot of $-log_{10}(p)$-values of 15,656 genes tested in the TWAS of white blood cell count in the participants of the TOPMed program. The dashed horizontal line represents the Bonferroni-corrected significance threshold Previously associated genes labels are shown in gray and novel genes labels are shown in color.

6.4 Discussion

Linear mixed models have become a popular approach for conducting GWAS. Widely used LMMs make an assumption of homogeneous phenotypic variance in the analyzed sample. However, modern GWAS and TWAS often use samples pooled from multiple studies and ancestries and it would not be uncommon for one subgroup to have phenotypic variance higher than another subgroup in the sample. If the assumption of variance homogeneity does not hold, it can lead to miscalibrated tests and decreased statistical power.

We have described an approach to account for heterogeneity in phenotypic variance across groups in TWAS. In the LMM setting, we allow for the residual variance to differ across groups while keeping the genetic variance component the same. Analysis groups can be defined by race or ethnicity (self-reported or inferred), study or a combination of both. Via a simulation study, we illustrated our method is properly calibrated in terms of type I error and that accounting for heterogeneous variances across groups can lead to an increase in statistical power.

We have demonstrated the utility of our method in GWAS and TWAS settings. We applied our method to the multi-study multi-ethnic TOPMed program and performed a TWAS and a GWAS of the white blood cell count trait. We allowed for heteroscedasticity in the LMM by fitting different residual variances for each study by ancestry group. Out of 12,656 genes tested, 31 genes reached Bonferroni-adjusted significance threshold. However, further exploration needs to be done to support these associations as novel discoveries. Since previously identified GWAS signals can drive observed TWAS signals, we recommend performing conditional analyses by adjusting for previously identified variants in the vicinity (e.g., within $\pm$ 1Mb) of the gene transcript. Furthermore, TWA studies still require fine mapping to determine which statistically significant genes in a genomic region are more likely to be causally related to the phenotype. In case of TWAS, we need to account for how the genes in the

region are correlated and to account for the correlation among predictive variants in the region. Using software, such as FOCUS [36], we can perform fine-mapping and determine a credible set of genes in the region.

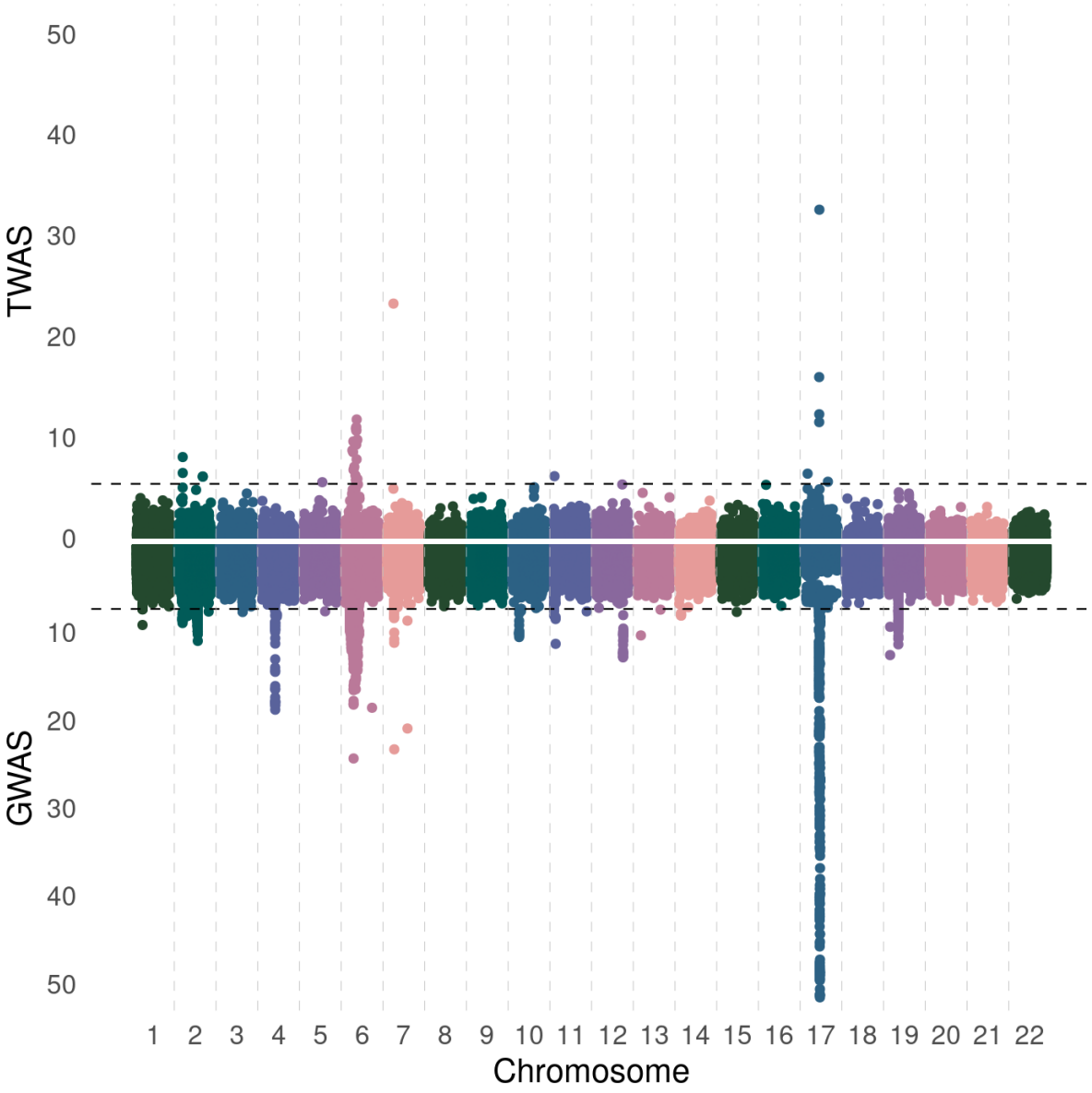


Figure 6.4: **TWAS-GWAS Manhattan mirror plot for white blood cell count in the TOPMed program.** Top: Manhattan plot of $-log_{10}(p)$-values of 15,656 genes tested in the TWAS of white blood cell count in the participants of the TOPMed program. The dashed horizontal line represents the Bonferroni-corrected significance threshold. Bottom: Manhattan plot of $-log_{10}(p)$-values of all the variants tested in the GWAS of white blood cell count in the participants of the TOPMed program. The dashed horizontal line represents the significance threshold of 5×10^{-8}.

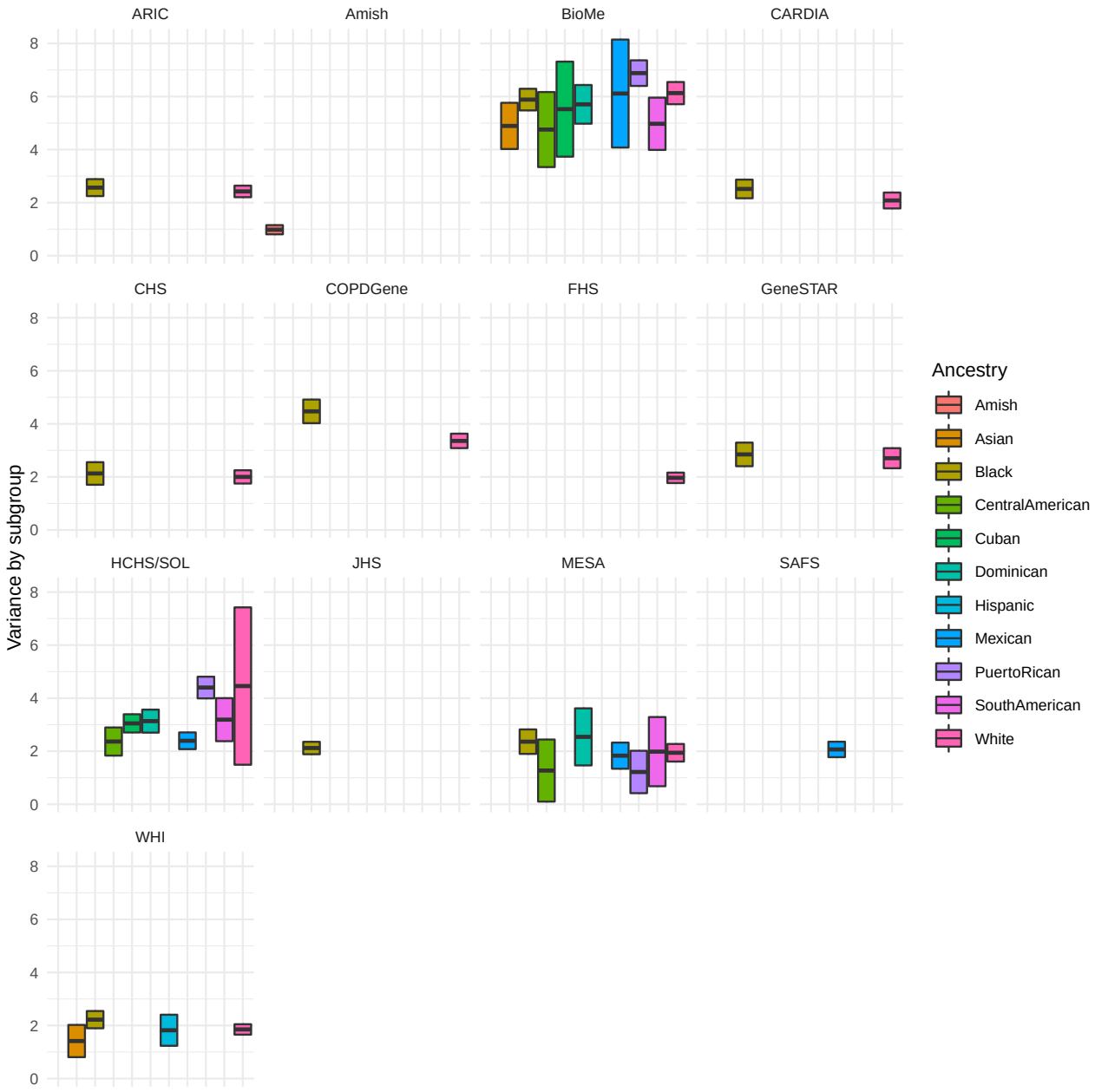


Figure 6.5: **Residual variance components in the null model for the TWAS of WBC count in the TOPMed program.** Residual variance components estimated from the heterogeneous residual variances mixed model. The colored boxes show the estimated residual variance components by subgroup. The range of each box shows the 95% confidence interval.

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, we focus on statistical methods for large-scale transcriptome-wide association studies (TWAS) for ancestrally diverse populations. TWAS is a gene-based association approach that investigates associations between genetically regulated gene expression and complex traits. TWAS gained popularity over the years due to their ability to provide some functional interpretation of genome-wide association studies (GWAS) results and understanding of biological disease mechanisms. We have evaluated the performance of an existing TWAS method in diverse populations and discussed the lack of cross-population generalizability of transcriptome prediction models. We have identified challenges and limitations of existing approaches and developed methodology to address them. In this chapter, we summarize what statistical methods we developed and discuss possible directions of future research to extend our methods.

In Chapter 3, we evaluated prediction accuracy of an existing TWAS method, PrediXcan, in diverse populations. We demonstrated that PrediXcan models, which were largely trained on European ancestry reference panels, have poor performance when applied to samples of African ancestry and have much better performance when applied to samples of European ancestry. Recent publications confirmed our findings and demonstrated that TWAS methods have the best prediction accuracy when training and testing sets are matched in ancestry[29, 41]. The lack of genetic diversity present in the reference datasets used for predictive model training poses a serious challenge to deriving prediction models appropriate for diverse populations. We emphasize the need for larger panels for diverse populations - African, Asian,

Hispanic/Latino, Middle Eastern, Native American and Pacific Islanders.

In Chapter 4, we discussed different ways of building transcriptome prediction models from reference panels. The training of prediction models can be performed in pooled multi-ethnic dataset or, alternatively, in ancestry-specific datasets. Each approach has advantages and disadvantages, and depends on the genetic architectures of a particular gene. We have more power to identify expression quantitative loci (eQTLs) shared among populations when using pooled datasets. In contrast, ancestry-specific eQTLs with weaker effect sizes may be missed in pooled analyses, but detected in ancestry-specific analyses. To addresses the lack of transcriptome prediction methods that focus on leveraging multi-ethnic reference datasets, we adopted a novel, powerful, and flexible fused lasso framework to perform gene-level association analysis. Our method allows for ancestry-specific effects while borrowing information across different ancestral groups for more efficient eQTL detection. Thus, we have increased power to detect eQTLs that are common to all the ancestral groups as well as detect eQTLs that have strong associations in one of the ancestries.

There is a number of possible directions for future research here. As more ancestrally diverse reference datasets become available, our method can be easily extended to incorporate three or more population groups. An additional challenge would be to account for two-way and three-way admixture in populations such as African Americans and Hispanic/Latinos. Our method can incorporate estimates of local ancestry on a gene-by-gene basis. We can build transcriptome prediction models by estimating ancestry-specific effect sizes at each variant by jointly modeling a set of variants together with their local ancestry estimates.

Another possible extension of our method is incorporating transcriptome data from multiple tissues. Previous research has shown that a proportion of eQTLs are shared across all tissues and have correlated effect sizes across tissues [18]. Instead of using population labels, we can use appropriate tissue labels and follow identical fused lasso framework to identify cross-tissue and tissue-specific eQTLs.

So far, TWAS methods have been using *cis*-eQTLs within a certain distance from genes. However, gene expression can be regulated by both *cis* and *trans*- regulatory elements. Previous studies estimated that up to 70% of the genetic heritability of transcriptome levels can be attributed to *trans*-eQTLs and highlighted the importance of distant regulatory elements in transcriptome regulation [33]. However, the identification of *trans*-eQTLs is hindered by an enormous multiple testing burden, since the number of statistical tests of all possible intra and inter-chromosome variant-gene pairs is orders of magnitude greater than of *cis*-eQTLs.

In Chapter 5, we examined the performance of linear mixed model (LMM) methods for TWAS. Traditional GWAS methods account for the presence of population structure and relatedness among the samples via by modeling the phenotype covariance structure with a genetic relationship matrix (GRM). Alternatively, we can adjust for the population stratification by using principal components in the mean model, and we can account for the relatedness in the sample by fitting a kinship matrix as a random effect. We developed a method, which extends traditional LMM methods used for GWAS by including an additional random random effect, a transcriptome relatedness matrix (TRM). When accounting for a TRM in a linear mixed model, our test is properly calibrated in terms of type I error and yields higher power than the model that accounts for a GRM only. Moreover, our model can incorporate additional random effects, e.g. effects that account for shared environment.

All analyses considered in Chapter 5 excluded the X-chromosome. Publicly available TWAS prediction models primarily focus on autosomal genes. However, with available X-chromosome reference panels, one can train prediction models and perform TWAS of the X-chromosome genes. Since males have only one copy of the X-chromosome, while females have two, extra considerations are needed when accounting for the correlation structure on chromosome X [40]. Our method can account for X-chromosome structure by including an kinship matrix representing X-chromosome structure, in addition to a kinship matrix representing autosomal structure. This will allow for better calibrated

association tests at the genes on chromosome X.

In Chapter 6, we examined the issue of heterogeneity in phenotypic variance across subgroups in large samples. This issue arises in large-scale analyses where data are pooled from multiple studies and multiple ethnic groups. When unaccounted for, it can lead to increased false positives and decreased statistical power. To address the issue of differential phenotypic variances by subgroup, we developed a method of modeling heteroscedasticity within the linear mixed model framework. Through a simulation study, we illustrated that our method yields higher power to detect gene-trait associations. We applied our method to participants of the TOPMed program with white blood cell count measurements and identified 31 statistically significant genes under the heterogeneous variances model.

Overall, TWAS has shown to be a powerful tool for identifying complex trait - gene associations. Although various TWAS methods have been developed in the last five years, more work is expected in this field of research. We see evidence that TWA analyses can aid in our understanding of disease etiology and translational medicine. In the future, we hope to see more research in diverse populations that uncovers complex disease mechanisms and translates genomic discoveries to clinical settings and precision medicine.

BIBLIOGRAPHY

[1] A. Adeyemo and C. Rotimi. Genetic variants associated with complex human diseases show wide variation across multiple populations. *Public Health Genomics*, 13(2):72–79, 2009.

[2] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[3] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[4] Alvaro N Barbeira, Rodrigo Bonazzola, Eric R Gamazon, Yanyu Liang, YoSon Park, Sarah Kim-Hellmuth, Gao Wang, Zhuoxun Jiang, Dan Zhou, Farhad Hormozdiari, et al. Exploiting the gtex resources to decipher the mechanisms at gwas loci. *Genome biology*, 22(1):1–24, 2021.

[5] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications*, 9(1):1–20, 2018.

[6] Alvaro N Barbeira, Milton Pividori, Jiamao Zheng, Heather E Wheeler, Dan L Nicolae, and Hae Kyung Im. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS genetics*, 15(1):e1007889, 2019.

[7] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, Kenneth B. Beckman, Jianxin Shi, Rui Mei, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24, 2014.

[8] Amy R. Bentley, Shawneequa Callier, and Charles N. Rotimi. Diversity and inclusion in genomic research: why the uneven progress? *Journal of Community Genetics*, 8(4):255–266, 2017.

[9] Diane E. Bild, David A. Bluemke, Gregory L. Burke, Robert Detrano, Ana V. Diez Roux, Aaron R. Folsom, Philip Greenland, David R. Jacobs, Richard Kronmal, Kiang Liu, et al. Multi-Ethnic Study of Atherosclerosis: Objectives and design. *American Journal of Epidemiology*, 156(9):871–881, 2002.

[10] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[11] Carlos D Bustamante, Esteban González Burchard, and Francisco M De la Vega. Genomics for the world. *Nature*, 475(14 July):163–165, 2011.

[12] Minal Çalkan, Jonathan K. Pritchard, Carole Ober, and Yoav Gilad. The Effect of Freeze-Thaw Cycles on Gene Expression Levels in Lymphoblastoid Cell Lines. *PLoS ONE*, 9(9):e107166, 2014.

[13] Christopher S. Carlson, Tara C. Matise, Kari E. North, Christopher A. Haiman, Megan D. Fesinmeyer, Steven Buyske, Fredrick R. Schumacher, Ulrike Peters, Nora Franceschini, Marylyn D. Ritchie, et al. Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS Biology*, 11(9), 2013.

[14] Christopher S Carlson, Tara C Matise, Kari E North, Christopher A Haiman, Megan D Fesinmeyer, Steven Buyske, Fredrick R Schumacher, Ulrike Peters, Nora Franceschini, Marylyn D Ritchie, et al. Generalization and dilution of association results from european gwas in populations of non-european ancestry: the page study. *PLoS biology*, 11(9):e1001661, 2013.

[15] Matthew P. Conomos, Cecelia A. Laurie, Adrienne M. Stilp, Stephanie M. Gogarten, Caitlin P. McHugh, Sarah C. Nelson, Tamar Sofer, Lindsay Fernández-Rhodes, Anne E. Justice, Mariaelisa Graff, et al. Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *American Journal of Human Genetics*, 98(1):165–184, 2016.

[16] Matthew P. Conomos, Michael B. Miller, and Timothy A. Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39(4):276–293, 2015.

[17] Matthew P. Conomos, Alexander P. Reiner, Bruce S. Weir, and Timothy A. Thornton. Model-free Estimation of Recent Genetic Relatedness. *American Journal of Human Genetics*, 98(1):127–148, 2016.

[18] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[19] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[20] Michael D Gallagher and Alice S Chen-Plotkin. The post-gwas era: from association to function. *The American Journal of Human Genetics*, 102(5):717–730, 2018.

[21] Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, Dan L. Nicolae, Nancy J. Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

[22] Stephanie M Gogarten, Tamar Sofer, Han Chen, Chaoyu Yu, Jennifer A Brody, Timothy A Thornton, Kenneth M Rice, and Matthew P Conomos. Genetic association testing using the genesis r/bioconductor package. *Bioinformatics*, 35(24):5346–5348, 2019.

[23] Assaf Gottlieb, Roxana Daneshjou, Marianne DeGorter, Stephane Bourgeois, Peter J. Svensson, Mia Wadelius, Panos Deloukas, Stephen B. Montgomery, and Russ B. Altman. Cohort-specific imputation of gene expression improves prediction of warfarin dose for African Americans. *Genome Medicine*, 9(1):1–9, 2017.

[24] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-po Yang, Eshwar Meduri, Amy Barrett, et al. Articles Mapping cis- and trans -regulatory effects across multiple tissues in twins. 44(10), 2012.

[25] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W.J.H. Penninx, Rick Jansen, Eco J.C. De Geus, Dorret I. Boomsma, Fred A. Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2016.

[26] Lucia A. Hindorff, Vence L. Bonham, Lawrence C. Brody, Margaret E.C. Ginoza, Carolyn M. Hutter, Teri A. Manolio, and Eric D. Green. Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, 19(3):175–185, 2018.

[27] Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M Zekavat, Zhaolong Yu, Boyang Li, Jianlei Gu, Sydney Muchnik, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics*, 51(3):568–576, 2019.

[28] Derek E. Kelly, Matthew E.B. Hansen, and Sarah A. Tishkoff. Global variation in gene expression and the value of diverse sampling. *Current Opinion in Systems Biology*, 1:102–108, 2017.

[29] Kevin L Keys, Angel CY Mak, Marquitta J White, Walter L Eckalbar, Andrew W Dahl, Joel Mefford, Anna V Mikhaylova, María G Contreras, Jennifer R Elhawary, Celeste Eng, et al. On the cross-population generalizability of gene expression prediction models. *PLoS genetics*, 16(8):e1008927, 2020.

[30] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A.C. 'T Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[31] Binglan Li, Shefali S. Verma, Yogasudha C Veturi, Anurag Verma, Yuki Bradford, David W. Haas, and Marylyn D Ritchie. Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. In *Biocomputing 2018*, pages 448–459, 2018.

[32] Yun R. Li and Brendan J. Keating. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Medicine*, 6(10):91, oct 2014.

[33] Xuanyao Liu, Yang I Li, and Jonathan K Pritchard. Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv*, page 425108, 2018.

[34] J Lonsdale, J Thomas, M Salvatore, R Phillips, and E Lo. The Genotype-Tissue Expression (GTEx) project : Nature Genetics : Nature Publishing Group. *Nature*, 2013.

[35] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 2017.

[36] Nicholas Mancuso, Malika K Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Probabilistic fine-mapping of transcriptome-wide association studies. *Nature genetics*, 51(4):675–682, 2019.

[37] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[38] Arjun K. Manrai, Birgit H. Funke, Heidi L. Rehm, Morten S. Olesen, Bradley A. Maron, Peter Szolovits, David M. Margulies, Joseph Loscalzo, and Isaac S. Kohane. Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*, 375(7):655–665, 2016.

[39] Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100(4):635–649, 2017.

[40] Caitlin Patricia McHugh. *Statistical Methods for the Analysis of Autosomal and X Chromosome Genetic Data in Samples with Unknown Structure*. PhD thesis, 2016.

[41] Lauren S Mogil, Angela Andaleon, Alexa Badalamenti, Scott P Dickinson, Xiuqing Guo, Jerome I Rotter, W Craig Johnson, Hae Kyung Im, Yongmei Liu, and Heather E Wheeler. Genetic architecture of gene expression traits across diverse populations Author summary. *PLoS Genetics*, pages 1–17, 2018.

[42] Cassandra E Murcray, Juan Pablo Lewinger, and W James Gauderman. Gene-environment interaction in genome-wide association studies. *American journal of epidemiology*, 169(2):219–226, 2009.

[43] Shaila Musharoff, Danny Park, Andy Dahl, Joshua Galanter, Xuanyao Liu, Scott Huntsman, Celeste Eng, Esteban G Burchard, Julien F Ayroles, and Noah Zaitlen. Existence and implications of population variance structure. *bioRxiv*, page 439661, 2018.

[44] Sini Nagpal, Xiaoran Meng, Michael P Epstein, Lam C Tsoi, Matthew Patrick, Greg Gibson, Philip L De Jager, David A Bennett, Aliza P Wingo, Thomas S Wingo, et al. Tigar: an improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *The American Journal of Human Genetics*, 105(2):258–266, 2019.

[45] Anna C. Need and David B. Goldstein. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, 25(11):489–494, 2009.

[46] NHLBI. NHLBI Trans-Omics for Precision Medicine Whole Genome Sequencing Program. TOPMed. 2016.

[47] Alexandra C. Nica, Stephen B. Montgomery, Antigone S. Dimas, Barbara E. Stranger, Claude Beazley, Inês Barroso, and Emmanouil T. Dermitzakis. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, 6(4), 2010.

[48] Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*, 6(4), 2010.

[49] Evangelia E Ntzani, George Liberopoulos, Teri A Manolio, and John PA Ioannidis. Consistency of genome-wide associations across major ancestral groups. *Human genetics*, 131(7):1057–1071, 2012.

[50] Sam S. Oh, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E. Barcelo, Marquitta J. White, Danielle M. de Bruin, Ruth M. Greenblatt, Kirsten Bibbins-Domingo, Alan H.B. Wu, et al. Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. *PLoS Medicine*, 12(12):1–9, 2015.

[51] Sam S. Oh, Marquitta J. White, Christopher R. Gignoux, and Esteban G. Burchard. Making precision medicine socially precise: Take a deep breath. *American Journal of Respiratory and Critical Care Medicine*, 193(4):348–350, 2016.

[52] Yongjin Park, Abhishek Sarkar, Kunal Bhutani, and Manolis Kellis. Multi-tissue polygenic models for transcriptome-wide association studies. *bioRxiv*, page 107623, 2017.

[53] Bogdan Pasaniuc and Alkes L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18(2):117–127, 2016.

[54] Slavé Petrovski and David B. Goldstein. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biology*, 17(1):157, 2016.

[55] Alice B. Popejoy and Stephanie M. Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 2016.

[56] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[57] Lars Rönnegård and William Valdar. Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics*, 188(2):435–447, 2011.

[58] Lars Rönnegård and William Valdar. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC genetics*, 13(1):1–7, 2012.

[59] Lulu Shang, Jennifer A Smith, Wei Zhao, Minjung Kho, Stephen T Turner, Thomas H Mosley, Sharon LR Kardia, and Xiang Zhou. Genetic architecture of gene expression in european and african americans: an eqtl mapping study in genoa. *The American Journal of Human Genetics*, 106(4):496–512, 2020.

[60] Xingjie Shi, Xiaoran Chai, Yi Yang, Qing Cheng, Yuling Jiao, Haoyue Chen, Jian Huang, Can Yang, and Jin Liu. A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. *Nucleic acids research*, 48(19):e109–e109, 2020.

[61] Tamar Sofer, Xiuwen Zheng, Stephanie M Gogarten, Cecelia A Laurie, Kelsey Grinde, John R Shaffer, Dmitry Shungin, Jeffrey R OConnell, Ramon A Durazo-Arvizo, Laura Raffield, et al. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic epidemiology*, 43(3):263–275, 2019.

[62] Tamar Sofer, Xiuwen Zheng, Cecelia A Laurie, Stephanie M Gogarten, Jennifer A Brody, Matthew P Conomos, Joshua C Bis, Timothy A Thornton, Adam Szpiro, Jeffrey R. O'Connell, et al. Variant-specific inflation factors for assessing population stratification at the phenotypic variance level. *Nature Communications*, 12(1):3506, dec 2021.

[63] Barbara E. Stranger, Stephen B. Montgomery, Antigone S. Dimas, Leopold Parts, Oliver Stegle, Catherine E. Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez-Arcelus, et al. Patterns of Cis regulatory variation in diverse human populations. *PLoS Genetics*, 8(4), 2012.

[64] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk

biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[65] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

[66] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[67] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[68] Jason M. Torres, Eric R. Gamazon, Esteban J. Parra, Jennifer E. Below, Adan Valladares-Salgado, Niels Wacher, Miguel Cruz, Craig L. Hanis, and Nancy J. Cox. Cross-tissue and tissue-specific eQTLs: Partitioning the heritability of a complex trait. *American Journal of Human Genetics*, 95(5):521–534, 2014.

[69] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010.

[70] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American journal of human genetics*, 101(1):5–22, 2017.

[71] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599, 2019.

[72] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[73] Can Yang, Xiang Wan, Xinyi Lin, Mengjie Chen, Xiang Zhou, and Jin Liu. Comm: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*, 35(10):1644–1652, 2019.

[74] Yuan Yuan, Lei Tian, Dongsheng Lu, and Shuhua Xu. Analysis of Genome-Wide RNA-Sequencing Data Suggests Age of the CEPH/Utah (CEU) Lymphoblastoid Cell Lines Systematically Biases Gene Expression Profiles. *Scientific Reports*, 5:1–5, 2015.

[75] Zhongshang Yuan, Huanhuan Zhu, Ping Zeng, Sheng Yang, Shiquan Sun, Can Yang, Jin Liu, and Xiang Zhou. Testing and controlling for horizontal pleiotropy with probabilistic mendelian randomization in transcriptome-wide association studies. *Nature communications*, 11(1):1–14, 2020.

[76] Ping Zeng and Xiang Zhou. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nature communications*, 8(1):1–11, 2017.

[77] Hui Zou and Trevor Hastie. Regression shrinkage and selection via the elastic net, with applications to microarrays. *JR Stat Soc Ser B*, 67:301–20, 2003.

Appendix A

SUPPLEMENTARY TABLES AND FIGURES

We present the results for the four PrediXcan weight databases omitted from the main text. The databases are GTEx v6 1KG whole blood (GTEx v6 1KG WB), GTEx v6 1KG LCL (GTEx v6 1KG LCL), GTEx v6 HapMap whole blood (GTEx v6 HM WB), and GTEx v6 HapMap LCL (GTEx v6 HM LCL).

Table A.1: Number of genes for which Pearson correlation coefficients are available by population and by GTEx v6 database.

PrediXcan database	1KG WB	1KG LCL	HM WB	HM LCL
Genes with observed and predicted expression values	6,179	3,662	6,039	3,363
By population:				
CEU	6,136	3,636	6,017	3,361
FIN	6,132	3,637	6,020	3,361
GBR	6,133	3,637	6,018	3,361
TSI	6,141	3,637	6,025	3,361
YRI	6,039	3,612	5,997	3,351
Genes before filtering	6,010	3,604	5,978	3,350
Genes after filtering	2,198	1,889	2,207	1,847

Table A.2: Binned gene correlation coefficients for the five populations using GTEx v6 weight databases.

	Unfiltered					Filtered				
	CEU	FIN	GBR	TSI	YRI	CEU	FIN	GBR	TSI	YRI
	GTEx v6 1KG WB database									
$r < 0$	2,094	2,031	1,993	2,051	2,285	347	313	329	334	492
$0 < r < 0.2$	2,804	2,747	2,673	2,770	2,826	870	809	730	804	976
$0.2 < r < 0.4$	856	892	961	867	708	727	737	757	739	540
$0.4 < r < 0.6$	183	249	280	240	147	181	248	279	239	146
$0.6 < r < 0.8$	64	82	88	69	39	64	82	88	69	39
$0.8 < r < 1$	9	9	15	13	5	9	9	15	13	5
	GTEx v6 1KG LCL database									
$r < 0$	841	806	799	804	1030	97	99	86	100	236
$0 < r < 0.2$	1,570	1,492	1,460	1,544	1,666	673	570	551	612	817
$0.2 < r < 0.4$	849	841	833	829	676	775	755	740	750	604
$0.4 < r < 0.6$	267	339	376	320	178	267	339	376	320	178
$0.6 < r < 0.8$	69	112	117	89	48	69	112	117	89	48
$0.8 < r < 1$	8	14	19	18	6	8	14	19	18	6
	GTEx v6 HapMap WB database									
$r < 0$	2,092	2,052	1,986	2,065	2,246	350	331	329	368	504
$0 < r < 0.2$	2,770	2,705	2,679	2,760	2,883	885	804	763	830	1,004
$0.2 < r < 0.4$	879	892	942	846	685	736	743	745	703	535
$0.4 < r < 0.6$	176	248	279	229	128	175	248	278	228	128
$0.6 < r < 0.8$	55	72	79	68	33	55	72	79	68	33
$0.8 < r < 1$	6	9	13	10	3	6	9	13	10	3
	GTEx v6 HapMap LCL database									
$r < 0$	750	691	676	688	947	87	77	89	81	256
$0 < r < 0.2$	1,471	1,405	1,365	1,434	1,568	689	593	536	601	820
$0.2 < r < 0.4$	810	797	847	834	628	752	720	760	771	564
$0.4 < r < 0.6$	257	350	328	299	164	257	350	328	299	164
$0.6 < r < 0.8$	53	93	119	78	39	53	93	119	78	39
$0.8 < r < 1$	9	14	15	17	4	9	14	15	17	4

Table A.3: Results from linear mixed models for population category (with CEU as a reference) and change in gene correlation coefficient among filtered genes.

Regression	GTEx v6 1KG WB		GTEx v6 1KG LCL		GTEx v6 HM WB		GTEx v6 HM LCL	
parameter	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI		
FIN	0.025	(0.018, 0.032)	0.031	(0.024, 0.039)	0.020	(0.013, 0.027)	0.034	(0.026, 0.041)
GBR	0.031	(0.023, 0.038)	0.043	(0.035, 0.050)	0.028	(0.021, 0.035)	0.044	(0.036, 0.051)
TSI	0.015	(0.008, 0.022)	0.023	(0.016, 0.031)	0.011	(0.004, 0.018)	0.027	(0.019, 0.034)
YRI	-0.040	(-0.047, -0.033)	-0.061	(-0.069, -0.054)	-0.047	(-0.054, -0.040)	-0.064	(-0.071, -0.056)

Table A.4: Results from linear mixed models for population category (excluding CEU, with FIN as a reference) and change in gene correlation coefficient among filtered genes.

Regression	GTEx v6 1KG WB		GTEx v6 1KG LCL		GTEx v6 HM WB		GTEx v6 HM LCL	
parameter	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI		
GBR	0.006	(-0.001, 0.013)	0.012	(0.004, 0.019)	0.008	(0, 0.015)	0.010	(0.002, 0.018)
TSI	-0.009	(-0.017, -0.002)	-0.008	(-0.016, 0)	-0.010	(-0.017, -0.002)	-0.007	(-0.015, 0)
YRI	-0.064	(-0.072, -0.057)	-0.092	(-0.100, -0.085)	-0.067	(-0.075, -0.060)	-0.099	(-0.107, -0.091)

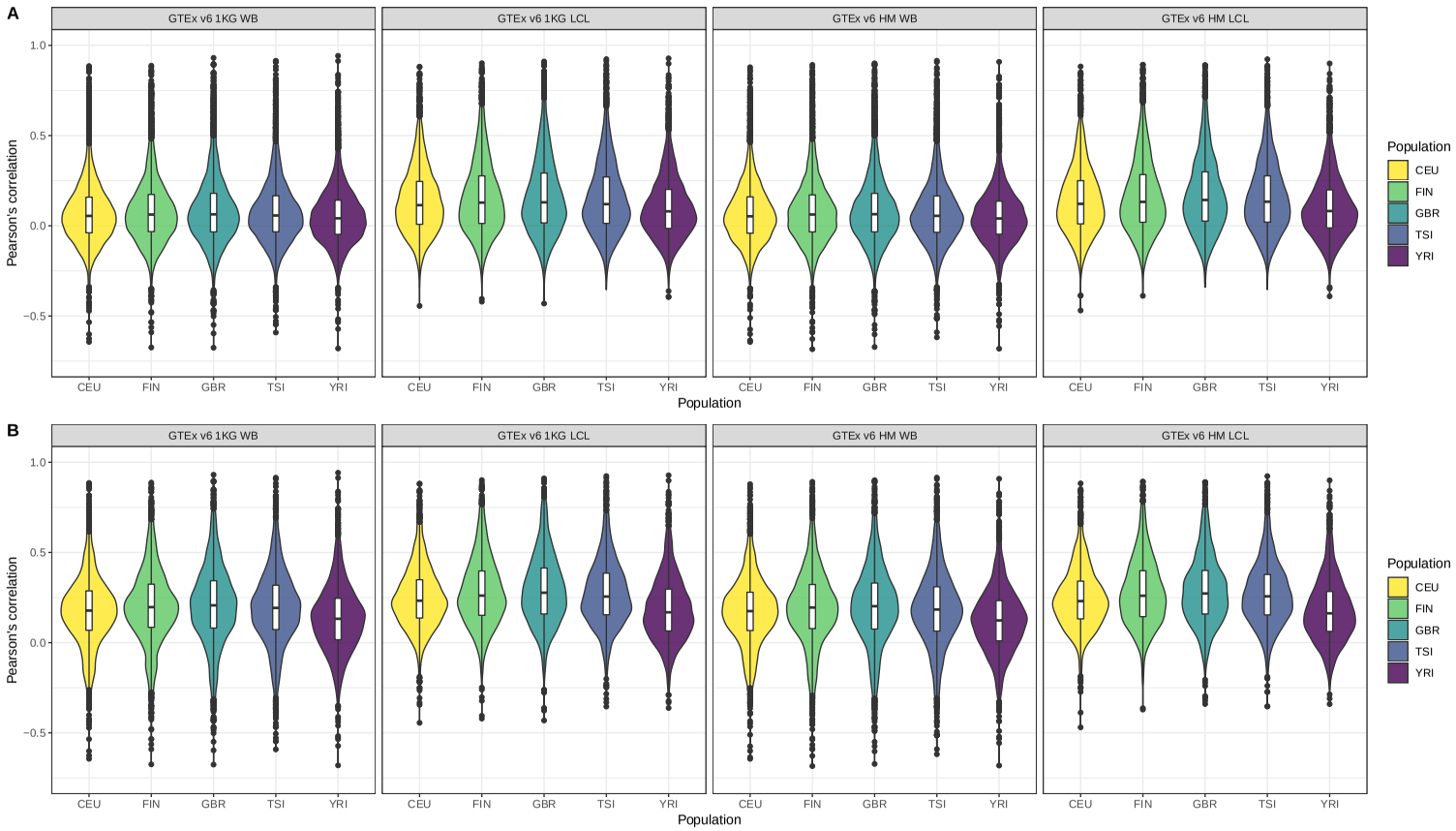


Figure A.1: Violin plots of gene expression correlation coefficients by five populations using GTEx weight databases; **(A)** before and **(B)** after filtering out poorly predicted genes.

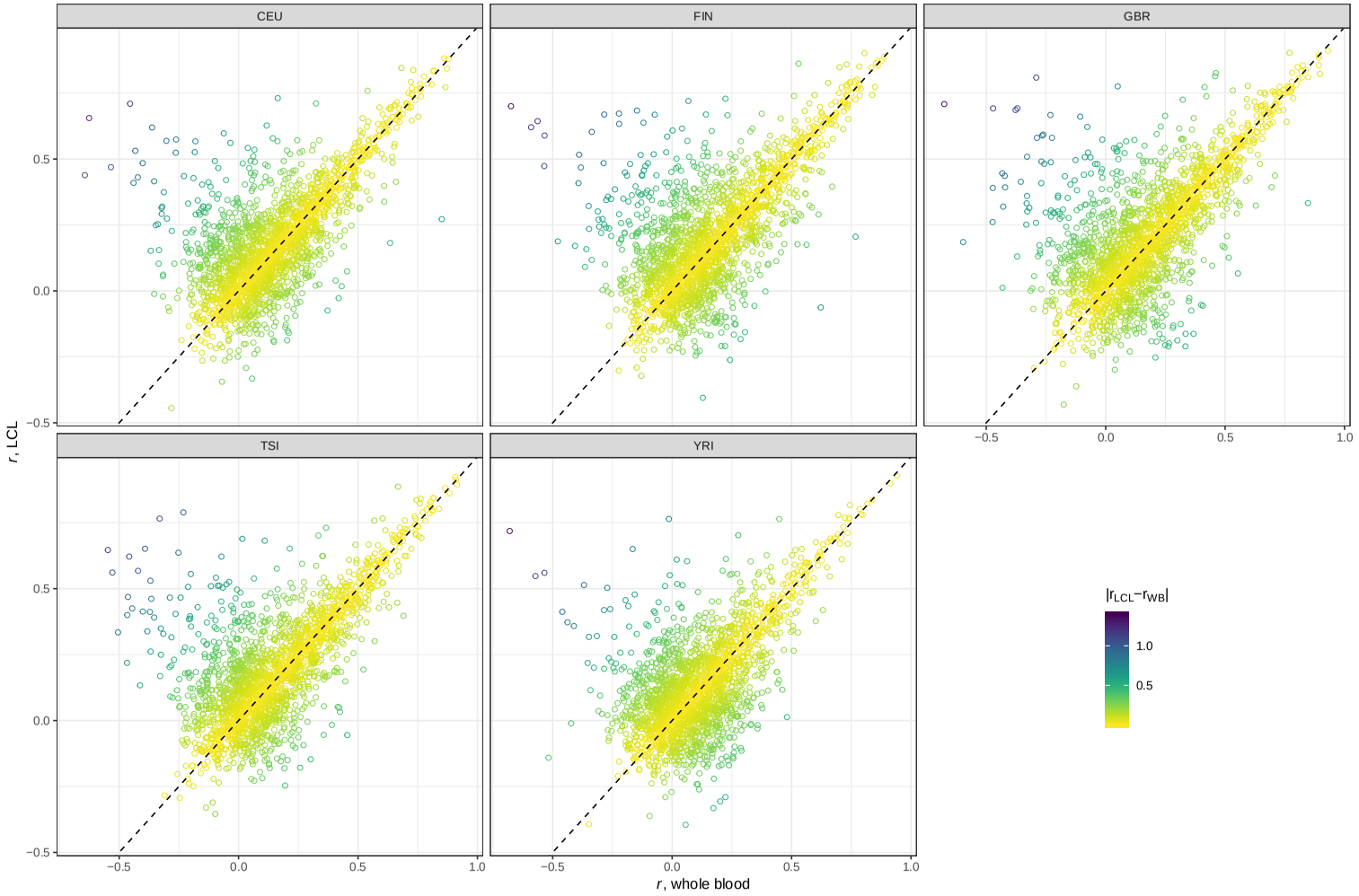


Figure A.2: Scatter plots comparing gene correlation coefficients by population using GTEx v6 1KG LCL vs GTEx v6 1KG WB databases.

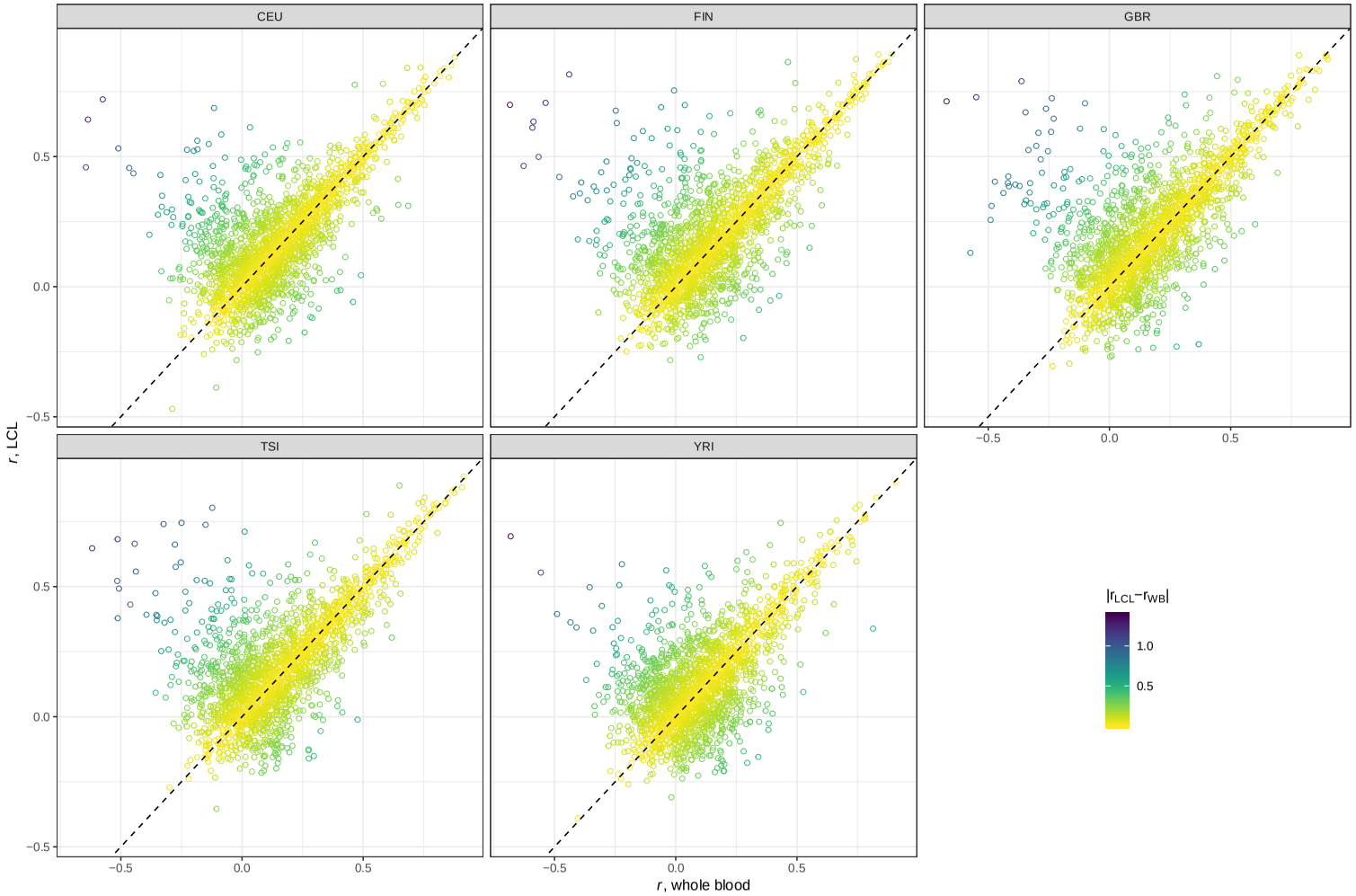


Figure A.3: Scatter plots comparing gene correlation coefficients by population using GTEx v6 HapMap LCL vs GTEx v6 HapMap WB databases.

Appendix B

SUPPLEMENTARY TABLES AND FIGURES

Table B.1: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of eQTLs with all eQTLs and the corresponding effect sizes shared between populations.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
EN anc	ESN	1	1	1	0.316 (0.229, 0.398)
EN anc	GBR	1	1	1	0.326 (0.236, 0.41)
EN comb	ESN	1	1	1	0.36 (0.292, 0.419)
EN comb	GBR	1	1	1	0.366 (0.289, 0.431)
Fused lasso	ESN	1	1	1	0.339 (0.248, 0.404)
Fused lasso	GBR	1	1	1	0.338 (0.258, 0.407)
Lasso anc	ESN	1	1	1	0.327 (0.234, 0.402)
Lasso anc	GBR	1	1	1	0.332 (0.241, 0.41)
Lasso comb	ESN	1	1	1	0.366 (0.298, 0.426)
Lasso comb	GBR	1	1	1	0.369 (0.295, 0.436)
EN anc	ESN	2	1	1	0.267 (0.171, 0.351)
EN anc	GBR	2	1	1	0.282 (0.191, 0.368)
EN comb	ESN	2	1	1	0.33 (0.259, 0.396)
EN comb	GBR	2	1	1	0.339 (0.267, 0.404)
Fused lasso	ESN	2	1	1	0.3 (0.213, 0.373)
Fused lasso	GBR	2	1	1	0.319 (0.233, 0.387)
Lasso anc	ESN	2	1	1	0.274 (0.17, 0.354)

Table B.1: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of eQTLs with all eQTLs and the corresponding effect sizes shared between populations.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
Lasso anc	GBR	2	1	1	0.282 (0.183, 0.367)
Lasso comb	ESN	2	1	1	0.328 (0.259, 0.398)
Lasso comb	GBR	2	1	1	0.34 (0.266, 0.404)
EN anc	ESN	10	1	1	0.153 (0.068, 0.244)
EN anc	GBR	10	1	1	0.199 (0.108, 0.276)
EN comb	ESN	10	1	1	0.219 (0.139, 0.298)
EN comb	GBR	10	1	1	0.249 (0.165, 0.328)
Fused lasso	ESN	10	1	1	0.219 (0.131, 0.293)
Fused lasso	GBR	10	1	1	0.245 (0.165, 0.324)
Lasso anc	ESN	10	1	1	0.149 (0.067, 0.241)
Lasso anc	GBR	10	1	1	0.189 (0.091, 0.273)
Lasso comb	ESN	10	1	1	0.213 (0.131, 0.296)
Lasso comb	GBR	10	1	1	0.241 (0.16, 0.328)

Table B.2: Correlations between predicted and simulated phenotypes for different methods and test populations across different proportions of shared eQTLs and different number of causal eQTLs with same direction effect sizes.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
EN anc	ESN	1	0	1	0.32 (0.238, 0.396)
EN anc	GBR	1	0	1	0.339 (0.255, 0.413)
EN comb	ESN	1	0	1	0.181 (0.058, 0.286)
EN comb	GBR	1	0	1	0.188 (0.078, 0.296)
Fused lasso	ESN	1	0	1	0.268 (0.147, 0.357)
Fused lasso	GBR	1	0	1	0.29 (0.168, 0.382)
Lasso anc	ESN	1	0	1	0.317 (0.228, 0.399)
Lasso anc	GBR	1	0	1	0.342 (0.251, 0.412)
Lasso comb	ESN	1	0	1	0.177 (0.05, 0.285)
Lasso comb	GBR	1	0	1	0.187 (0.07, 0.296)
EN anc	ESN	2	0	1	0.284 (0.188, 0.359)
EN anc	GBR	2	0	1	0.278 (0.192, 0.351)
EN anc	ESN	2	0.5	1	0.271 (0.182, 0.356)
EN anc	GBR	2	0.5	1	0.287 (0.188, 0.357)
EN comb	ESN	2	0	1	0.178 (0.072, 0.286)
EN comb	GBR	2	0	1	0.17 (0.07, 0.277)
EN comb	ESN	2	0.5	1	0.263 (0.168, 0.346)
EN comb	GBR	2	0.5	1	0.264 (0.16, 0.351)
Fused lasso	ESN	2	0	1	0.225 (0.111, 0.31)
Fused lasso	GBR	2	0	1	0.22 (0.11, 0.318)
Fused lasso	ESN	2	0.5	1	0.252 (0.157, 0.345)
Fused lasso	GBR	2	0.5	1	0.269 (0.162, 0.354)

Table B.2: Correlations between predicted and simulated phenotypes for different methods and test populations across different proportions of shared eQTLs and different number of causal eQTLs with same direction effect sizes.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
Lasso anc	ESN	2	0	1	0.286 (0.183, 0.361)
Lasso anc	GBR	2	0	1	0.276 (0.184, 0.357)
Lasso anc	ESN	2	0.5	1	0.279 (0.179, 0.357)
Lasso anc	GBR	2	0.5	1	0.284 (0.194, 0.368)
Lasso comb	ESN	2	0	1	0.167 (0.065, 0.283)
Lasso comb	GBR	2	0	1	0.164 (0.062, 0.268)
Lasso comb	ESN	2	0.5	1	0.26 (0.162, 0.348)
Lasso comb	GBR	2	0.5	1	0.27 (0.161, 0.351)
EN anc	ESN	10	0	1	0.208 (0.115, 0.288)
EN anc	GBR	10	0	1	0.194 (0.114, 0.281)
EN anc	ESN	10	0.5	1	0.182 (0.086, 0.273)
EN anc	GBR	10	0.5	1	0.186 (0.098, 0.272)
EN comb	ESN	10	0	1	0.173 (0.082, 0.259)
EN comb	GBR	10	0	1	0.149 (0.059, 0.234)
EN comb	ESN	10	0.5	1	0.199 (0.094, 0.284)
EN comb	GBR	10	0.5	1	0.196 (0.105, 0.278)
Fused lasso	ESN	10	0	1	0.175 (0.083, 0.259)
Fused lasso	GBR	10	0	1	0.164 (0.078, 0.244)
Fused lasso	ESN	10	0.5	1	0.184 (0.096, 0.273)
Fused lasso	GBR	10	0.5	1	0.212 (0.126, 0.282)
Lasso anc	ESN	10	0	1	0.199 (0.106, 0.286)
Lasso anc	GBR	10	0	1	0.191 (0.112, 0.276)
Lasso anc	ESN	10	0.5	1	0.18 (0.079, 0.267)

Table B.2: Correlations between predicted and simulated phenotypes for different methods and test populations across different proportions of shared eQTLs and different number of causal eQTLs with same direction effect sizes.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
Lasso anc	GBR	10	0.5	1	0.187 (0.099, 0.262)
Lasso comb	ESN	10	0	1	0.164 (0.078, 0.258)
Lasso comb	GBR	10	0	1	0.148 (0.056, 0.234)
Lasso comb	ESN	10	0.5	1	0.193 (0.087, 0.279)
Lasso comb	GBR	10	0.5	1	0.196 (0.096, 0.272)

Table B.3: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of causal eQTLs with the same location but different direction of effect sizes between populations.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
EN anc	ESN	1	1	0	0.322 (0.234, 0.396)
EN anc	GBR	1	1	0	0.331 (0.244, 0.401)
EN comb	ESN	1	1	0	0.049 (-0.046, 0.147)
EN comb	GBR	1	1	0	0.09 (-0.018, 0.188)
Fused lasso	ESN	1	1	0	0.291 (0.173, 0.377)
Fused lasso	GBR	1	1	0	0.309 (0.21, 0.393)
Lasso anc	ESN	1	1	0	0.329 (0.234, 0.407)
Lasso anc	GBR	1	1	0	0.333 (0.25, 0.402)
Lasso comb	ESN	1	1	0	0.049 (-0.047, 0.151)
Lasso comb	GBR	1	1	0	0.086 (-0.018, 0.187)
EN anc	ESN	2	1	0	0.267 (0.17, 0.348)
EN anc	GBR	2	1	0	0.29 (0.199, 0.37)
EN anc	ESN	2	1	0.5	0.275 (0.171, 0.355)
EN anc	GBR	2	1	0.5	0.289 (0.202, 0.374)
EN comb	ESN	2	1	0	0.044 (-0.044, 0.13)
EN comb	GBR	2	1	0	0.09 (0.003, 0.184)
EN comb	ESN	2	1	0.5	0.155 (0.048, 0.253)
EN comb	GBR	2	1	0.5	0.196 (0.081, 0.29)
Fused lasso	ESN	2	1	0	0.219 (0.101, 0.32)
Fused lasso	GBR	2	1	0	0.25 (0.127, 0.349)
Fused lasso	ESN	2	1	0.5	0.226 (0.112, 0.325)
Fused lasso	GBR	2	1	0.5	0.267 (0.16, 0.353)

Table B.3: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of causal eQTLs with the same location but different direction of effect sizes between populations.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
Lasso anc	ESN	2	1	0	0.268 (0.171, 0.351)
Lasso anc	GBR	2	1	0	0.285 (0.19, 0.368)
Lasso anc	ESN	2	1	0.5	0.272 (0.162, 0.358)
Lasso anc	GBR	2	1	0.5	0.285 (0.188, 0.373)
Lasso comb	ESN	2	1	0	0.045 (-0.045, 0.128)
Lasso comb	GBR	2	1	0	0.08 (0.001, 0.182)
Lasso comb	ESN	2	1	0.5	0.156 (0.045, 0.255)
Lasso comb	GBR	2	1	0.5	0.196 (0.084, 0.289)
EN anc	ESN	10	1	0	0.155 (0.07, 0.248)
EN anc	GBR	10	1	0	0.195 (0.108, 0.276)
EN anc	ESN	10	1	0.5	0.151 (0.074, 0.243)
EN anc	GBR	10	1	0.5	0.194 (0.103, 0.281)
EN comb	ESN	10	1	0	0.034 (-0.041, 0.121)
EN comb	GBR	10	1	0	0.091 (0, 0.182)
EN comb	ESN	10	1	0.5	0.097 (0.014, 0.173)
EN comb	GBR	10	1	0.5	0.146 (0.059, 0.23)
Fused lasso	ESN	10	1	0	0.104 (0.017, 0.181)
Fused lasso	GBR	10	1	0	0.145 (0.063, 0.234)
Fused lasso	ESN	10	1	0.5	0.114 (0.017, 0.202)
Fused lasso	GBR	10	1	0.5	0.175 (0.078, 0.257)
Lasso anc	ESN	10	1	0	0.157 (0.069, 0.241)
Lasso anc	GBR	10	1	0	0.185 (0.099, 0.271)
Lasso anc	ESN	10	1	0.5	0.149 (0.057, 0.239)

Table B.3: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of causal eQTLs with the same location but different direction of effect sizes between populations.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
Lasso anc	GBR	10	1	0.5	0.191 (0.09, 0.28)
Lasso comb	ESN	10	1	0	0.039 (-0.039, 0.119)
Lasso comb	GBR	10	1	0	0.089 (0, 0.177)
Lasso comb	ESN	10	1	0.5	0.092 (0.013, 0.174)
Lasso comb	GBR	10	1	0.5	0.141 (0.05, 0.225)

Table B.4: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of causal eQTLs with different locations and different direction of effect sizes between populations.

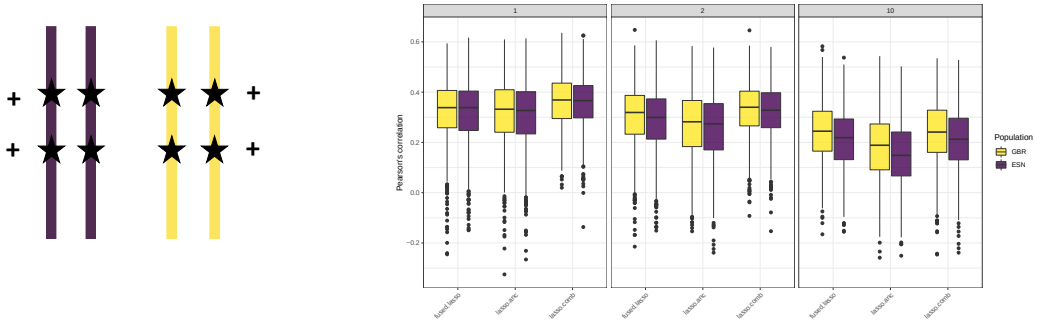
Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
EN anc	ESN	1	0	0	0.319 (0.225, 0.393)
EN anc	GBR	1	0	0	0.334 (0.242, 0.411)
EN comb	ESN	1	0	0	0.114 (0.013, 0.216)
EN comb	GBR	1	0	0	0.124 (0.029, 0.22)
Fused lasso	ESN	1	0	0	0.281 (0.163, 0.368)
Fused lasso	GBR	1	0	0	0.302 (0.175, 0.391)
Lasso anc	ESN	1	0	0	0.312 (0.225, 0.395)
Lasso anc	GBR	1	0	0	0.331 (0.246, 0.41)
Lasso comb	ESN	1	0	0	0.115 (0.012, 0.214)
Lasso comb	GBR	1	0	0	0.123 (0.036, 0.217)
EN anc	ESN	2	0	0	0.274 (0.184, 0.36)
EN anc	GBR	2	0	0	0.287 (0.195, 0.372)
EN anc	ESN	2	0.5	0.5	0.272 (0.167, 0.352)
EN anc	GBR	2	0.5	0.5	0.289 (0.203, 0.36)
EN comb	ESN	2	0	0	0.1 (0.001, 0.193)
EN comb	GBR	2	0	0	0.121 (0.029, 0.216)
EN comb	ESN	2	0.5	0.5	0.187 (0.085, 0.271)
EN comb	GBR	2	0.5	0.5	0.21 (0.113, 0.295)
Fused lasso	ESN	2	0	0	0.224 (0.111, 0.327)
Fused lasso	GBR	2	0	0	0.251 (0.139, 0.344)
Fused lasso	ESN	2	0.5	0.5	0.224 (0.127, 0.322)
Fused lasso	GBR	2	0.5	0.5	0.258 (0.164, 0.337)

Table B.4: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of causal eQTLs with different locations and different direction of effect sizes between populations.

Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
Lasso anc	ESN	2	0	0	0.277 (0.181, 0.357)
Lasso anc	GBR	2	0	0	0.293 (0.195, 0.371)
Lasso anc	ESN	2	0.5	0.5	0.277 (0.172, 0.356)
Lasso anc	GBR	2	0.5	0.5	0.293 (0.206, 0.36)
Lasso comb	ESN	2	0	0	0.094 (-0.001, 0.186)
Lasso comb	GBR	2	0	0	0.115 (0.028, 0.209)
Lasso comb	ESN	2	0.5	0.5	0.187 (0.088, 0.274)
Lasso comb	GBR	2	0.5	0.5	0.208 (0.112, 0.292)
EN anc	ESN	10	0	0	0.209 (0.122, 0.293)
EN anc	GBR	10	0	0	0.196 (0.104, 0.276)
EN anc	ESN	10	0.5	0.5	0.171 (0.088, 0.264)
EN anc	GBR	10	0.5	0.5	0.206 (0.112, 0.287)
EN comb	ESN	10	0	0	0.083 (-0.009, 0.173)
EN comb	GBR	10	0	0	0.1 (0.023, 0.185)
EN comb	ESN	10	0.5	0.5	0.119 (0.031, 0.202)
EN comb	GBR	10	0.5	0.5	0.161 (0.073, 0.243)
Fused lasso	ESN	10	0	0	0.133 (0.039, 0.228)
Fused lasso	GBR	10	0	0	0.126 (0.033, 0.216)
Fused lasso	ESN	10	0.5	0.5	0.126 (0.031, 0.229)
Fused lasso	GBR	10	0.5	0.5	0.174 (0.087, 0.262)
Lasso anc	ESN	10	0	0	0.207 (0.114, 0.293)
Lasso anc	GBR	10	0	0	0.191 (0.1, 0.27)
Lasso anc	ESN	10	0.5	0.5	0.172 (0.081, 0.256)

Table B.4: Correlations between predicted and simulated phenotypes for different methods and test populations across different number of causal eQTLs with different locations and different direction of effect sizes between populations.

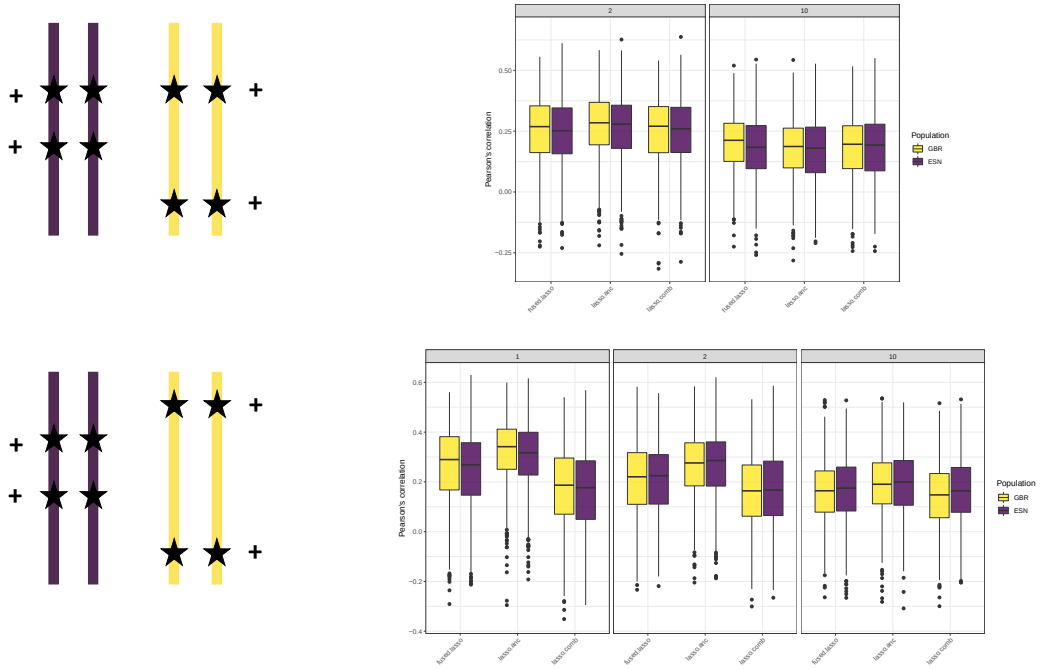
Method	Test popn	k	Prop eQTLs	Prop effects	Median (IQR)
Lasso anc	GBR	10	0.5	0.5	0.202 (0.106, 0.287)
Lasso comb	ESN	10	0	0	0.078 (-0.012, 0.172)
Lasso comb	GBR	10	0	0	0.102 (0.027, 0.181)
Lasso comb	ESN	10	0.5	0.5	0.114 (0.025, 0.203)
Lasso comb	GBR	10	0.5	0.5	0.161 (0.071, 0.24)



(a) genetic architecture

(b) simulation results

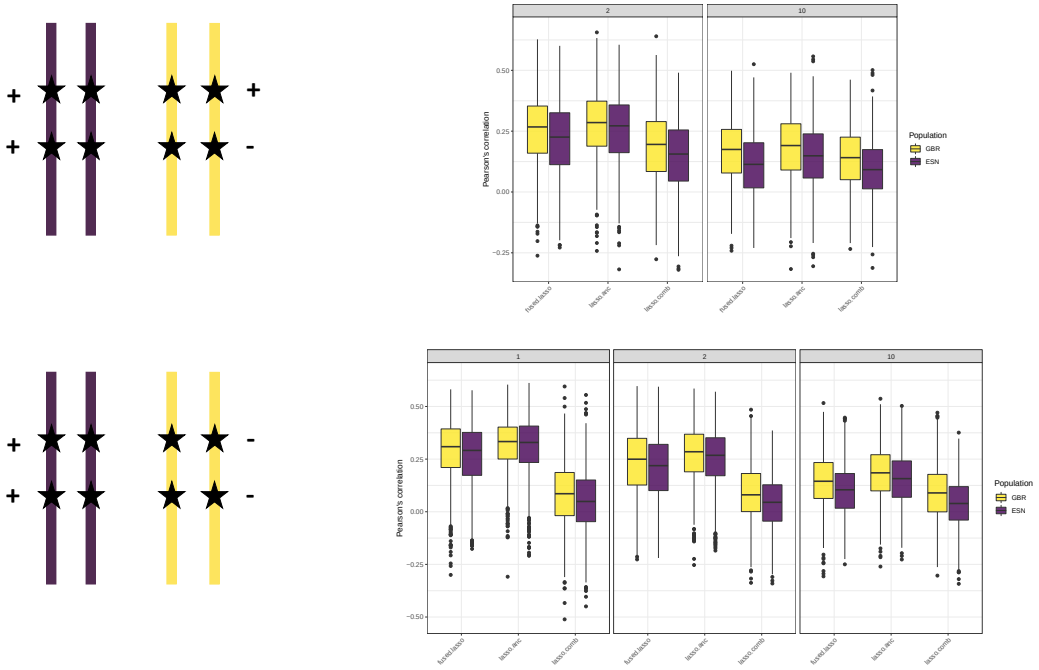
Figure B.1: **Simulation results comparing fused lasso vs standard lasso: similar genetic architecture.**



(a) genetic architecture

(b) simulation results

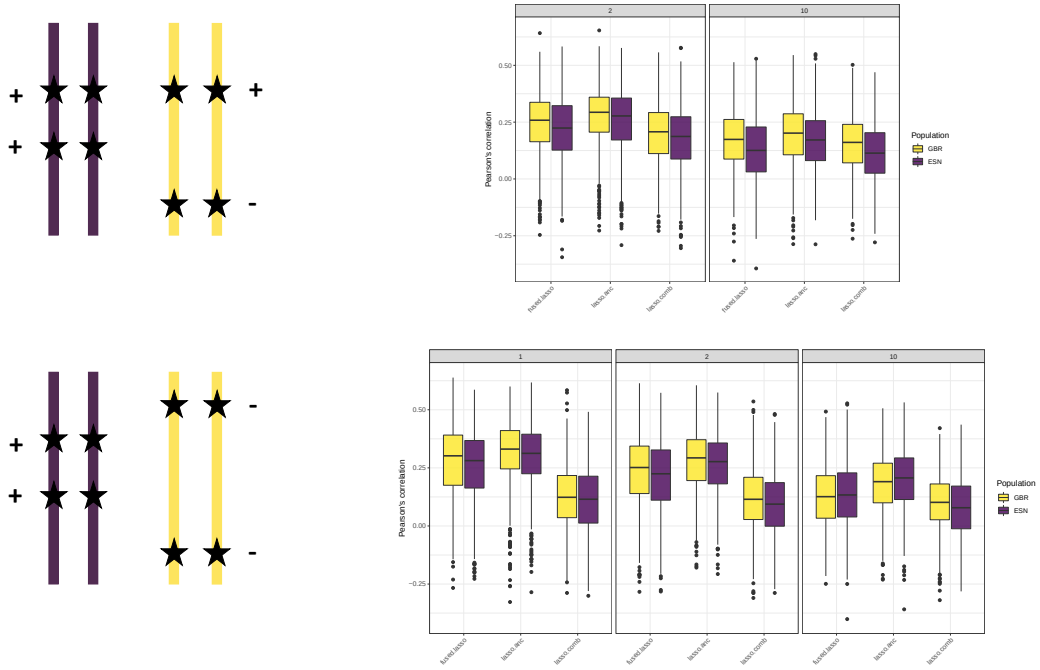
Figure B.2: **Simulation results comparing fused lasso vs standard lasso: different location of eQTLs, same direction of effects.**



(a) genetic architecture

(b) simulation results

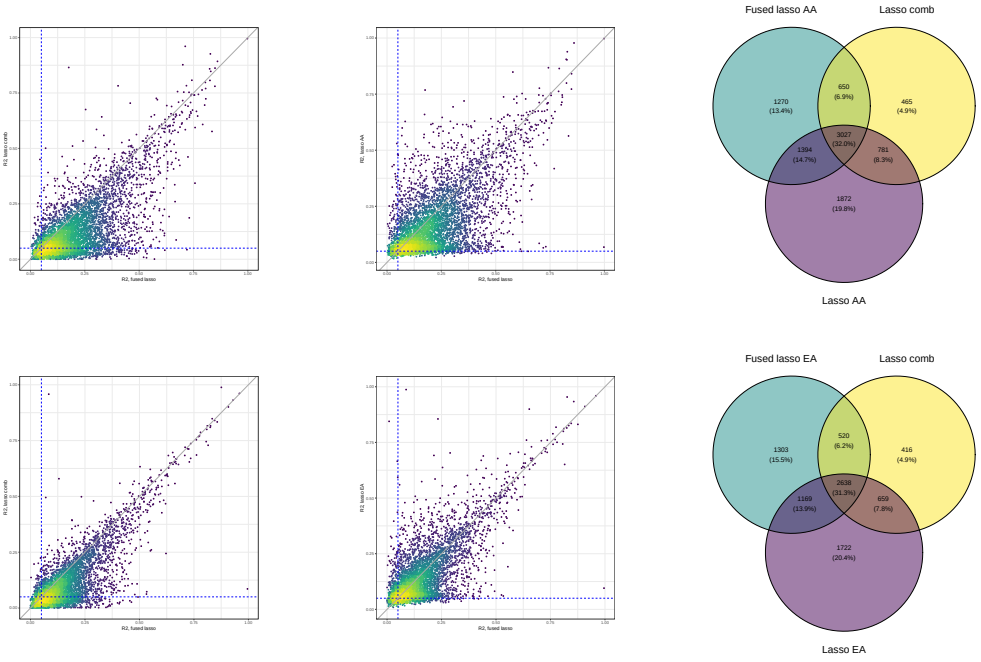
Figure B.3: **Simulation results comparing fused lasso vs standard lasso: same location of eQTLs, different direction of effects.**



(a) genetic architecture

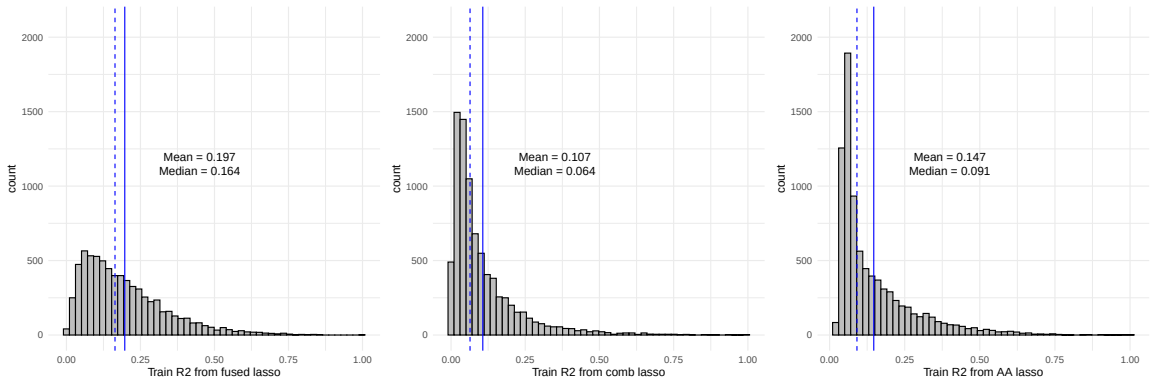
(b) simulation results

Figure B.4: **Simulation results comparing fused lasso vs standard lasso: different location of eQTLs, different direction of effects.**

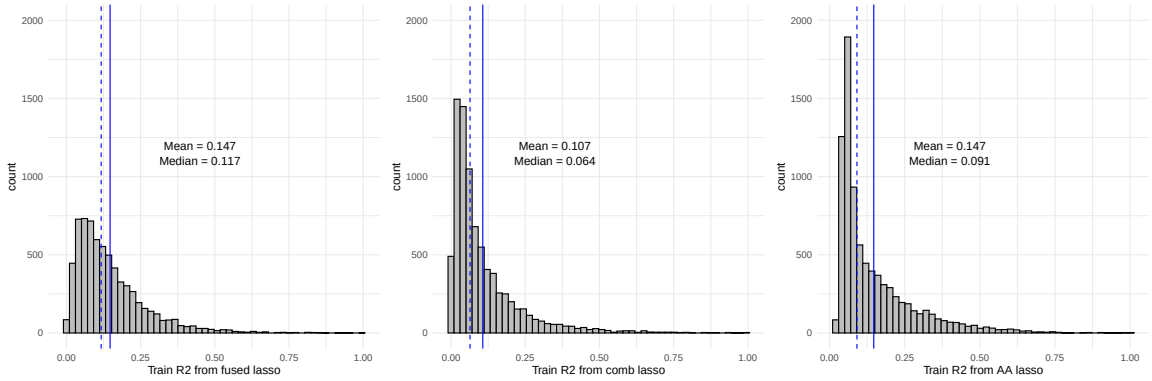


(a) (b) (c)

Figure B.5: **Performance of fused lasso and standard lasso models in training sets, based on model R^2 of all genes.** Smooth scatter plots of R^2 values between measured gene expression and predicted transcriptome values in MESA AA (top panel) and MESA EA (bottom panel) training sets. (a) Comparison of model R^2 values using lasso comb vs fused lasso models; (b) Comparison of model R^2 values using ancestry-specific lasso vs fused lasso models; (c) Venn diagram showing the overlap of well-predicted genes (with $R^2 > 0.05$) by fused lasso and standard lasso models in training sets.

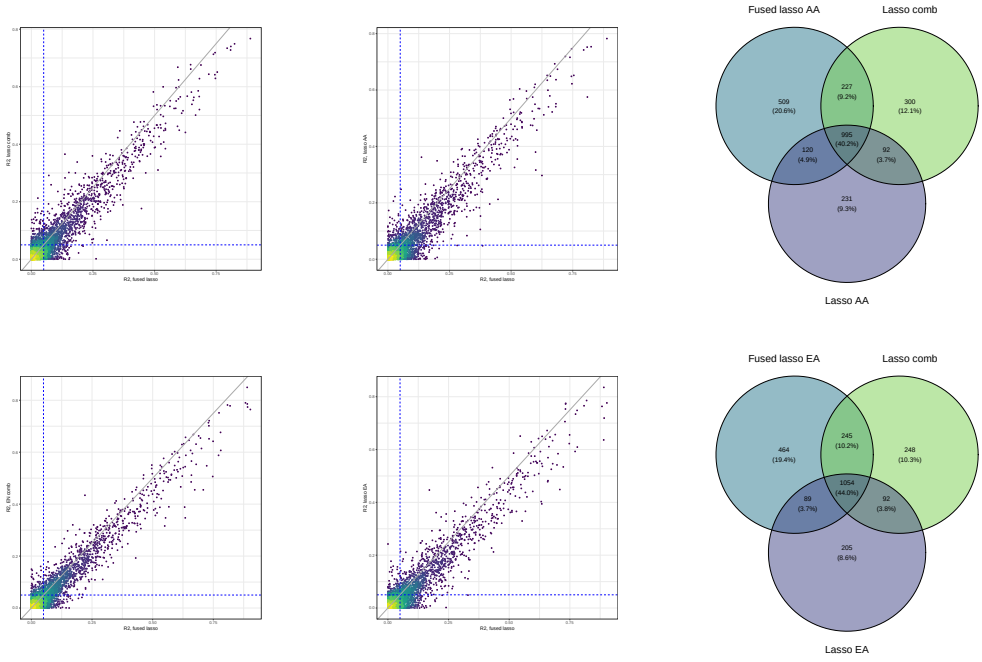


(a) MESA AA training set



(b) MESA EA training set

Figure B.6: **Histograms of the model R^2 distribution of all genes comparing fused lasso and standard lasso.** Histograms showing model R^2 for three methods in (a) African American and (b) European ancestry training sets. The blue solid line denotes the mean of model R^2; the blue dashed line denotes the median of model R^2. All genes, including those that fall below the $R^2 = 0.05$ cut-off, are displayed.



(a) (b) (c)

Figure B.7: **Performance of fused lasso and standard lasso models in test sets, based on true R^2 of all genes.** Smooth scatter plots of R^2 values between measured gene expression and predicted gene expression values in MESA AA (top panel) and MESA EA (bottom panel) test sets. (a) Comparison of true R^2 values using lasso comb vs fused lasso models; (b) Comparison of true R^2 values using ancestry-specific lasso vs fused lasso models; (c) Venn diagram of well-predicted ($R^2 > 0.05$) genes in test sets.

Appendix C

SUPPLEMENTARY METHODS AND FIGURES

C.0.1 Participating studies

Amish

The Amish Complex Disease Research Program includes a set of large community-based studies focused largely on cardiometabolic health carried out in the Old Order Amish (OOA) community of Lancaster, Pennsylvania (http://medschool.umaryland.edu/endocrinology/amish/research-program.asp). The OOA population of Lancaster County, PA immigrated to the Colonies from Western Europe in the early 1700s. There are now over 30,000 OOA individuals in the Lancaster area, nearly all of whom can trace their ancestry back 12-14 generations to approximately 700 founders. Investigators at the University of Maryland School of Medicine have been studying the genetic determinants of cardiometabolic health in this population since 1993. To date, over 7,000 Amish adults have participated in one or more of our studies.

Due to their ancestral history, the OOA are enriched for rare exonic variants that arose in the population from a single founder (or small number of founders) and propagated through genetic drift. Many of these variants have large effect sizes and identifying them can lead to new biological insights about health and disease. The parent study for this WGS project provides one (of multiple) examples. In our parent study, we identified through a genome-wide association analysis a haplotype that was highly enriched in the OOA that is associated with very high LDL-cholesterol levels. At the present time, the identity of the causative SNP and even the implicated gene is not known because the associated haplotype contains numerous genes, none of which

are obvious lipid candidate genes. A major goal of the WGS that will be obtained through the NHLBI TOPMed Consortium will be to identify functional variants that underlie some of the large effect associations observed in this unique population.

ARIC

The ARIC study is a population-based cohort study consisting of 15,792 men and women that were drawn from four U.S. communities (Suburban Minneapolis, Minnesota; Washington County, Maryland; Forsyth County, North Carolina, and Jackson, Mississippi) 1. It was designed to investigate the causes of atherosclerosis and its clinical outcomes, and variation in cardiovascular risk factors, medical care, and disease by race, sex, location, and date. Participants were between age 45 and 64 years at their baseline examination in 1987-1989 when blood was drawn for DNA extraction and participants consented to genetic testing.

BioMe

The Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center (MSMC), BioMe Biobank, founded in September 2007, is an ongoing, broadly-consented electronic health record-linked clinical care biobank that enrolls participants non-selectively from the Mount Sinai Medical Center patient population. The MSMC serves diverse local communities of upper Manhattan, including Central Harlem (86% African American), East Harlem (88% Hispanic/Latino), and Upper East Side (88% Caucasian/White) with broad health disparities.

CARDIA

The Coronary Artery Risk Development in Young Adults (CARDIA) Study is a study examining the development and determinants of clinical and subclinical cardiovascular disease and their risk factors. It began in 1985-1986 with a group of 5,115 black and

white men and women aged 18-30 years. The participants were selected so that there would be approximately the same number of people in subgroups of race, gender, education (high school or less and more than high school) and age (18-24 and 25-30) in each of 4 centers: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA.

CHS

The Cardiovascular Health Study (CHS) is a population-based cohort study of risk factors for coronary heart disease and stroke in adults 65 years and older conducted across four field centers 2. The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of people on Medicare eligibility lists from four US communities. Subsequently, an additional predominantly African-American cohort of 687 persons was enrolled for a total sample of 5,888. Institutional review committees at each field center approved the CHS, and participants gave informed consent. Blood samples were drawn from all participants at their baseline examination, and DNA was subsequently extracted from available samples. These analyses were limited to participants with available DNA who also consented to genetic studies. Participants were examined annually from enrollment to 1999 and continued to be under surveillance for stroke following 1999.

COPDGene

COPDGene (also known as the Genetic Epidemiology of COPD Study) is an NIH-funded, multicenter study. A study population of more than 10,000 smokers (1/3 African American and 2/3 non-Hispanic White) has been characterized with a study protocol including pulmonary function tests, chest CT scans, six minute walk testing, and multiple questionnaires. Five years after this initial visit, all available study participants are being brought back for a follow-up visit with a similar study protocol. This study has been used for epidemiologic and genetic studies. Previous genetic analysis in

this study has been based on genome-wide SNP genotyping data. Approximately 1,900 subjects underwent whole genome sequencing in this NHLBI WGS project, including severe COPD subjects and resistant smoking controls. The COPDGene Study web site is: http://www.copdgene.org/.

FHS

FHS is a three-generation, single-site, community-based, ongoing cohort study that was initiated in 1948 to investigate prospectively the risk factors for CVD including stroke. It now comprises 3 generations of participants: the Original cohort followed since 1948; their Offspring and spouses of the Offspring, followed since 1971; and children from the largest Offspring families enrolled in 2002 (Gen 3). The Original cohort enrolled 5,209 men and women who comprised two-thirds of the adult population then residing in Framingham, MA. Survivors continue to receive biennial examinations. The Offspring cohort comprises 5,124 persons (including 3,514 biological offspring) who have been examined approximately once every 4 years. The Gen 3 cohort contains 4,095 participants.

GeneSTAR

In 1982 The Johns Hopkins Sibling and Family Heart Study was created to study patterns of coronary heart disease and related risk factors in families with early-onset coronary disease, identified from10 Baltimore area Hospitals. GeneSTAR continues to study mechanisms of coronary heart disease and stroke in families using novel models and exciting new methods. GeneSTAR is a family-based study in initially healthy brothers and sisters, and offspring of people with early-onset coronary disease, The goal is to discover and amplify mechanisms of stroke and coronary heart disease. Our African American and European American family cohort has undergone extensive screening, genetic testing, and follow-up for new cardiovascular disease, stroke, and other clinical events for 5 to 32 years.

HCHS/SOL

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a multi-center study of Hispanic/Latino populations with the goal of determining the role of acculturation in the prevalence and development of diseases, and to identify other traits that impact Hispanic/Latino health. The study is sponsored by the National Heart, Lung, and Blood Institute (NHLBI) and other institutes, centers, and offices of the National Institutes of Health (NIH). Recruitment began in 2006 with a target population of 16,000 persons of Cuban, Puerto Rican, Dominican, Mexican or Central/South American origin. Participants were recruited through four sites affiliated with San Diego State University, Northwestern University in Chicago, Albert Einstein College of Medicine in Bronx, New York, and the University of Miami. Recruitment was implemented through a two-stage area household probability design. The study enrolled 16,415 participants who were self-identified Hispanic/Latino and aged 18-74 years and the extensive psycho-social and clinical assessments were conducted during 2008-2011. Annual telephone follow-up interviews are ongoing since study inception. During the 2014-2017 second visit, the participants were re-examined again of various health outcomes of interest.

JHS

The Jackson Heart Study (JHS, https://www.jacksonheartstudy.org/jhsinfo/) is a large, community-based, observational study whose participants were recruited from urban and rural areas of the three counties (Hinds, Madison and Rankin) that make up the Jackson, MS metropolitan statistical area (MSA). Participants were enrolled from each of 4 recruitment pools: random, 17%; volunteer, 30%; currently enrolled in the Atherosclerosis Risk in Communities (ARIC) Study, 31% and secondary family members, 22%. Recruitment was limited to non-institutionalized adult African Americans 35-84 years old, except in a nested family cohort where those 21 to 34

years of age were also eligible. The final cohort of 5,301 participants included 6.59% of all African American Jackson MSA residents aged 35-84 during the baseline exam (N=76,426, US Census 2000). Among these, approximately 3,700 gave consent that allows genetic research and deposition of data into dbGaP. Major components of three clinic examinations (Exam 1 2000-2004; Exam 2 2005-2008; Exam 3 2009-2013) include medical history, physical examination, blood/urine analytes and interview questions on areas such as: physical activity; stress, coping and spirituality; racism and discrimination; socioeconomic position; and access to health care. Extensive clinical phenotyping includes anthropometrics, electrocardiography, carotid ultrasound, ankle-brachial blood pressure index, echocardiography, CT chest and abdomen for coronary and aortic calcification, liver fat, and subcutaneous and visceral fat measurement, and cardiac MRI. At 12-month intervals after the baseline clinic visit (Exam 1), participants have been contacted by telephone to: update information; confirm vital statistics; document interim medical events, hospitalizations, and functional status; and obtain additional sociocultural information. Questions about medical events, symptoms of cardiovascular disease and functional status are repeated annually. Ongoing cohort surveillance includes abstraction of medical records and death certificates for relevant International Classification of Diseases (ICD) codes and adjudication of nonfatal events and deaths. CMS data are currently being incorporated into the dataset.

MESA

The MESA study is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease. MESA researchers study a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84. Thirty-eight percent of the recruited participants are white, 28 percent African-American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants

were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and the University of California - Los Angeles.

SAFS

The San Antonio Family Study (SAFS) is a complex pedigree-based mixed longitudinal study designed to identify low frequency or rare variants influencing susceptibility to cardiovascular disease, using WGS information from 2,590 individuals in large Mexican American pedigrees from San Antonio, Texas. The major objectives of this study are to identify low frequency or rare variants in and around known common variant signals for CVD, as well as to find novel low frequency or rare variants influencing susceptibility to CVD.

WHI

The Womens Health Initiative (WHI) is a long-term, prospective, multi-center cohort study that investigates post-menopausal womens health. WHI was funded by the National Institutes of Health and the National Heart, Lung, and Blood Institute to study strategies to prevent heart disease, breast cancer, colon cancer, and osteoporotic fractures in women 50-79 years of age. WHI involves 161,808 women recruited between 1993 and 1998 at 40 centers across the US. The study consists of two parts: the WHI Clinical Trial which was a randomized clinical trial of hormone therapy, dietary modification, and calcium/Vitamin D supplementation, and the WHI Observational Study, which focused on many of the inequities in womens health research and provided practical information about the incidence, risk factors, and interventions related to heart disease, cancer, and osteoporotic fractures. For TOPMed WGS, the study over-sampled participants with incident stroke and VTE. The remaining samples were age- and ethnicity-matched controls without stroke or VTE.

C.0.2 Genetic ancestry and relatedness

Principal components (PCs) of genetic ancestry and pairwise relatedness measures were estimated for all 140,062 samples included in the TOPMed freeze 8 genotype release. Autosomal genetic variants passing the quality filter with a MAF > 0.01 and missing call rate < 0.01 were LD-pruned with an r^2 threshold of 0.1 to obtain a set of 638,486 effectively independent variants for genetic ancestry and relatedness estimation. PC-AiR was used to obtain ancestry informative PCs robust to familial relatedness; the first 11 PCs showed evidence of population structure. PC-Relate was then used to estimate pairwise kinship coefficients (KCs) for all pairs of samples, conditional on the genetic ancestry captured by PC-AiR PCs 1-11; these KC estimates reflect only recent genetic relatedness, e.g. due to pedigree structure. The PC-Relate KC estimates were used to construct a 4th degree sparse, block-diagonal, empirical kinship matrix (KM) for association testing, using the procedure recommended in Gogarten et al: any pair of samples with estimated KC $> 2^{(-11/2)} \sim 0.022$ were clustered in the same block; all KC estimates within a block of samples were kept, regardless of value; and all KC estimates between blocks were set to 0. By using a sparse block-diagonal KM, the association tests are more computationally efficient yet recent genetic relatedness is still accounted for. We subset the freeze-wide PCs and sparse KM to the appropriate set of participants for each analysis.

C.0.3 Race imputation using HARE

Ancestry groups were based on a combination of participants reported race/ethnicity and genetic ancestry represented by PCs from PC-AiR. To infer race/population group membership for participants with missing values, we used the HARE method, a machine learning algorithm that uses a support vector machine (SVM) to determine stratum assignment, taking as input genetically estimated PC values and reported race/ethnicity for each participant. Strata are defined by the unique reported

race/ethnicity values provided, then the HARE SVM uses the input (training) data to learn the probability of stratum membership across the entire PC space. The output of HARE consists of multinomial probability vectors of stratum membership for each participant. HARE was run on a subset of samples included in the TOPMed freeze 8 genotype release; specifically, samples for participants from non-US-based studies and the Amish participants (because they were very distinct in PC space) were excluded from the HARE analysis. HARE was run using the first 9 PC-AiR PCs generated on this subset of samples to represent genetic ancestry with the following reported race/population groups: Asian, Black, Central American, Cuban, Dominican, Mexican, Puerto Rican, South American, and White. The genetic data from the 31,918 participants with either unreported or non-specific (e.g. Multiple or Other) race and population membership was included in the HARE analysis, but they were not used to train the SVM. These participants were assigned to a population stratum based on their highest HARE output probability of membership. All other participants remained in the population stratum corresponding to their reported race/population group. Amish participants were assigned to their own stratum.

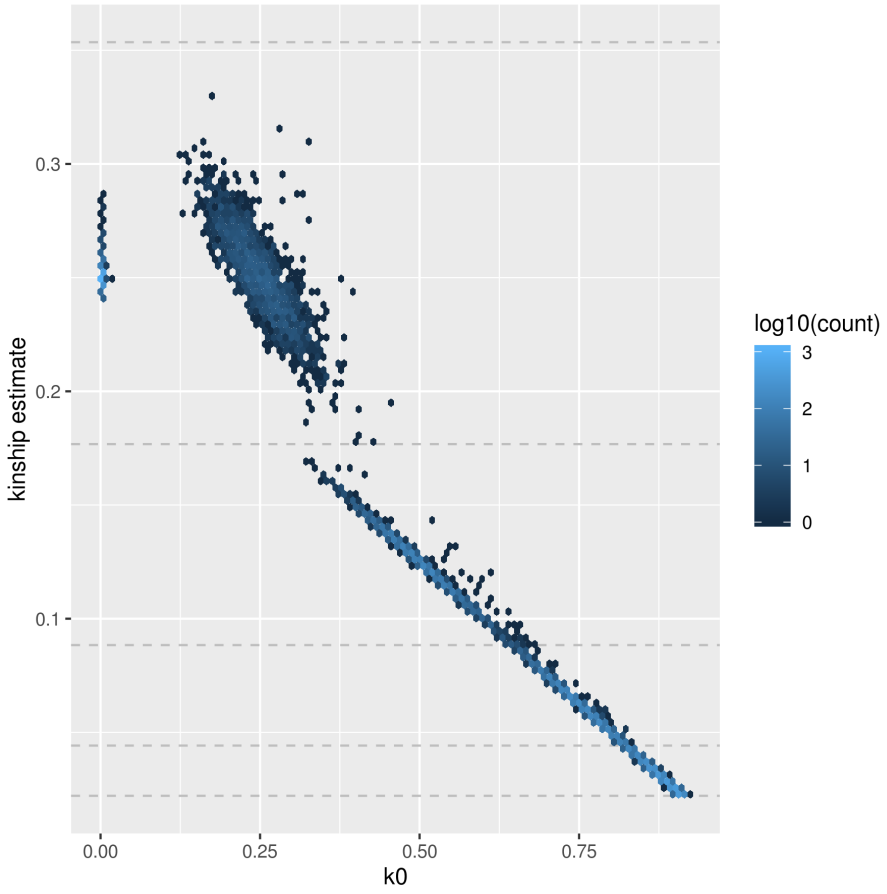


Figure C.1: **Relationship estimation in 6 studies from TOPMed freeze8.** Scatter plots of the estimated kinship coefficients against the estimated probabilities of sharing zero alleles IBD, $k^{(0)}$ and of the estimated probabilities of sharing two alleles IBD, $k^{(2)}$, from PC-Relate.

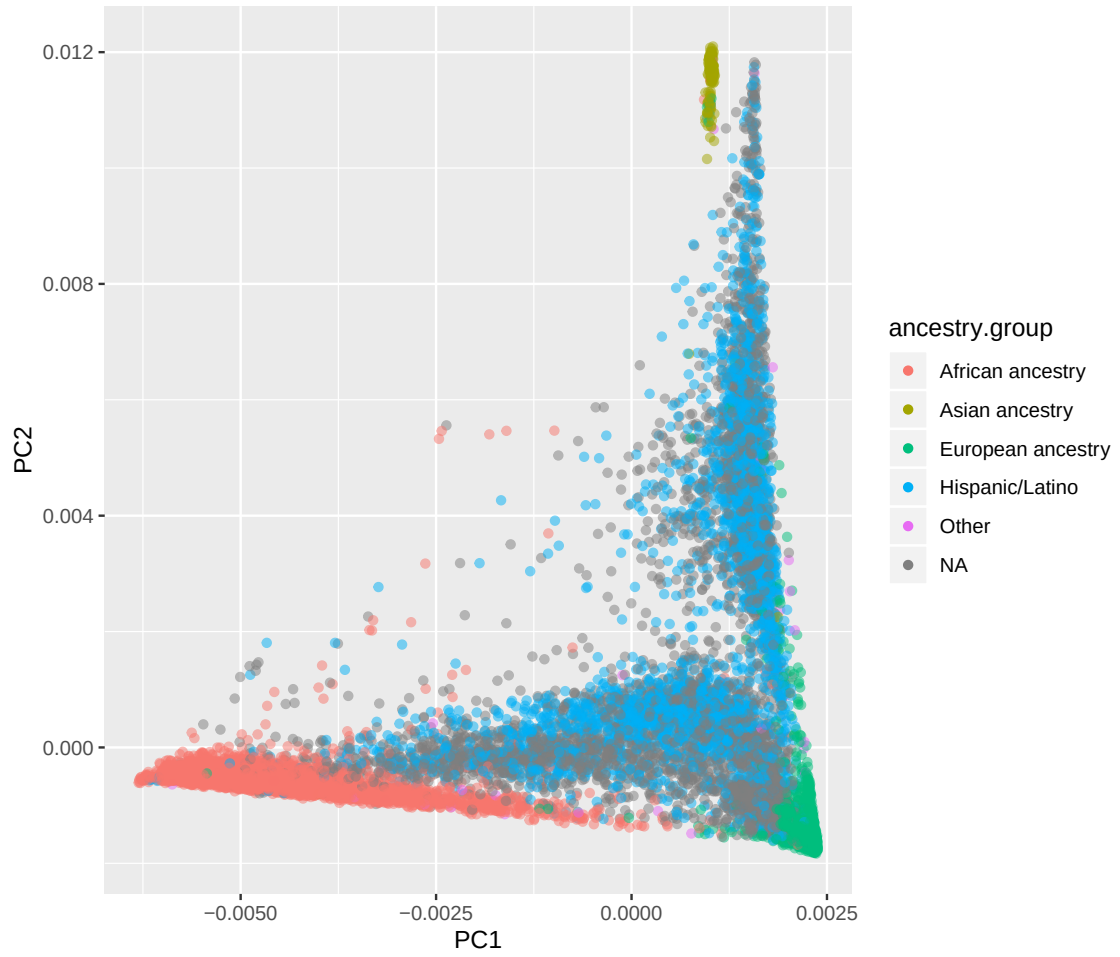


Figure C.2: **Population structure inference for 6 studies from TOPMed freeze8.** Scatter plots of the top two principal components from PC-AiR. The color of each point represents that individual's self-reported ancestry.

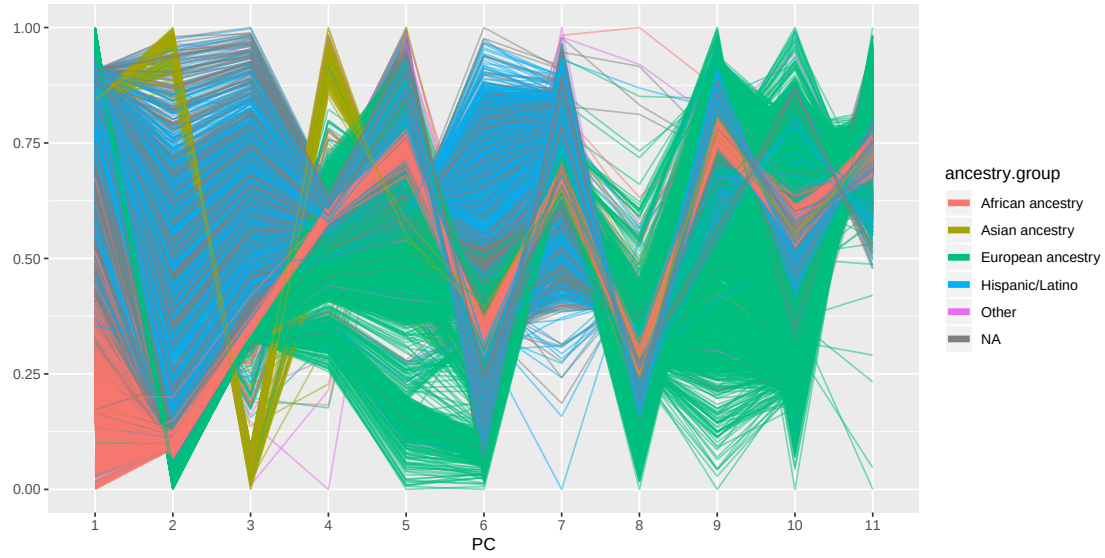


Figure C.3: **Parallel coordinates plots 6 studies from TOPMed freeze8.** Parallel coordinates plots of the top eleven principal components from PC-AiR. Each vertical bar represents one of the eigenvectors, and each line traces out the coordinates for an individuals across all eleven eigenvectors. The color of each line represents that individual's self-reported ancestry.

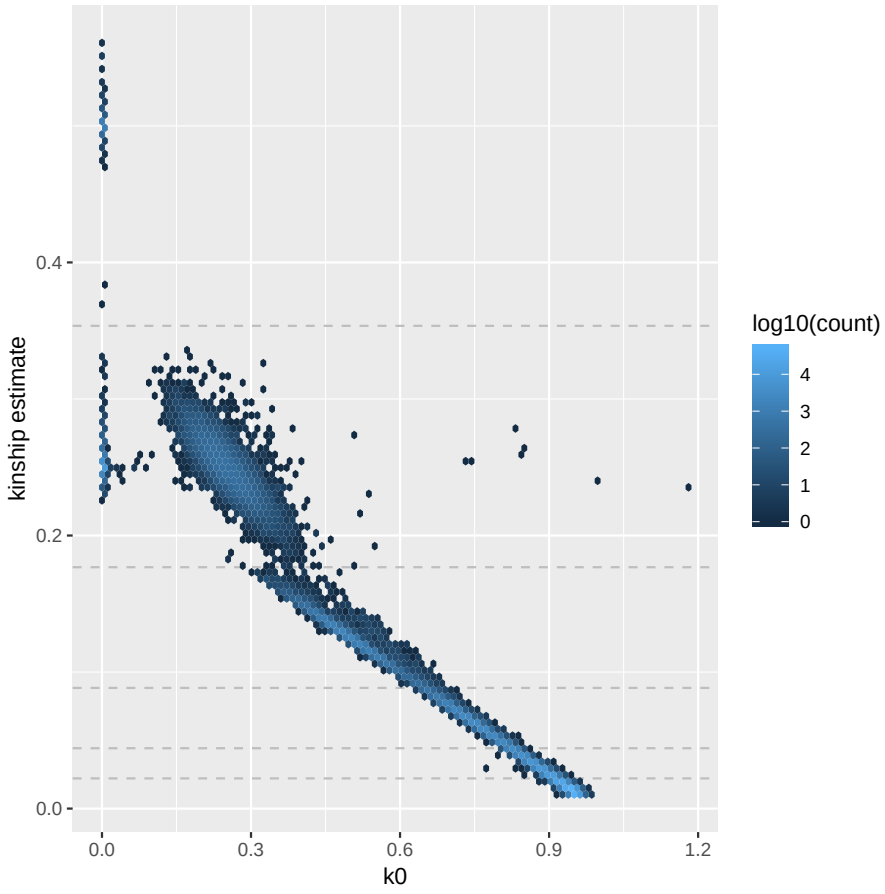


Figure C.4: **Relationship estimation for all individuals from TOPMed freeze8.** Scatter plots of the estimated kinship coefficients against the estimated probabilities of sharing zero alleles IBD, $k^{(0)}$ and of the estimated probabilities of sharing two alleles IBD, $k^{(2)}$, from PC-Relate.

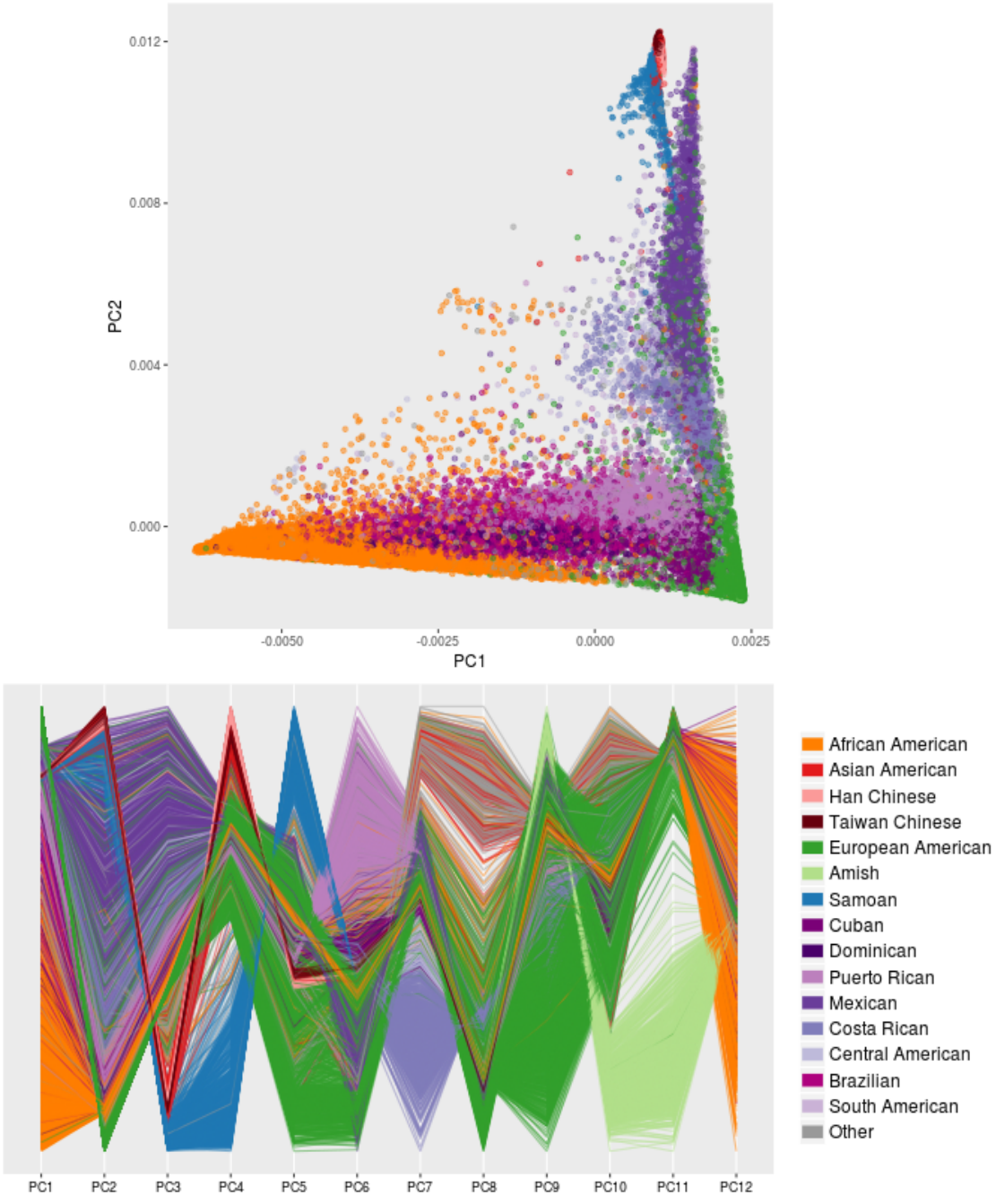


Figure C.5: **Population structure inference and parallel coordinates plots for all individuals from TOPMed freeze8.** Top: Scatter plots of the top two principal components from PC-AiR. The color of each point represents that individual's self-reported ancestry. Bottom: Parallel coordinates plots of the top eleven principal components from PC-AiR. Each vertical bar represents one of the eigenvectors, and each line traces out the coordinates for an individuals across all eleven eigenvectors. The color of each line represents that individual's self-reported ancestry.