2023

Incorporating Sex Chromosomes in Transcriptome Prediction Models and Improving Cross-Population Prediction Performance

Daniel S. Araujo

LOYOLA UNIVERSITY CHICAGO

INCORPORATING SEX CHROMOSOMES IN TRANSCRIPTOME PREDICTION MODELS AND IMPROVING CROSS-POPULATION PREDICTION PERFORMANCE

A THESIS SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

MASTER OF SCIENCE

PROGRAM IN BIOINFORMATICS

BY

DANIEL S. ARAUJO

CHICAGO, IL

MAY 2023

ACKNOWLEDGMENTS

First and foremost, I would like to thank Dr. Heather Wheeler for giving me the chance to join her research group, as well as for her incredible mentorship over the years. I am beyond grateful for all the learning opportunities given to me throughout the years, and for the space and resources provided to me so I could learn more about a subject I became deeply passionate about. Additionally, I would also like to thanks all members of the Wheeler Lab who have directly or indirectly helped me during my thesis project. More specifically, I wish to thank Henry Wittich, who welcomed me to Chicago and helped me to adapt to a new chapter of my life, and with whom I was capable to develop a strong friendship outside of the lab. Furthermore, I would like to acknowledge the support of the Bioinformatics Program for granting me the Graduate Research Fellowship throughout my time at Loyola.

Lastly, I would like to thanks my family and friends back in Brazil who believed in me and helped me realize I was ready to take an opportunity to study abroad. In particular, I would like to thank my parents for all the unconditional support throughout my whole life, which made it possible for me to pursue my own dreams and goals. Everything I have ever accomplished and will ever do is a consequence of their encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFA	African American population in MESA
ALL	All individuals of Geuvadis combined
CEU	Utah residents with Northern and Western European ancestry in Geuvadis
CHN	Chinese population in MESA
EN	Elastic Net
eQTL	Expression quantitative trait locus
EUR	European population in MESA
FIN	Finnish in Finland in Geuvadis
GBR	British in England and Scotland in Geuvadis
GTEx	Genotype-Tissue Expression Project
GWAS	Genome-Wide Association Study
HIS	Hispanic population in MESA
HWE	Hardy-Weinberg Equilibrium
INDEL	Insertion or deletion of nucleotides
LD	Linkage-disequilibrium
MAF	Minor allele frequency
MASHR	Multivariate adaptive shrinkage in R
MB	Megabase (1,000,000 base pairs)
MESA	Multi-Ethnic Study of Atherosclerosis
PAGE	Population Architecture using Genomics and Epidemiology study
PanUKBB	Pan-ancestry genetic analysis of the UK Biobank

SNP	Single nucleotide polymorphism
TOPMed	NHLBI Trans-Omics for Precision Medicine consortium
TPM	Transcripts per million
TSI	Toscani in Italy in Geuvadis
TWAS	Transcriptome-Wide Association Study
YRI	Yoruba in Ibadan, Nigeria in Geuvadis

ABSTRACT

Transcriptome prediction models built with data from European-descent individuals are less accurate when applied to different populations because of differences in linkage disequilibrium patterns and allele frequencies. We hypothesized multivariate adaptive shrinkage may improve cross-population transcriptome prediction, as it leverages effect size estimates across different conditions - in this case, different populations. To test this hypothesis, we made transcriptome prediction models for use in transcriptome-wide association studies (TWAS) using different methods (Elastic Net, Matrix eQTL and Multivariate Adaptive Shrinkage in R (MASHR)) and tested their transcriptome prediction accuracy in population-matched and cross-population scenarios. Additionally, to evaluate model applicability in TWAS, we integrated publicly available multi-ancestry genome-wide association study (GWAS) summary statistics from the Population Architecture using Genomics and Epidemiology Study (PAGE) and Pan-UK Biobank with our developed transcriptome prediction models. In regard to transcriptome prediction accuracy, MASHR models had similar performance to other methods when the training population ancestry closely matched the test population, but outperformed other methods in cross-population predictions. Furthermore, in multi-ancestry TWAS, MASHR models yielded more discoveries that replicate in both PAGE and PanUKBB across all methods analyzed. Overall, we demonstrate the importance of using methods that incorporate effect size estimates from multiple populations in order to improve TWAS for multi-ancestry or underrepresented populations.

INTRODUCTION

Human Genetic Variation

All current living humans share an ancestral population that lived in Africa approximately 200,000 years ago (Cann, Stoneking, and Wilson 1987). Members of that ancestral population and their descendants migrated to other regions of the globe, which is known nowadays as the “out-of-Africa” dispersal (Klein 2008). However, there is a debate regarding where modern humans originated within the African continent. Some authors defend a model of multiregionalism that postulates that distinct populations co-evolved, with multiple gene flow events due to migration, while others believe that one population expanded its range and predominated over others, with possibility of regional admixtures (Henn, Steele, and Weaver 2018). Still, the theory that modern humans originated in Africa and migrated out of the continent, known as recent African origin, has widespread acceptance (Stringer 2014).

However, although humans present a wide range of distinct phenotypes, genetically, we are extremely similar. In fact, it is estimated that on average, two individuals share 99.9% of their DNA (Fine, Ibrahim, and Thomas 2005). This is because humans are more related to each other than one might think. An individual’s genealogical tree grows exponentially for every generation, which eventually would become larger than the number of all humans who have ever lived (Derrida, Manrubia, and Zanette 2000). This issue is solved if taken into account that genealogical trees coalesce and collapse on themselves, with many ancestors being relatives and occupying multiple positions (Derrida, Manrubia, and Zanette 2000).

Consequently, is it possible to estimate a point in time in which all current living humans share every ancestor in common, known as the genetic isopoint. According to previous studies, the global genetic isopoint is fairly recent, occurring 3,400 years ago (Rutherford 2020).

The 0.1% of differences in the DNA sequence among individuals, also known as DNA polymorphisms, can have different causes, and depending on different factors, may be lost or fixed in a population. The migration out of Africa is a recent event in human history, and thus, most of the natural history of humans occurred in Africa. Consequently, African populations harbor the highest levels of genetic diversity in the world (Tishkoff and Williams 2002). Such discrepancies between African and non-African populations are due to the multiple population bottleneck events that happened throughout human history (Campbell and Tishkoff 2008). As individuals migrated to different regions, they carried only a fraction of the genetic diversity of the original population.

Ever since the early 2000s, when the human genome was finally sequenced and sequencing costs became progressively cheaper, many efforts to sample the genetic diversity across human populations have been made. One such example is the 1000 Genomes Project Consortium, which initially sequenced and analyzed the genomes of 1,092 individuals from 14 populations (The 1000 Genomes Project Consortium 2012). The project showed that most common human genetic variations are found almost in every population, but rare variants (frequency < 1%) tend to be population-specific. Furthermore, most of the rarer DNA polymorphisms were found in current African populations, which is in agreement to the out-of-African human dispersal theory, as those groups did not go through major genetic bottleneck events. The 1000 Genomes Project was later incorporated into the International Genome Sample Resource (IGSR), and its numbers have greatly increased over the years, reaching 2,706 samples across 26 populations (Clarke et al. 2017). Besides data from the 1000 Genomes Project, IGSR also contains data from other initiatives, such as The Gambian Genome Variation Project, Simons Diversity Project, and Human Genome Diversity Project.

Genome-wide Association Studies

DNA polymorphisms can be of different types, such as variation in the number of copies of tandem repeats, insertions or deletions of multiple nucleotides, or single nucleotide differences, also known as single nucleotide polymorphisms (SNPs). More specifically, SNPs happen in cases in which at the same base pair position in the genome, two or more nucleotides are found when comparing the genome of distinct people. These different “versions” are called alleles. In comparison to other polymorphisms, SNPs are more common, stable, and dispersed throughout the genome (Shastry 2002). However, although very simple (it is only a single nucleotide change), SNPs have subtypes. These variants can be found outside of genes, in what is known as non-coding regions of the genome, or inside of genes. When inside of genes, SNPs can either be intronic or exonic – inside introns or exons, respectively. Lastly, if exonic, SNPs are synonymous mutations if the encoded amino acid remains the same, nonsynonymous mutations if the encoded amino acid is not the same as the original, or nonsense if the mutation changes the codon to a stop codon (Shastry 2002).

In situations in which SNPs are found inside of genes, it is fairly simple to test the association of changes in the DNA sequence to gene function. However, only a small fraction of the genome is comprised of protein-encoding genes (International Human Genome Sequencing Consortium 2004). Thus, different tools to understand the influence of SNPs have been developed. One example is genome-wide association studies (GWAS), in which millions of SNPs are tested for associations with a trait. The phenotypes tested for association can be either discrete, in which human subjects are split into control and case groups, or continuous. The reasoning behind GWAS is that if a SNP has a higher frequency in the case group, for instance, then it is likely that that SNP is associated to the phenotype of interest (Cano-Gamez and Trynka 2020). However, GWAS are heavily influenced by different factors, such as sample size, allele frequencies, linkage disequilibrium patterns, and heritability of the investigated trait (Visscher et al. 2017).

Most phenotypes investigated in GWAS are complex, meaning that they may be influenced by many SNPs, such as heart diseases or psychiatric disorders (Visscher et al. 2012). Those SNPs, in turn, occur at different rates, which are known as allele frequencies. Common SNPs (minor allele frequencies [MAF] higher than 1%) are usually well studied, as they do not require the same degree of statistical power conferred by larger sample sizes in comparison to rarer SNPs, especially those with frequency less than 1% (Altshuler et al. 2010). This is one example of why sample size is an important factor for GWAS, and the reason GWAS are usually done in larger consortiums, such as the Trans-Omics for Precision Medicine Program, the Million Veteran Program, and the Global Lipids Genetic Consortium (Uffelmann et al. 2021; Taliun et al. 2021; Gaziano et al. 2016; Willer et al. 2013). However, it is crucial to recognize that simply because a SNP is associated to a phenotype, it does not mean that it is causal. Causal inferences are hard to confirm due to linkage disequilibrium between SNPs, that is, the non-independent correlation between two physically close SNPs in the DNA (Uffelmann et al. 2021). Consequently, in a typical GWAS, usually it is observed that groups of physically close SNPs are associated to a phenotype, with perhaps just one member of the group being the true causal SNP (Dandine-Roulland and Perdry 2015). Attempts to distinguish the true causal SNP among groups of linked SNP are refereed as fine-mapping, and different statistical methods can be applied, such as using marginal association statistics or posterior probabilities (Kichaev et al. 2014).

Additionally, many complex traits, such as human diseases, are influenced by a combination of genetic and environmental factors (Chakravarti and Little 2003). The interplay between genetics and environment is not necessarily evenly split, as some traits have a bigger influence of genetic or environmental factors. The proportion of variance observed in a phenotype that is explained by genetic factors is deemed heritability (often represented by h^2), and can range from 0 to 1 – therefore, the bigger h^2 is, the stronger the correlation between the phenotype and genetics (Visscher, Hill, and Wray 2008). Thus, as GWAS test SNPs for association with a trait, it is important to be mindful that they can only be used to understand the genetically influenced component of a phenotype.

Transcriptome-wide Association Studies

As aforementioned, most SNPs are found outside of genes, which makes the biological interpretation difficult for most associations. Often, most authors will assign phenotype-associated SNPs to the nearest gene in the genome, which may not reflect true biological meanings (Petersen et al. 2013). Thus, different approaches have been designed to understand the link between genes and phenotypes of interest, such as PrediXcan and FUSION, which perform transcriptome-wide association studies (TWAS) (Gamazon et al. 2015; Mancuso et al. 2018).

In a GWAS, SNPs are tested for association with a trait. In a TWAS, genetically predicted gene expression levels (RNA levels) are tested for association for a trait. To achieve this, TWAS rely on gene expression prediction models built using expression quantitative trait loci (eQTLs), which are SNPs associated to the expression of certain genes (Nica and Dermitzakis 2013). eQTLs can be classified as *cis*-acting if they influence the expression of a nearby gene, or *trans*-acting eQTLs if they play a part in modulating the expression of a gene far away, such as on a different chromosome, although they tend to have smaller effect sizes and thus larger samples are needed to detect them (Westra and Franke 2014). As *cis*-eQTLs are easier to detect and have higher effect sizes in comparison to *trans*-eQTLs, gene expression prediction models rely on them. Thus, using *cis*-eQTLs, gene expression is estimated by performing a weighted-sum of allele dosages, as shown in the following equation, in which $\hat{Y}$ is the estimated gene expression for a particular gene, *n* is the number of *cis*-eQTLs in the model, W_k is the effect size for SNP *k*, and X_k is the dosage of the SNP *k*.

$$\hat{Y} = \sum_{k=1}^{n} W_k X_k$$

In contrast with GWAS, as TWAS test RNA levels for association, they provide information about which genes are up- or down-regulated in regards to a phenotype of interest (Barbeira et al. 2019). This helps to pinpoint specific genes that might be the target for possible therapeutic drugs (Mulford et al. 2021). Furthermore, gene expression prediction models for use in TWAS tend to be tissue-specific. Unlike DNA, which is virtually the same in all cells in the body with the exception of somatic mutations that occur throughout an individual’s life, RNA levels can naturally differ between tissues (Zhu et al. 2016). In agreement to that, studies have found tissue-specific eQTLs, although most *cis*-eQTLs are shared between tissues (Aguet et al. 2017; Kirsten et al. 2015). However, some eQTLs may have opposite direction of effect in different tissues (Mizuno and Okada 2019). One major scientific effort made to help identify eQTLs across different human tissues is the Genotype-Tissue Expression (GTEx) project, which has collected and analyzed samples from over 40 different tissues in order to investigate gene expression profiles across all of them (The GTEx Consortium et al. 2015). Moreover, one limitation of gene expression prediction models is that they only account for the expression influenced by genetic factors (Wainberg et al.

2019). Similarly to other complex traits, gene expression can also be affected by environmental factors (Gibson 2008). Consequently, gene expression prediction models have higher accuracy when estimating the expression of genes whose expression is highly heritable (Li et al. 2018).

Underrepresentation in Association Studies

Over the years, due to the wide applicability and popularity of GWAS, it became necessary to gather all generated results in a single platform to facilitate access to them. Thus, the National Human Genome Research Institute (NHGRI) and the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) joined efforts to create the NHGRI-EBI GWAS Catalog (Welter et al. 2014; MacArthur et al. 2017). As of July 2022, the GWAS Catalog contained information about more than 400,000 SNP-phenotype associations across over 6,000 publications (Sollis et al. 2023). However, by analyzing the data in the GWAS Catalog, it is possible to notice gaps within GWAS. For instance, the majority of GWAS focus only on the autosomal chromosomes, ignoring the genetic content of the X chromosome (Kukurba et al. 2016). A survey done in 2013 analyzing the GWAS Catalog revealed that only 33% of studies included the X chromosome in their analysis (Wise, Gyi, and Manolio 2013). Ten years later, the scenario has not changed much. A recent study investigated GWAS summary statistics published in 2021, and found out that only 25% of them provided results for the X chromosome (Sun et al. 2023). The proportion of studies that provide results for Y chromosome is even lower – only 3% (Sun et al. 2023). This underrepresentation is likely due do the fact that analysis on the sex chromosomes require specialized methods due to different dosages between males and females (Wise, Gyi, and Manolio 2013). Consequently, most GWAS fail to acknowledge the relationship between the sex chromosomes and complex traits (Kukurba et al. 2016; Brumpton and Ferreira 2016).

Additionally, another gap found within studies in the GWAS Catalog is the population underrepresentation. As previously discussed, different human populations may have distinct allele frequencies and linkage disequilibrium patterns due to genetic bottleneck events as consequence of distinct migration patterns (Campbell and Tishkoff 2008). Although the average genetic difference between individuals is extremely small (around 0.1%), undersampling the genetic diversity that exists among populations can negatively impact the portability of association studies results, that is, associations found in one population may not happen in other populations (Martin et al. 2019). In 2009, 96% of all individuals in studies in the GWAS Catalog were of European ancestry (Popejoy and Fullerton 2016). Almost ten years later, in 2018, European-descent individuals comprised almost 80% of all individuals in the GWAS Catalog, even though they corresponded to 16% of the world's total population (Martin et al. 2019). Many efforts have been made to try to increase genetic diversity in human genetics with hopes to reduce health disparities among individuals of different ancestries, such as the NHLBI Trans-Omics for Precision Medicine consortium, the Human Heredity and Health in Africa initiative, and the All of Us Research program (Taliun et al. 2021; The H3Africa Consortium et al. 2014; The All of Us Research Program Investigators 2019). Note, although individuals that participate in association studies are often clustered into continental ancestries groups (*e.g.* African, Asian, European, etc.), genetic ancestry is actually multi-dimensional and continuous – depending on the timescale, individuals will have multiple ancestries (Lewis et al. 2022).

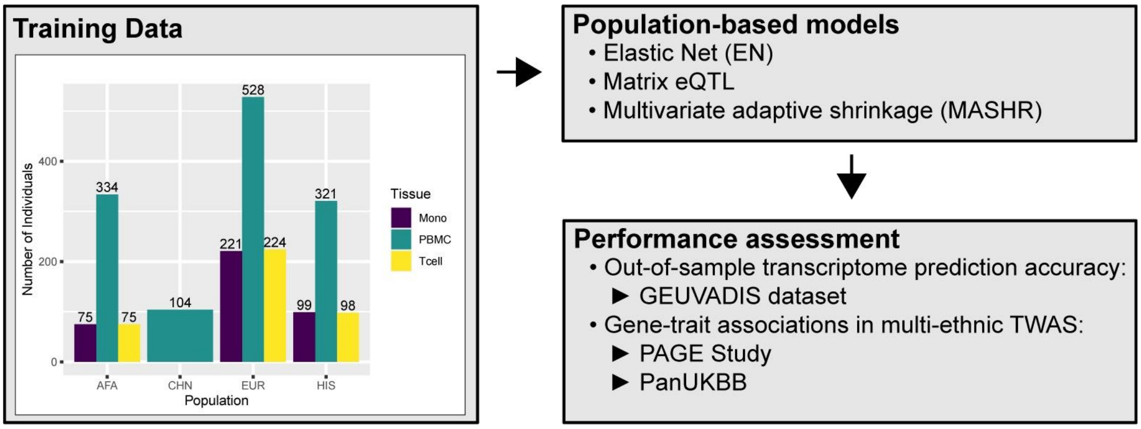
Moreover, similarly to GWAS, TWAS also suffer from the same underrepresentation. Gene expression prediction models are often trained using data from individuals of European descent due to the data availability bias, such as from GTEx, and as previous studies have shown, those prediction models have lower prediction accuracy when applied to non-

European population datasets (Keys et al. 2020; Mikhaylova and Thornton 2019; Mogil et al. 2018). In fact, TWAS have higher power for discovery and replication when the gene expression prediction model was trained in a cohort of similar ancestry of the test dataset (Geoffroy, Gregga, and Wheeler 2020). Similar results have been observed in protein-wide association studies as well (Schubert et al. 2022). Nevertheless, the biological mechanisms behind complex traits are expected to be conserved across all populations (Qiao et al. 2022). Thus, it is important to build gene expression prediction models that account for the allelic differences among populations and better estimate effect sizes to better understand the genetics of complex traits across all human populations (Geoffroy, Gregga, and Wheeler 2020).

Summary

In this thesis, we sought to develop gene expression prediction models with a higher cross-population prediction accuracy for use in multi-ethnic TWAS. As aforementioned, non-European genetic ancestry representation has been increasing in GWAS over the years, although it still is a small fraction (Martin et al. 2019). Likewise, gene expression prediction models for use in TWAS are often trained in European-descent individuals data and show poor cross-population prediction performance due to differences in allele frequencies, eQTL effect sizes and linkage-disequilibrium patterns between populations (Keys et al. 2020; Mikhaylova and Thornton 2019). Thus, as many scientific efforts have been trying to increase genetic data diversity in GWAS, it is important to develope new gene expression prediction models that will estimate gene expression levels across different populations with a higher accuracy than current methods.

For this, we used whole genome genotyping and RNA-sequencing data from the TOPMed Multi-Ethnic Study of Atherosclerosis (MESA), which includes X chromosome data, to build gene expression prediction models for TWAS (Bild et al. 2002). The training dataset contains data from three cell types (CD16+ monocytes, CD4+ T-cells, and peripheral blood mononuclear cells [PBMC]). Furthermore, each cell type dataset contains individuals of up to four distinct populations (African American [AFA], Chinese [CHN], European [EUR], or Hispanic/Latino [HIS) (Figure 1). To build population-specific gene expression prediction models, we used three distinct methods: elastic net, unadjusted Matrix eQTL, and multivariate adaptive shrinkage in R (MASHR) (Zou and Hastie 2005; Friedman, Hastie, and Tibshirani 2010; Shabalin 2012; Urbut et al. 2019). Later, we assessed population-matched and cross-population gene expression prediction performance using a test dataset that contained individuals of distinct continental ancestries. Lastly, we assessed the applicability of our models in a multi-ethnic TWAS, using data from two large multi-ethnic studies.



Population-based models
- Elastic Net (EN)
- Matrix eQTL
- Multivariate adaptive shrinkage (MASHR)

Performance assessment
- Out-of-sample transcriptome prediction accuracy:
  - ▶ GEUVADIS dataset
- Gene-trait associations in multi-ethnic TWAS:
  - ▶ PAGE Study
  - ▶ PanUKBB

Figure 1: Overall study methodology. Using TOPMed MESA as a training dataset, we built population-based transcriptome prediction models using three different methods (Elastic Net, Matrix eQTL, and Multivariate adaptive shrinkage). With these transcriptome models, we evaluated their out-of-sample transcriptome prediction accuracy using the GEUVADIS dataset. Additionally, we assessed their applicability in multi-ethnic TWAS using GWAS summary statistics from the PAGE Study and PanUKBB. AFA = African American, CHN = Chinese, EUR = European, HIS = Hispanic/Latino.

METHODS

Publication disclaimer

Part of this work is available as a preprint at bioRxiv (doi.org/10.1101/2023.02.09.527747) and is under review for publication with the following authors:

Daniel S. Araujo[1], Chris Nguyen[2], Xiaowei Hu[3], Anna V. Mikhaylova[4], Chris Gignoux[5], Kristin Ardlie[6], Kent D. Taylor[7], Peter Durda[8], Yongmei Liu[9], George Papanicolaou[10], Michael H. Cho[11], Stephen S. Rich[3], Jerome I. Rotter[7], NHLBI TOPMed Consortium, Hae Kyung Im[12], Ani Manichaikul[3], Heather E. Wheeler[1,2,*]

[1]Program in Bioinformatics, Loyola University Chicago, Chicago, IL, 60660, USA; [2]Department of Biology, Loyola University Chicago, Chicago, IL, 60660, USA; [3]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, 22908, USA; [4]Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA; [5]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, UC Denver Anschutz Medical Campus, Aurora, CO, 80045, USA; [6]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA; [7]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, 90502, USA; [8]Laboratory for Clinical Biochemistry Research, University of Vermont, Colchester, VT, 05446, USA; [9]Department of Medicine, Duke University School of Medicine, Durham,

NC, 27710, USA; [10]Epidemiology Branch, Division of Cardiovascular Sciences, National Heart, Lung and Blood Institute, Bethesda, MD, 20892, USA; [11]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, 02115, USA; [12]Section of Genetic Medicine, The University of Chicago, Chicago, IL, 60637, USA.

Training dataset

To build our transcriptome prediction models, we used data from the Multi-Ethnic Study of Atherosclerosis (MESA) multi-omics pilot study of the NHLBI Trans-Omics for Precision Medicine (TOPMed) consortium (Bild et al. 2002). This data set includes genotypes derived from whole genome sequencing and transcripts per million (TPM) values derived from RNA-Seq for individuals of four different populations – African American (AFA), Chinese (CHN), European (EUR), and Hispanic/Latino (HIS) – for three different blood cell types: peripheral blood mononuclear cells (PBMC, ALL n = 1287, AFA n = 334, CHN n = 104, EUR n = 528, HIS n= 321), CD16+ monocytes (Mono, ALL n = 395, AFA n = 75, EUR n = 221, HIS n = 99), and CD4+ T-cells (T cells, ALL n = 397, AFA n = 75, EUR n = 224, HIS n = 98).

Genotype and RNA-Seq QC

We performed QC on each MESA tissue-population pair separately. For the genotype data (Freeze 8, phs001416.v2.p1), we excluded INDELs, multi-allelic SNPs, and ambiguous-strand SNPs (A/T, C/G), and removed the remaining variants with minor allele frequency (MAF) < 0.01 and Hardy-Weinberg equilibrium (HWE) $p < 1 \times 10^{-6}$ using PLINK v1.9 (Purcell et al. 2007). For chromosome X, filtering by HWE was only applied in variants found within the pseudoautosomal regions based on GRCh38 positions. Furthermore, for the non-pseudoautosomal region of X, male dosages were assigned either 0 or 2. After QC, the numbers of non-ambiguous SNPs remaining were: AFA = 15.7M; CHN = 8.4M; EUR = 9.7M; HIS = 13.2M.

For the RNA-Seq data, we also performed QC separately by tissue-population. First, we removed genes with average TPM values < 0.1. For some individuals, RNA expression levels were measured at two different time points (Exam 1 and Exam 5); thus, after log-transforming each measurement and adjusting for age and sex as covariates, we took the mean of the two time points (or the single adjusted log-transformed value, if expression levels were only measured once), performed rank-based inverse normal transformation, and adjusted for the first 10 genotype and 10 expression PCs. To estimate genotype and expression principal components, we used PC-AiR, which accounts for sample relatedness, known or not (Conomos, Miller, and Thornton 2015). For each tissue, we removed genes absent in at least one population. After QC, we had 17,585 genes in PBMC, 14,503 in Mono, and 16,647 in T cells.

Gene Expression Cis-Heritability Estimation

We estimated gene expression heritability (h^2) using cis-SNPs within the 1Mb region upstream of the transcription start site and 1Mb region downstream of the transcription end site. Using the genotype data filtered only by HWE P-value > 1×10^{-6}, for each tissue-population pair, we first performed LD-pruning with a 500 variants count window, a 50 variants count step, and a 0.2 r^2 threshold using PLINK v1.9 (Purcell et al. 2007). Then, for each gene, we extracted cis-SNPs and excluded SNPs with MAF < 0.01. Finally, to assess cis-SNP expression heritability, we estimated the genetic relationship matrix and h^2 using GCTA-GREML with the "--reml-no-constrain" option (Yang et al. 2010). We considered a gene heritable if it had a positive h^2 estimate (h^2 - 2*S.E. > 0.01 and p-value < 0.05) in at least one MESA population. In total, 9,206 genes were heritable in PBMC, 3,804 in Mono, and 4,053 in T cells. Only these genes are included in the final models and were analyzed in the results.

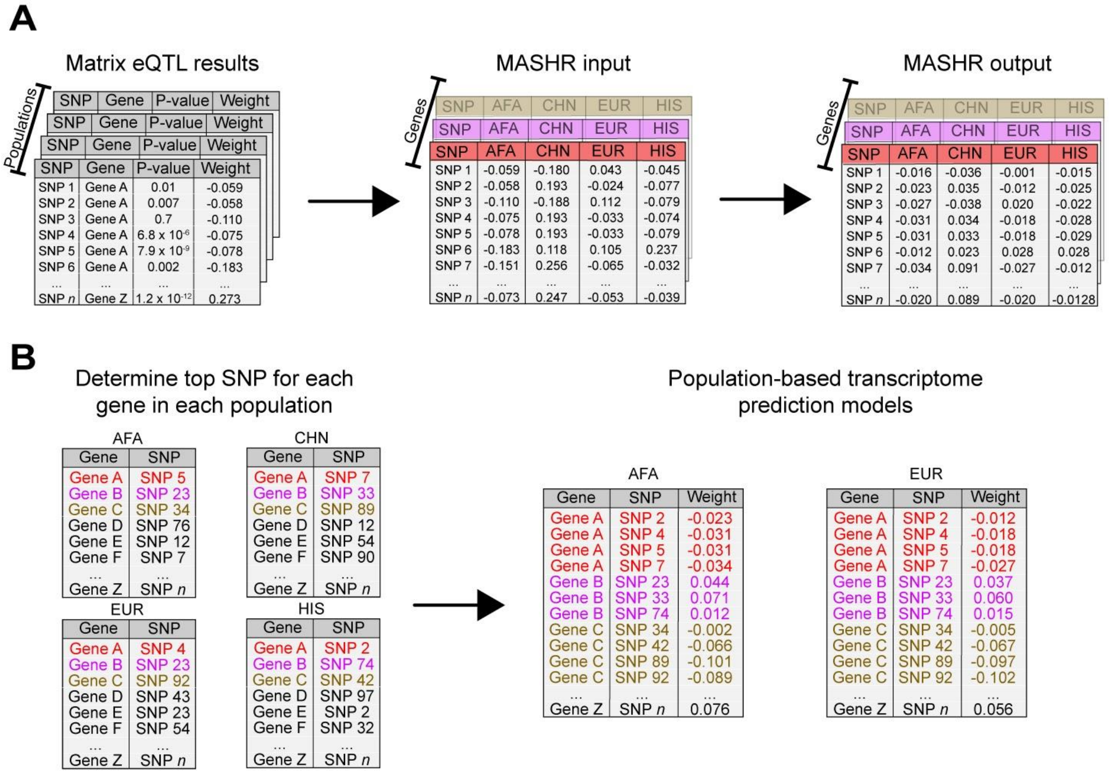
Transcriptome Prediction Models

With the aforementioned genotype and gene expression data, we built transcriptome prediction models for each MESA tissue-population pair, and for each gene we considered cis-SNPs as defined in the previous section. Additionally, we only considered SNPs present in the GWAS summary statistics of the Population Architecture using Genomics and Epidemiology (PAGE) study to build our prediction models (Wojcik et al. 2019). This step is important to make sure that there would be a high overlap between SNPs in the transcriptome models and SNPs in the GWAS summary statistics. After merging with PAGE SNPs, the average numbers of SNPs left in our dataset were: AFA = 12.8M; CHN = 6.2M; EUR = 7.4M; HIS = 10.5M.

We built our population-based models using three different approaches. The first one consists of a cross-validated elastic-net (EN) regression using the *glmnet* package in R, with mixing parameter $\alpha = 0.5$ (Zou and Hastie 2005; Friedman, Hastie, and Tibshirani 2010). We considered EN as our baseline model, as it has been previously used to make transcriptome prediction models for TOPMed MESA data (Mogil et al. 2018).

The second method implemented was mash (Multivariate Adaptive Shrinkage) in R (MASHR) (Urbut et al. 2019). Unlike EN, MASHR does not estimate weights by itself; rather, it takes zscore (or weight and standard error) matrices as input and adjusts them based on correlation patterns present in the data, allowing for both shared and population-specific effects. We ran MASHR for each gene at a time, using cis-SNPs weights estimated by Matrix eQTL and MESA populations as different conditions (Figure 2A) (Shabalin 2012). Then, we split MASHR-adjusted weights according to their respective populations, and selected the top

SNP (lowest local false sign rate) per gene to determine which SNPs would end up in the final models (Figure 2B). In order to make population-based models, we used population-specific effect sizes, taken from the corresponding MASHR output matrices.



Matrix eQTL results

SNP	Gene	P-value	Weight
SNP 1	Gene A	0.01	-0.059
SNP 2	Gene A	0.007	-0.058
SNP 3	Gene A	0.7	-0.110
SNP 4	Gene A	6.8×10^{-6}	-0.075
SNP 5	Gene A	7.9×10^{-9}	-0.078
SNP 6	Gene A	0.002	-0.183
...	...	...	...
SNP *n*	Gene Z	1.2×10^{-12}	0.273

MASHR input

SNP	AFA	CHN	EUR	HIS
SNP 1	-0.059	-0.180	0.043	-0.045
SNP 2	-0.058	0.193	-0.024	-0.077
SNP 3	-0.110	-0.188	0.112	-0.079
SNP 4	-0.075	0.193	-0.033	-0.074
SNP 5	-0.078	0.193	-0.033	-0.079
SNP 6	-0.183	0.118	0.105	0.237
SNP 7	-0.151	0.256	-0.065	-0.032
...	...	...	...	...
SNP *n*	-0.073	0.247	-0.053	-0.039

MASHR output

SNP	AFA	CHN	EUR	HIS
SNP 1	-0.016	-0.036	-0.001	-0.015
SNP 2	-0.023	0.035	-0.012	-0.025
SNP 3	-0.027	-0.038	0.020	-0.022
SNP 4	-0.031	0.034	-0.018	-0.028
SNP 5	-0.031	0.033	-0.018	-0.029
SNP 6	-0.012	0.023	0.028	0.028
SNP 7	-0.034	0.091	-0.027	-0.012
...	...	...	...	...
SNP *n*	-0.020	0.089	-0.020	-0.0128

Determine top SNP for each gene in each population

AFA

Gene	SNP
Gene A	SNP 5
Gene B	SNP 23
Gene C	SNP 34
Gene D	SNP 76
Gene E	SNP 12
Gene F	SNP 7
...	...
Gene Z	SNP *n*

CHN

Gene	SNP
Gene A	SNP 7
Gene B	SNP 33
Gene C	SNP 89
Gene D	SNP 12
Gene E	SNP 54
Gene F	SNP 90
...	...
Gene Z	SNP *n*

EUR

Gene	SNP
Gene A	SNP 4
Gene B	SNP 23
Gene C	SNP 92
Gene D	SNP 43
Gene E	SNP 23
Gene F	SNP 54
...	...
Gene Z	SNP *n*

HIS

Gene	SNP
Gene A	SNP 2
Gene B	SNP 74
Gene C	SNP 42
Gene D	SNP 97
Gene E	SNP 2
Gene F	SNP 32
...	...
Gene Z	SNP *n*

Population-based transcriptome prediction models

AFA

Gene	SNP	Weight
Gene A	SNP 2	-0.023
Gene A	SNP 4	-0.031
Gene A	SNP 5	-0.031
Gene A	SNP 7	-0.034
Gene B	SNP 23	0.044
Gene B	SNP 33	0.071
Gene B	SNP 74	0.012
Gene C	SNP 34	-0.002
Gene C	SNP 42	-0.066
Gene C	SNP 89	-0.101
Gene C	SNP 92	-0.089
...	...	...
Gene Z	SNP *n*	0.076

EUR

Gene	SNP	Weight
Gene A	SNP 2	-0.012
Gene A	SNP 4	-0.018
Gene A	SNP 5	-0.018
Gene A	SNP 7	-0.027
Gene B	SNP 23	0.037
Gene B	SNP 33	0.060
Gene B	SNP 74	0.015
Gene C	SNP 34	-0.005
Gene C	SNP 42	-0.067
Gene C	SNP 89	-0.097
Gene C	SNP 92	-0.102
...	...	...
Gene Z	SNP *n*	0.056

Figure 2: Design of the methodology implemented to make MASHR models. (A) Using effect sizes estimated using Matrix eQTL within each population dataset, we combined them across genes, with the different populations as conditions, to use as input for MASHR. The output matrixes contain adjusted effect sizes. (B) For each population, we selected the top SNP (lowest local false sign rate) per gene. Then, we concatenated the Gene-top SNP pairs across populations to determine which SNPs would end up in the final models. Lastly, to make our population-based transcriptome prediction models, we used population-specific effect sizes, taken from the corresponding MASHR output matrices. AFA = African American, CHN = Chinese, EUR = European, HIS = Hispanic/Latino.

The third and last method was based on the effect sizes estimated by Matrix eQTL using the linear regression model (Shabalin 2012). We used the same approach taken to build the MASHR models, but the key difference is that we made the models using the unadjusted effect sizes.

Assessing Transcriptome Prediction Performance

To evaluate the gene expression prediction performance of all our transcriptome prediction models, we used DNA and lymphoblastoid cell lines RNA-Seq data from 449 individuals in the Geuvadis study (Lappalainen et al. 2013). Individuals within the testing dataset belong to five different populations (Utah residents with Northern and Western European ancestry (CEU), n = 91; Finnish in Finland (FIN), n = 92; British in England and Scotland (GBR), n = 86; Toscani in Italy (TSI), n = 91; Yoruba in Ibadan, Nigeria (YRI), n = 89), which we analyzed both separately and together (ALL). As with our training dataset, we performed rank-based inverse normal transformation on the gene expression levels and adjusted for the first 10 genotype and 10 expression PCs. With the Geuvadis genotype data and our transcriptome prediction models, we used PrediXcan to estimate gene expression levels, and compared the estimated values to the adjusted, measured expression levels using Spearman correlation (Gamazon et al. 2015).

Applications in Association Studies

To test the applicability of our transcriptome prediction models in multi-ethnic association studies, we applied S-PrediXcan to GWAS summary statistics from the Population Architecture using Genomics and Epidemiology (PAGE) study (Barbeira et al. 2018; Wojcik et al. 2019). The PAGE study consists of different phenotypes tested for association with variants within a multi-ethnic, non-European cohort of 49,839 individuals (Hispanic/Latino [n=22,216], African American [n=17,299], Asian [n=4,680], Native Hawaiian [n=3,940], Native American [n=652] or Other [n=1,052]). The phenotypes investigated are included in the next table (Table 1).

Table 1. Matched PAGE and PanUKBB phenotypes.

PAGE Phenotypes	PanUKBB Phenotypes
Body mass index	Calculated or estimated body mass index
Coffee consumption	Coffee intake
C-reactive protein levels	C-reactive protein
Diastolic blood pressure	Automated or manual diastolic blood pressure
End-stage renal disease	End-stage renal disease
Estimated glomerular filtration rate	Glomerular filtration rate serum creatinine, glomerular filtration rate cystain C, glomerular filtration rate serum creatinine and cystain C
Fasting blood glucose	Fasting blood glucose, impaired or not
HDL cholesterol levels	HDL cholesterol levels
Height	Sitting or standing height
Hemoglobin A1c levels	Hemoglobin A1c
Hypertension	Hypertension or non-cancer hypertension
LDL cholesterol levels	LDL cholesterol levels
Mean corpuscular hemoglobin concentration	Mean corpuscular hemoglobin
Platelet count	Platelet count
PR interval	PR interval
QRS duration	QRS duration
Smoking behavior	Smoking behavior
Systolic blood pressure	Automated or manual systolic blood pressure
Total cholesterol levels	Total cholesterol levels
Triglyceride levels	Triglyceride levels
Type II diabetes	Type II diabetes
Waist-to-rip ratio	Waist-hip ratio hip circumference, waist-hip ratio waist circumference
White blood cell count	White blood cell count

Since we tested multiple phenotypes and transcriptome prediction models, we considered genes significantly associated with a phenotype if the association p-value was less than the Bonferroni corrected GWAS significance threshold of 5e-8.

To replicate the associations found in PAGE, we also applied S-PrediXcan to PanUKBB GWAS summary statistics (N=441,331; European [n=420,531], Central/South Asian [n=8,876], African [n=6,636], East Asian [n=2,709], Middle Eastern [n=1,599] or Admixed American [n=980]) (Barbeira et al. 2018; Pan UKBB Team 2022). For similarity purposes, we selected summary statistics of phenotypes that overlap with the ones tested in PAGE (Table 1). As previously described, a gene-trait pair association was considered significant if its p-value was less than the Bonferroni corrected GWAS significance threshold of 5e-8. Furthermore, we deemed significant gene-trait pair associations as replicated if they were detected by the same MESA tissue-population model and had the same direction of effect in PAGE and PanUKBB. To assess if the gene-trait association pairs reported in our study are novel or not, we compared them to studies found in the GWAS Catalog (All associations v1.0.2 file downloaded on 11/9/2022) (Buniello et al. 2019).

RESULTS

Publication disclaimer

Part of this work is available as a preprint at bioRxiv (doi.org/10.1101/2023.02.09.527747) and is under review for publication.

Increased Sample Sizes Improve Gene Expression Cis-Heritability Estimation

With the goal of improving transcriptome prediction in diverse populations, we first determined which gene expression traits were heritable and thus amenable to genetic prediction, using genome-wide genotype and RNA-Seq data from three blood cell types (PBMCs, monocytes, T cells) in TOPMed MESA. We estimated cis-heritability (h^2) using data from four different populations (African American - AFA, Chinese - CHN, European - EUR, and Hispanic/Latino - HIS). Variation in h^2 estimation between populations is expected due to differences in allele frequencies and LD patterns; however, we show that larger population sample sizes yield more h^2 estimates with $p < 0.05$ (Figure 3). For instance, with the EUR dataset (n = 528), we estimated h^2 for 10,228 genes, however, we estimated h^2 for 8,765 genes using the AFA dataset (n = 334) (Figure 3A). Moreover, we see a great impact on the CHN population, which has the smallest sample size. For that population, we managed to estimate h^2 for only 3,448 genes. The same pattern repeats when counting only the heritable genes (h^2 95% confidence interval lower bound > 0.01). In EUR, 6,902 genes were deemed heritable, whereas in AFA and CHN the number of heritable genes is 5,537 and 1,367, respectively (Figure 3B). Thus, larger sample sizes are needed to better pinpoint h^2 estimates, especially in non-European populations. In total, analyzing the union across all populations' results, we detected 9,206 heritable genes in PBMCs, 3,804 in monocytes, and 4,053 in T Cells.

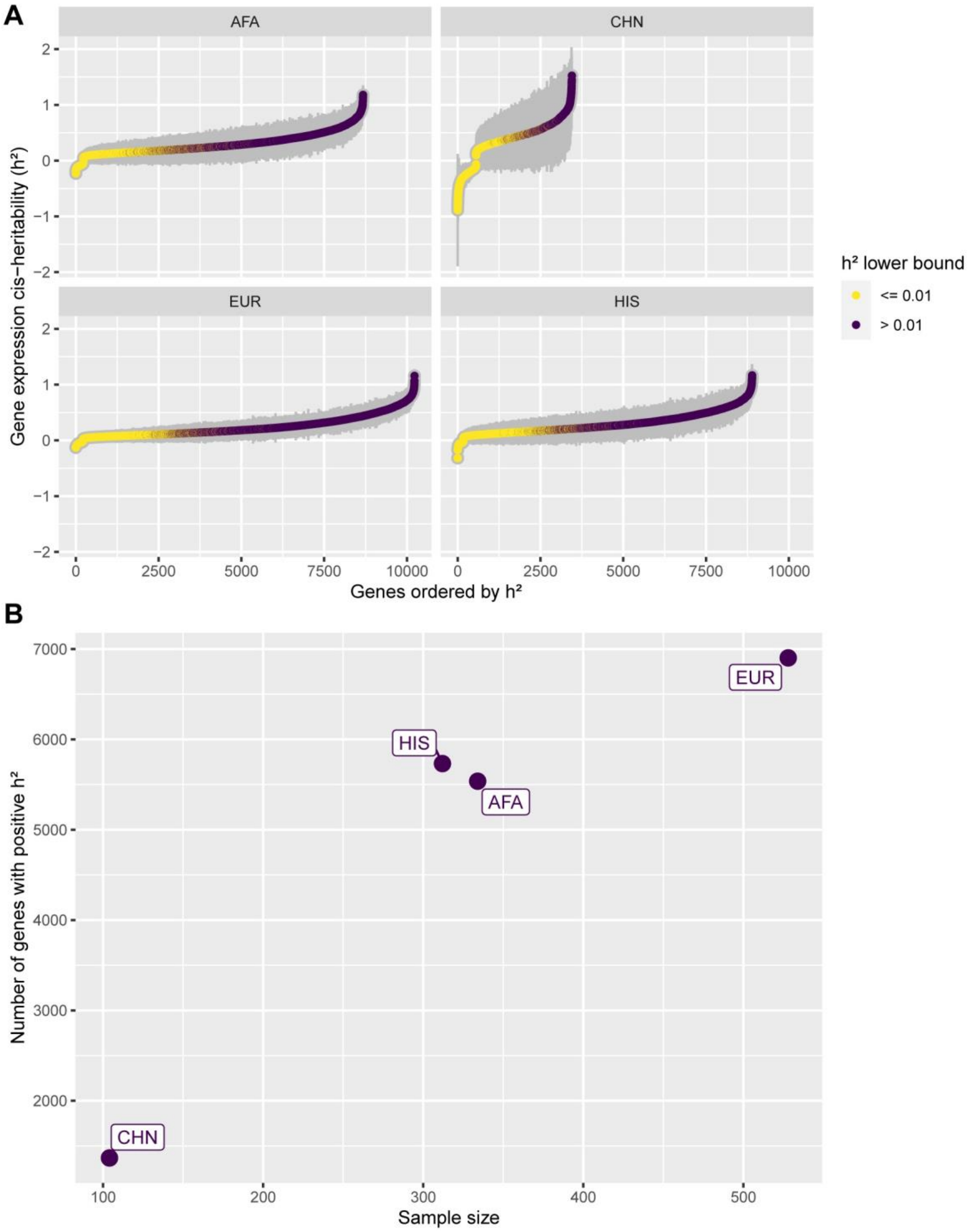


Figure 3: PBMC gene expression cis-heritability estimates across MESA populations. (A) Gene expression cis-heritability (h^2) estimated for different genes across different MESA population datasets. Only genes with significant estimated h^2 (p-value < 0.05) are shown.

Gray bars represent the standard errors (2*S.E.). Genes are ordered on the x-axis in ascending h^2 order, and colored according to the h^2 lower bound (h^2 - 2*S.E.). (B) Number of significant heritable genes (p-value < 0.05 and h^2 lower bound > 0.01) within each population dataset, by sample size. AFA = African American, CHN = Chinese, EUR = European, HIS = Hispanic/Latino.

MASHR Models Improve Cross-Population Prediction Performance

To improve TWAS power for discovery and replication across all populations, we sought to improve cross-population transcriptome prediction accuracy. For this, we used data from four different populations and built gene expression prediction models using three different methods (Elastic Net (EN), Matrix eQTL, and multivariate adaptive shrinkage in R (MASHR)). We chose EN as a baseline approach for comparison in our analysis, as it has been previously shown to have better performance than other common machine learning methods such as random forest, K-nearest neighbor, and support vector regression (Okoro et al. 2021). Matrix eQTL estimates univariate effect sizes for each cis-SNP-gene relationship and we developed an algorithm to include top SNPs from each population, but population-estimated effect sizes in each population’s model. Matrix eQTL effect sizes are the input for MASHR, which we hypothesized might better estimate cross-population effect sizes, due to its flexibility in allowing both shared and population-specific effects (Urbut et al. 2019; Barbeira et al. 2020). By filtering our models to include only genes with positive h^2 (h^2 lower bound > 0.01) in at least one population, we saw that among all methods used, we obtained more gene models in MatrixeQTL and MASHR in comparison to EN, especially in the CHN population model (Figure 4A). Specifically for chromosome X, EN models contained a low number of chrX genes for every population on average across all cell types analyzed (AFA=12, CHN=23, EUR=13, HIS=14). In comparison, both MatrixeQTL and MASHR had over 100 chrX genes for every population model on average across all cell types analyzed

(MatrixeQTL: AFA=111, CHN=191, EUR=108, HIS=111; MASHR: AFA=107, CHN=187, EUR=108, HIS=108).

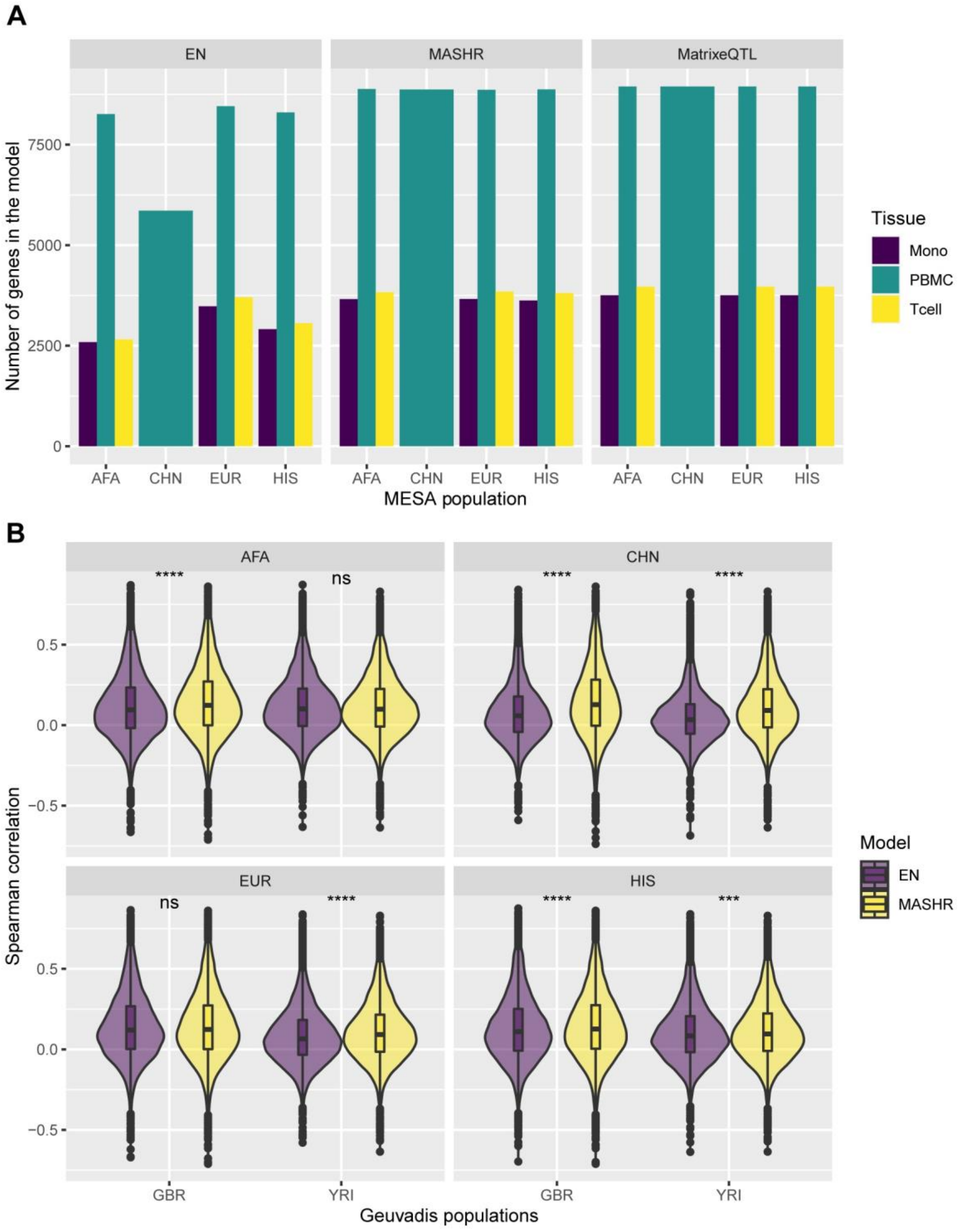


Figure 4: Comparison of MESA population transcriptome prediction models. (A) The number of genes in each MESA population model, by method and tissue. (B) Prediction performance (Spearman's rho) of MASHR and EN PBMC MESA population models in Geuvadis GBR

and YRI populations. Only genes with expression predicted by both methods for each MESA-Geuvadis population pair are shown. Differences in performance assessed through Wilcoxon rank sum tests; ns = not significant, *** = p-value ≤ 0.001, **** = p-value ≤ 0.0001.

To evaluate model performance at population-matched and cross-population transcriptome predictions, we used data from the Geuvadis study, which comprises individuals of West African or European descent. We defined "population-matched predictions" as the scenarios in which the transcriptome model MESA training data and Geuvadis test data have the closest genetic distance with available data, and we defined "cross-population predictions" as any other pairs (Figure 5).

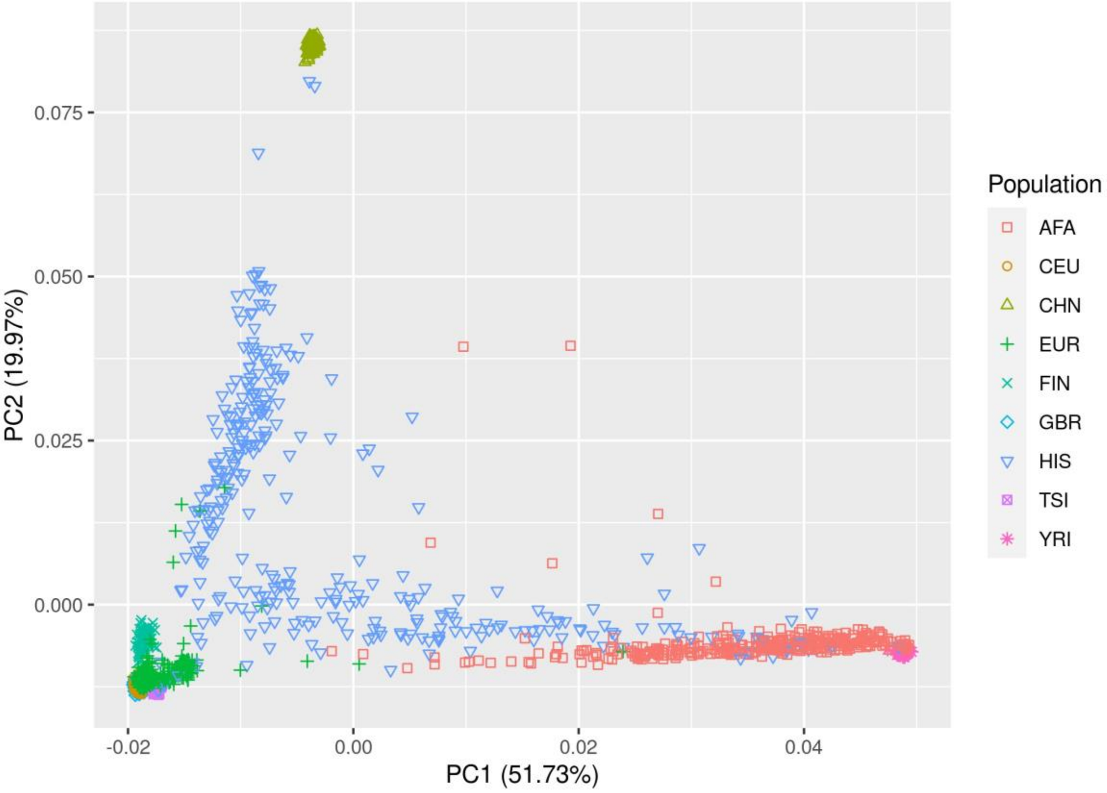


Figure 5: Genotype principal component analysis. Plot of the first two principal components of TOPMed MESA populations with Geuvadis populations. AFA = African American (TOPMed), CEU = Utah residents with Northern and Western European ancestry (Geuvadis), CHN = Chinese (TOPMed), EUR = European (TOPMed), FIN = Finnish in Finland

(Geuvadis), GBR = British in England and Scotland (Geuvadis), HIS = Hispanic/Latino (TOPMed), TSI = Toscani in Italy (Geuvadis), YRI = Yoruba in Ibadan, Nigeria (Geuvadis).

Focusing on Geuvadis GBR and YRI populations, which have similar sample sizes and are of distinct continental ancestries, we observed that MASHR models significantly outperform EN models in cross-population transcriptome predictions, considering genes with expression predicted by both methods, as seen in the AFA-GBR and EUR-YRI MESA-Geuvadis populations pairs (Figure 4B). We also see a higher prediction performance by the CHN and HIS MASHR models in comparison to EN, regardless of the Geuvadis population analyzed. However, in population-matched scenarios (AFA-YRI and EUR-GBR), prediction performance does not significantly differ between MASHR and EN methods. Similar results were obtained when comparing Matrix eQTL and EN (Figure 6A). Regarding MASHR and Matrix eQTL models, both methods perform the same in almost all cases, except for EUR-YRI and all CHN predictions, in which MASHR performed better (Figure 6B).

Overall, across all Geuvadis populations, MASHR models either performed better or the same as EN and MatrixeQTL models in both population-matched or cross-population transcriptome prediction scenarios (Table 2).

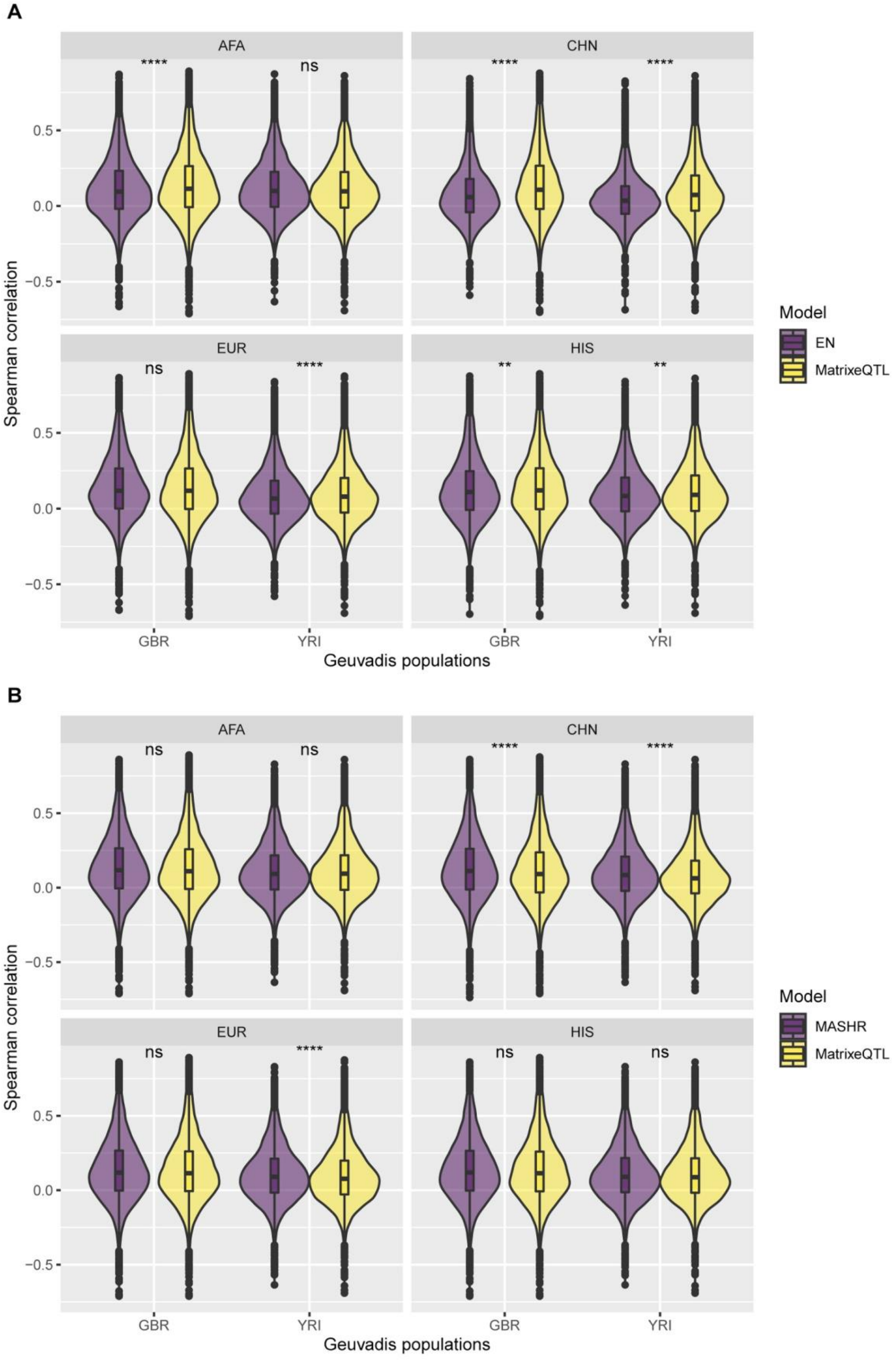

A
AFA
CHN
EUR
HIS
****
ns
****
****
ns
****
**
**
Spearman correlation
0.5
0.0
−0.5
GBR
YRI
Geuvadis populations
Model
EN
MatrixeQTL
B
AFA
CHN
EUR
HIS
ns
ns
****
****
ns
****
ns
ns
Model
MASHR
MatrixeQTL

Figure 6: Prediction performance of MESA population models in Geuvadis GBR and YRI populations. (A) Prediction performance (Spearman's rho) of EN and MatrixeQTL PBMC MESA population models in Geuvadis GBR and YRI populations. Only genes with expression predicted by both methods for each MESA-Geuvadis population pair are shown. Differences in performance assessed through Wilcoxon rank sum tests; ns = not significant, ** = p-value ≤ 0.01, **** = p-value ≤ 0.0001. (B) Prediction performance (Spearman's rho) of MASHR and MatrixeQTL PBMC MESA population models in Geuvadis GBR and YRI populations. Only genes with expression predicted by both methods for each MESA-Geuvadis population pair are shown. Differences in performance assessed through Wilcoxon rank sum tests; ns = not significant, **** = p-value ≤ 0.0001.

Table 2. Median gene expression prediction performance (Spearman's rho) of TOPMed MESA models in Geuvadis.

Tissue	Method	MESA Population	Geuvadis population	# of Genes	Median Spearman's rho
Mono	EN	AFA	ALL	2542	0.0343
Mono	EN	AFA	CEU	2437	0.0447
Mono	EN	AFA	FIN	2438	0.0423
Mono	EN	AFA	GBR	2434	0.0563
Mono	EN	AFA	TSI	2442	0.0472
Mono	EN	AFA	YRI	2536	0.0525
Mono	EN	EUR	ALL	3436	0.0966
Mono	EN	EUR	CEU	3434	0.1041
Mono	EN	EUR	FIN	3433	0.1171
Mono	EN	EUR	GBR	3434	0.1239
Mono	EN	EUR	TSI	3435	0.1082
Mono	EN	EUR	YRI	3414	0.0647
Mono	EN	HIS	ALL	2869	0.0643
Mono	EN	HIS	CEU	2841	0.0766
Mono	EN	HIS	FIN	2836	0.0852
Mono	EN	HIS	GBR	2842	0.0912
Mono	EN	HIS	TSI	2846	0.0797
Mono	EN	HIS	YRI	2839	0.0575
Mono	MASHR	AFA	ALL	3559	0.0905
Mono	MASHR	AFA	CEU	3520	0.0962
Mono	MASHR	AFA	FIN	3509	0.1056
Mono	MASHR	AFA	GBR	3518	0.1144
Mono	MASHR	AFA	TSI	3525	0.1030
Mono	MASHR	AFA	YRI	3461	0.0805
Mono	MASHR	EUR	ALL	3559	0.1028

Mono	MASHR	EUR	CEU	3525	0.1084
Mono	MASHR	EUR	FIN	3514	0.1195
Mono	MASHR	EUR	GBR	3524	0.1286
Mono	MASHR	EUR	TSI	3531	0.1148
Mono	MASHR	EUR	YRI	3453	0.0787
Mono	MASHR	HIS	ALL	3525	0.1015
Mono	MASHR	HIS	CEU	3490	0.1056
Mono	MASHR	HIS	FIN	3479	0.1166
Mono	MASHR	HIS	GBR	3489	0.1282
Mono	MASHR	HIS	TSI	3496	0.1142
Mono	MASHR	HIS	YRI	3426	0.0807
Mono	MatrixeQTL	AFA	ALL	3663	0.0715
Mono	MatrixeQTL	AFA	CEU	3543	0.0823
Mono	MatrixeQTL	AFA	FIN	3528	0.0926
Mono	MatrixeQTL	AFA	GBR	3537	0.0992
Mono	MatrixeQTL	AFA	TSI	3540	0.0880
Mono	MatrixeQTL	AFA	YRI	3629	0.0726
Mono	MatrixeQTL	EUR	ALL	3650	0.0918
Mono	MatrixeQTL	EUR	CEU	3637	0.1014
Mono	MatrixeQTL	EUR	FIN	3618	0.1127
Mono	MatrixeQTL	EUR	GBR	3636	0.1188
Mono	MatrixeQTL	EUR	TSI	3638	0.1042
Mono	MatrixeQTL	EUR	YRI	3418	0.0724
Mono	MatrixeQTL	HIS	ALL	3666	0.0853
Mono	MatrixeQTL	HIS	CEU	3607	0.0923
Mono	MatrixeQTL	HIS	FIN	3245	0.1186
Mono	MatrixeQTL	HIS	GBR	3607	0.1179
Mono	MatrixeQTL	HIS	TSI	3608	0.0997
Mono	MatrixeQTL	HIS	YRI	3581	0.0688
PBMC	EN	AFA	ALL	8115	0.0726
PBMC	EN	AFA	CEU	7983	0.0725
PBMC	EN	AFA	FIN	7972	0.0856
PBMC	EN	AFA	GBR	7979	0.0916
PBMC	EN	AFA	TSI	8002	0.0811
PBMC	EN	AFA	YRI	8102	0.0996
PBMC	EN	CHN	ALL	5578	0.0361
PBMC	EN	CHN	CEU	5506	0.0450
PBMC	EN	CHN	FIN	5541	0.0518
PBMC	EN	CHN	GBR	5499	0.0561
PBMC	EN	CHN	TSI	5515	0.0488

PBMC	EN	CHN	YRI	5488	0.0323
PBMC	EN	EUR	ALL	8312	0.0925
PBMC	EN	EUR	CEU	8310	0.0948
PBMC	EN	EUR	FIN	8298	0.1121
PBMC	EN	EUR	GBR	8308	0.1175
PBMC	EN	EUR	TSI	8307	0.1062
PBMC	EN	EUR	YRI	8242	0.0636
PBMC	EN	HIS	ALL	8161	0.0810
PBMC	EN	HIS	CEU	8096	0.0830
PBMC	EN	HIS	FIN	8083	0.0991
PBMC	EN	HIS	GBR	8087	0.1069
PBMC	EN	HIS	TSI	8095	0.0892
PBMC	EN	HIS	YRI	8127	0.0818
PBMC	MASHR	AFA	ALL	8642	0.0943
PBMC	MASHR	AFA	CEU	8410	0.0919
PBMC	MASHR	AFA	FIN	8378	0.1116
PBMC	MASHR	AFA	GBR	8393	0.1161
PBMC	MASHR	AFA	TSI	8434	0.1050
PBMC	MASHR	AFA	YRI	8452	0.0915
PBMC	MASHR	CHN	ALL	8625	0.0876
PBMC	MASHR	CHN	CEU	8398	0.0860
PBMC	MASHR	CHN	FIN	8366	0.1051
PBMC	MASHR	CHN	GBR	8381	0.1111
PBMC	MASHR	CHN	TSI	8422	0.0959
PBMC	MASHR	CHN	YRI	8434	0.0845
PBMC	MASHR	EUR	ALL	8618	0.0958
PBMC	MASHR	EUR	CEU	8391	0.0946
PBMC	MASHR	EUR	FIN	8359	0.1147
PBMC	MASHR	EUR	GBR	8374	0.1188
PBMC	MASHR	EUR	TSI	8415	0.1082
PBMC	MASHR	EUR	YRI	8428	0.0895
PBMC	MASHR	HIS	ALL	8628	0.0956
PBMC	MASHR	HIS	CEU	8401	0.0930
PBMC	MASHR	HIS	FIN	8369	0.1135
PBMC	MASHR	HIS	GBR	8384	0.1191
PBMC	MASHR	HIS	TSI	8425	0.1065
PBMC	MASHR	HIS	YRI	8437	0.0902
PBMC	MatrixeQTL	AFA	ALL	8733	0.0846
PBMC	MatrixeQTL	AFA	CEU	8527	0.0843
PBMC	MatrixeQTL	AFA	FIN	8519	0.1002

PBMC	MatrixeQTL	AFA	GBR	8528	0.1072
PBMC	MatrixeQTL	AFA	TSI	8547	0.0949
PBMC	MatrixeQTL	AFA	YRI	8662	0.0905
PBMC	MatrixeQTL	CHN	ALL	8331	0.0656
PBMC	MatrixeQTL	CHN	CEU	8203	0.0717
PBMC	MatrixeQTL	CHN	FIN	8250	0.0849
PBMC	MatrixeQTL	CHN	GBR	8193	0.0883
PBMC	MatrixeQTL	CHN	TSI	8211	0.0781
PBMC	MatrixeQTL	CHN	YRI	8058	0.0615
PBMC	MatrixeQTL	EUR	ALL	8687	0.0886
PBMC	MatrixeQTL	EUR	CEU	8666	0.0910
PBMC	MatrixeQTL	EUR	FIN	8640	0.1074
PBMC	MatrixeQTL	EUR	GBR	8660	0.1102
PBMC	MatrixeQTL	EUR	TSI	8670	0.1018
PBMC	MatrixeQTL	EUR	YRI	8312	0.0755
PBMC	MatrixeQTL	HIS	ALL	8721	0.0887
PBMC	MatrixeQTL	HIS	CEU	8609	0.0868
PBMC	MatrixeQTL	HIS	FIN	8601	0.1059
PBMC	MatrixeQTL	HIS	GBR	8610	0.1111
PBMC	MatrixeQTL	HIS	TSI	8620	0.0997
PBMC	MatrixeQTL	HIS	YRI	8602	0.0851
Tcell	EN	AFA	ALL	2601	0.0371
Tcell	EN	AFA	CEU	2499	0.0471
Tcell	EN	AFA	FIN	2500	0.0534
Tcell	EN	AFA	GBR	2503	0.0619
Tcell	EN	AFA	TSI	2511	0.0554
Tcell	EN	AFA	YRI	2584	0.0616
Tcell	EN	EUR	ALL	3645	0.1221
Tcell	EN	EUR	CEU	3643	0.1233
Tcell	EN	EUR	FIN	3640	0.1436
Tcell	EN	EUR	GBR	3643	0.1520
Tcell	EN	EUR	TSI	3643	0.1446
Tcell	EN	EUR	YRI	3610	0.0811
Tcell	EN	HIS	ALL	3002	0.0761
Tcell	EN	HIS	CEU	2973	0.0821
Tcell	EN	HIS	FIN	2959	0.1045
Tcell	EN	HIS	GBR	2966	0.1026
Tcell	EN	HIS	TSI	2972	0.0927
Tcell	EN	HIS	YRI	2959	0.0654
Tcell	MASHR	AFA	ALL	3713	0.1102

Tcell	MASHR	AFA	CEU	3669	0.1112
Tcell	MASHR	AFA	FIN	3669	0.1321
Tcell	MASHR	AFA	GBR	3677	0.1390
Tcell	MASHR	AFA	TSI	3682	0.1277
Tcell	MASHR	AFA	YRI	3622	0.0933
Tcell	MASHR	EUR	ALL	3727	0.1243
Tcell	MASHR	EUR	CEU	3693	0.1248
Tcell	MASHR	EUR	FIN	3693	0.1527
Tcell	MASHR	EUR	GBR	3699	0.1546
Tcell	MASHR	EUR	TSI	3703	0.1414
Tcell	MASHR	EUR	YRI	3628	0.0940
Tcell	MASHR	HIS	ALL	3692	0.1211
Tcell	MASHR	HIS	CEU	3657	0.1221
Tcell	MASHR	HIS	FIN	3657	0.1465
Tcell	MASHR	HIS	GBR	3663	0.1491
Tcell	MASHR	HIS	TSI	3667	0.1369
Tcell	MASHR	HIS	YRI	3602	0.0951
Tcell	MatrixeQTL	AFA	ALL	3865	0.0869
Tcell	MatrixeQTL	AFA	CEU	3736	0.0933
Tcell	MatrixeQTL	AFA	FIN	3734	0.1126
Tcell	MatrixeQTL	AFA	GBR	3747	0.1188
Tcell	MatrixeQTL	AFA	TSI	3747	0.1075
Tcell	MatrixeQTL	AFA	YRI	3809	0.0731
Tcell	MatrixeQTL	EUR	ALL	3848	0.1125
Tcell	MatrixeQTL	EUR	CEU	3833	0.1163
Tcell	MatrixeQTL	EUR	FIN	3826	0.1397
Tcell	MatrixeQTL	EUR	GBR	3840	0.1459
Tcell	MatrixeQTL	EUR	TSI	3839	0.1285
Tcell	MatrixeQTL	EUR	YRI	3608	0.0800
Tcell	MatrixeQTL	HIS	ALL	3856	0.1040
Tcell	MatrixeQTL	HIS	CEU	3806	0.1105
Tcell	MatrixeQTL	HIS	FIN	3804	0.1301
Tcell	MatrixeQTL	HIS	GBR	3813	0.1348
Tcell	MatrixeQTL	HIS	TSI	3821	0.1230
Tcell	MatrixeQTL	HIS	YRI	3734	0.0765

Leveraging Effect Sizes Across Different Populations Improves Discovery Rate in Multi-Ethnic TWAS

In order to investigate the applicability of the models we built in multi-ethnic TWAS, we used S-PrediXcan with GWAS summary statistics of complex traits from PAGE and PanUKBB. We show that across all tissue-population models, MASHR identified the highest number of gene-trait pair associations (208) that replicated in both PAGE and PanUKBB (P < 5e-8), followed by Matrix eQTL (173) and EN (94). Specifically for chromosome X, no EN model detected chrX genes-trait pair associations. In opposition to that, both MatrixeQTL and MASHR identified 5 chrX genes (*AVPR2*, *DNASE1L1*, *EMD*, *MECP2*, *RENBP*) associated to hemoglobin A1c levels. In addition to the aforementioned genes, MatrixeQTL also identified *VBP1* levels associated to hemoglobin A1c levels. Moreover, MASHR models were often the ones that reported a given association with the lowest p-value among all methods tested (Table 3). For instance, across 72 distinct gene-trait pairs associations found, MASHR had the lowest p-values for 39 of them (5 on chrX), followed by EN (20, 0 on chrX) and Matrix eQTL (13, 1 on chrX).

Table 3. All unique gene-trait association pairs that replicated in both PAGE and PanUKBB with same direction of effect, with the corresponding model that detected the association with the lowest p-value.

Gene	Phenotype	PAGE p-value	PanUKBB p-value	Model
ATP8B2	C-reactive protein	8.06E-12	6.91E-63	MatrixeQTL-HIS
AVPR2	Hemoglobin A1c	1.45E-42	2.85E-91	MASHR-AFA
BAK1	Platelet count	8.87E-25	4.80E-43	EN-AFA
BUD13	HDL	2.22E-08	9.32E-49	MASHR-AFA
BUD13	Triglycerides	5.65E-26	3.59E-108	MASHR-AFA
BUD23	Triglycerides	4.06E-13	5.71E-14	MatrixeQTL-AFA
C12orf43	C-reactive protein	1.95E-11	9.01E-133	MASHR-AFA
CAD	Triglycerides	2.84E-16	3.73E-32	MASHR-AFA
CBL	Platelet count	8.18E-09	8.72E-146	MASHR-EUR

CETP	Total cholesterol	7.70E-09	2.22E-57	MatrixeQTL-HIS
DNASE1L1	Hemoglobin A1c	3.99E-31	3.54E-49	MASHR-AFA
DOCK7	LDL	1.05E-08	2.97E-69	EN-AFA
DOCK7	Total cholesterol	6.88E-18	7.02E-141	EN-AFA
DOCK7	Triglycerides	2.07E-21	6.48E-304	EN-AFA
DPEP2	HDL	4.33E-11	1.12E-70	EN-EUR
DPEP3	HDL	1.15E-09	8.39E-58	MASHR-AFA
EMD	Hemoglobin A1c	1.92E-08	5.41E-10	MASHR-AFA
FADS1	LDL	9.47E-09	3.76E-28	MatrixeQTL-AFA
FADS1	Triglycerides	4.42E-08	5.99E-67	MatrixeQTL-AFA
FADS2	LDL	9.33E-12	1.65E-39	MASHR-CHN
FADS2	Triglycerides	2.70E-08	1.51E-58	MASHR-AFA
FCER1A	WBC count	4.49E-88	7.39E-17	MatrixeQTL-AFA
FCGR3B	WBC count	1.39E-10	8.31E-09	EN-HIS
FN3K	Hemoglobin A1c	5.32E-09	1.90E-82	EN-AFA
GFOD2	HDL	4.90E-08	6.96E-60	EN-EUR
GSDMA	WBC count	2.44E-08	1.29E-240	MASHR-CHN
GSDMB	HDL	4.69E-08	2.53E-20	MatrixeQTL-EUR
HHIP-AS1	Height	1.76E-09	4.35E-08	EN-AFA
IL6R	C-reactive protein	2.86E-20	1.31E-117	MASHR-AFA
KANK2	LDL	6.49E-14	1.29E-218	MatrixeQTL-AFA
KANK2	Total cholesterol	2.88E-10	2.20E-186	MatrixeQTL-AFA
KRTCAP3	C-reactive protein	3.63E-11	2.66E-116	MASHR-AFA
KRTCAP3	Fasting blood glucose	8.58E-09	3.29E-36	MASHR-AFA
KRTCAP3	Total cholesterol	8.46E-14	1.30E-41	MASHR-AFA
KRTCAP3	Triglycerides	1.16E-15	1.79E-12	EN-AFA
LA16c-349E10.1	Mean corpuscular hemoglobin	1.48E-09	4.42E-27	EN-HIS
LAMTOR2	WBC count	2.62E-24	3.97E-27	MASHR-AFA
LCAT	HDL	3.05E-08	3.24E-93	MASHR-AFA
LEPR	C-reactive protein	1.69E-10	0	MASHR-CHN
LMNA	WBC countWBC count	3.51E-35	1.00E-21	MASHR-AFA
LPL	HDL	1.80E-10	1.43E-50	EN-EUR
LPL	Triglycerides	2.43E-14	1.90E-10	MASHR-AFA
MECP2	Hemoglobin A1c	4.38E-08	2.62E-22	MASHR-HIS
MED24	WBC count	1.10E-17	0	MatrixeQTL-AFA
MEG3	Platelet count	5.13E-17	2.75E-139	MASHR-AFA
NRBF2	Platelet count	8.39E-22	7.05E-243	MASHR-AFA
NRBP1	C-reactive protein	2.80E-10	6.91E-70	MASHR-AFA
NRBP1	Fasting blood glucose	3.73E-08	2.14E-35	MASHR-AFA
NRBP1	Total cholesterol	1.35E-13	1.01E-19	MASHR-AFA
NRBP1	Triglycerides	2.08E-27	1.50E-153	MASHR-AFA

PAFAH1B2	Triglycerides	3.25E-14	1.53E-37	MASHR-AFA
PCSK7	Triglycerides	8.27E-14	2.54E-179	MASHR-CHN
POC5	Total cholesterol	1.71E-08	1.07E-39	MASHR-AFA
PSMD3	WBC count	5.30E-21	1.17E-211	MatrixeQTL-AFA
PSMD9	C-reactive protein	2.55E-10	1.14E-38	EN-AFA
PSRC1	LDL	1.25E-10	3.11E-12	EN-AFA
PSRC1	Total cholesterol	8.34E-55	1.97E-264	MASHR-AFA
RENBP	Hemoglobin A1c	2.87E-27	1.81E-44	MASHR-AFA
SCGB1C1	Platelet count	1.65E-08	8.16E-19	MatrixeQTL-EUR
SLC5A6	Triglycerides	6.65E-10	1.32E-21	EN-EUR
TMEM184B	C-reactive protein	4.29E-09	1.24E-08	MASHR-AFA
TMEM258	LDL	4.13E-10	4.15E-22	EN-AFA
TMEM258	Triglycerides	5.54E-10	2.32E-68	EN-AFA
TOMM40	C-reactive protein	5.31E-21	1.91E-138	EN-AFA
TOMM40	HDL	1.66E-11	4.46E-92	EN-AFA
TOMM40	LDL	3.97E-57	4.11E-16	EN-AFA
TPM4	Platelet count	1.64E-24	1.29E-142	MASHR-AFA
UQCC1	Height	1.92E-26	1.85E-11	MASHR-AFA
VBP1	Hemoglobin A1c	5.08E-54	4.57E-08	MatrixeQTL-HIS
YJEFN3	Triglycerides	6.07E-16	7.12E-86	MASHR-CHN
YKT6	Fasting blood glucose	9.62E-23	7.08E-235	MASHR-AFA
YKT6	Hemoglobin A1c	9.31E-11	2.61E-208	MASHR-AFA

When analyzing the total number of discoveries separately for each population, MASHR had the highest number of gene-trait pairs in most population models, with large discrepancies found in AFA and CHN models when comparing MASHR and EN (Figure 7A). Additionally, when comparing gene-trait pairs, we saw that most MASHR hits were shared between population models (Figure 7B), whereas in EN, the models have higher population-specific discoveries (Figure 7C). These findings suggest that MASHR models show high consistency and also suggest that TWAS results are not as affected by the MASHR population model used as compared to EN.

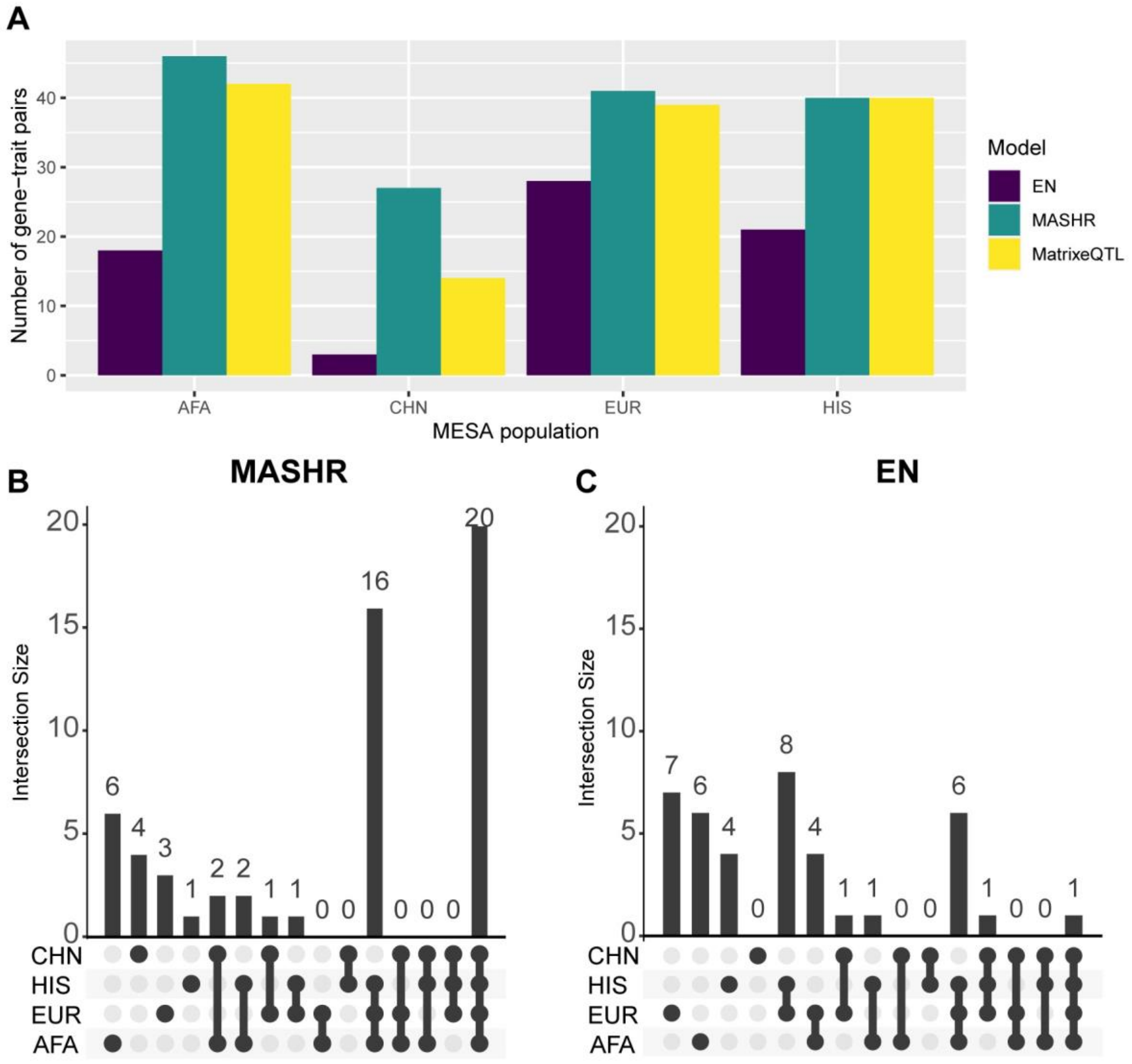


Figure 7: Number of significant S-PrediXcan gene-trait pairs in PAGE and PanUKBB GWAS summary statistics. (A) Total number of significant gene-trait pairs discovered by each MESA population model (considering the union of the three tissues), by method. (B) Number of significant gene-trait pairs discovered by MASHR MESA population models (considering the union of the three tissues). (C) Number of significant gene-trait pairs discovered by EN MESA population models (considering the union of the three tissues).

To contextualize our models' findings, we investigated whether the discovered gene-trait pairs had been previously reported in any studies in the GWAS Catalog (https://www.ebi.ac.uk/gwas/home). We saw that 19 out of the 72 (26.39%) distinct gene-trait association pairs have not been reported in the GWAS Catalog, and therefore may be novel associations that require further investigation (Table 4). Out of those potential new biological associations, most of them (13) were discovered with MASHR AFA models.

Table 4. Potentially novel gene-trait associations found in our TWAS and models that detected them.

Gene	Phenotype	Model
AVPR2	Hemoglobin A1c	MASHR-AFA, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-HIS
DNASE1L1	Hemoglobin A1c	MASHR-AFA, MatrixeQTL-AFA
EMD	Hemoglobin A1c	MASHR-AFA, MatrixeQTL-AFA
FCER1A	White blood cell count	MatrixeQTL-AFA
KRTCAP3	C-reactive protein	MASHR-AFA, MASHR-CHN, MASHR-EUR, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-CHN, MatrixeQTL-EUR, MatrixeQTL-HIS
KRTCAP3	Fasting blood glucose	MASHR-AFA, MASHR-CHN, MatrixeQTL-CHN
KRTCAP3	Total cholesterol	EN-EUR, MASHR-AFA, MASHR-CHN, MASHR-EUR, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-CHN, MatrixeQTL-EUR, MatrixeQTL-HIS
LA16c-3949E10.1	Mean corpuscular hemoglobin	EN-HIS
LAMTOR2	White blood cell count	EN-EUR, MASHR-AFA, MASHR-EUR, MASHR-HIS, MatrixeQTL-EUR, MatrixeQTL-HIS
MECP2	Hemoglobin A1c	MASHR-HIS, MatrixeQTL-HIS
MEG3	Platelet count	EN-EUR, EN-HIS, MASHR-AFA, MASHR-CHN, MASHR-EUR, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-EUR, MatrixeQTL-HIS
NRBF2	Platelet count	EN-AFA, EN-EUR, EN-HIS, MASHR-AFA, MASHR-CHN, MASHR-EUR, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-EUR, MatrixeQTL-HIS
NRBP1	C-reactive protein	EN-EUR, EN-HIS, MASHR-AFA, MASHR-EUR, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-EUR, MatrixeQTL-HIS
NRBP1	Fasting blood glucose	MASHR-AFA, MASHR-EUR, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-EUR, MatrixeQTL-HIS
PSMD9	C-reactive protein	EN-AFA
RENBP	Hemoglobin A1c	MASHR-AFA, MASHR-HIS, MatrixeQTL-AFA, MatrixeQTL-HIS

TMEM184B	C-reactive protein	MASHR-AFA
VBP1	Hemoglobin A1c	MatrixeQTL-HIS
YJEFN3	Triglycerides	MASHR-CHN

Furthermore, out of the 53 distinct known GWAS catalog associations discovered, MASHR models identified most of them. For instance, MASHR EUR models found 34 known associations, followed by MASHR AFA with 33, and MatrixeQTL with 32 (Figure 8).

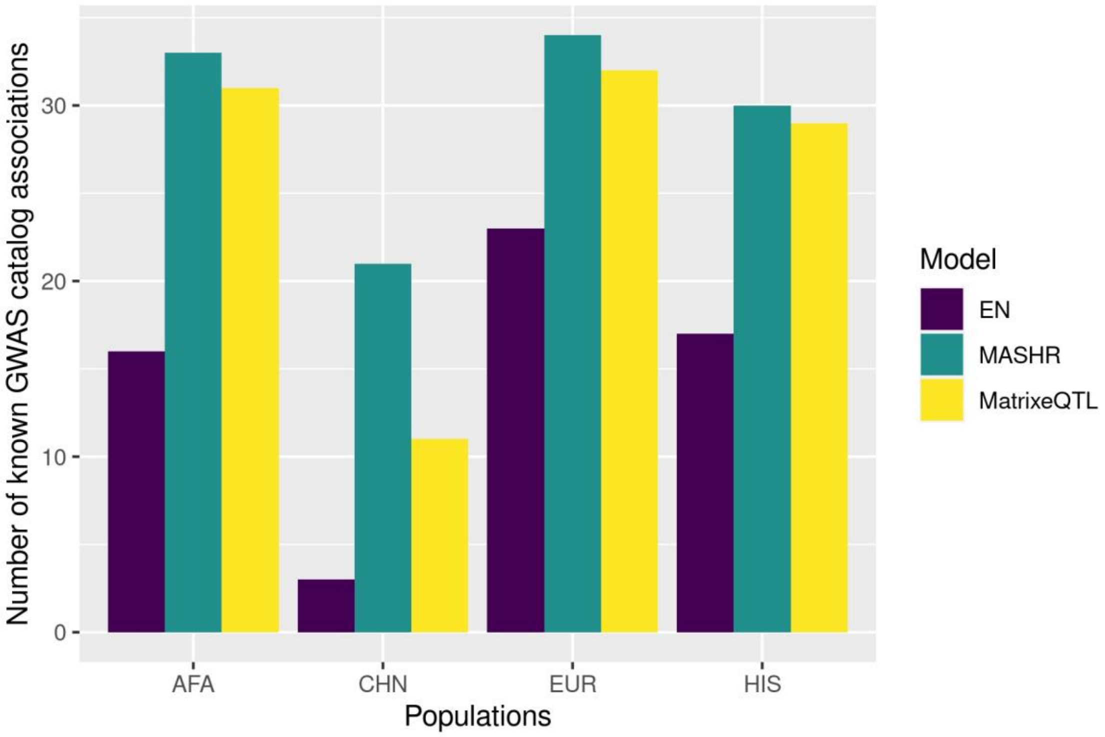


Figure 8: Number of significant S-PrediXcan gene-trait pairs in PAGE and PanUKBB GWAS summary statistics that have been reported in the GWAS catalog. Total number of significant gene-trait pairs discovered by each MESA population model (considering the union of the three tissues), by method.

DISCUSSION AND CONCLUSION

Publication disclaimer

Part of this work is available as a preprint at bioRxiv (doi.org/10.1101/2023.02.09.527747) and is under review for publication.

In this work, we sought to build population-based transcriptome prediction models for TWAS using data from the TOPMed MESA cohort using three distinct approaches. We saw that although the AFA and HIS populations' datasets contained the highest numbers of SNPs after quality control, EUR yielded the highest number of gene expression traits with significant heritability estimates across all tissues analyzed. This is most likely due to the higher sample size in EUR (n=528) in comparison to AFA (n=334) and HIS (n=321), as larger sample sizes provide higher statistical power to detect eQTLs with smaller effects (Aguet et al. 2017). Test data sample size has also been shown to positively correlate with gene expression prediction accuracy (Fryett, Morris, and Cordell 2020).

In addition to sample size, gene expression prediction accuracy is known to be greater when the training and testing datasets have similar ancestries (Keys et al. 2020; Mogil et al. 2018; Fryett, Morris, and Cordell 2020; Mikhaylova and Thornton 2019); however, non-European ancestries are vastly underrepresented in human genetics studies, which compromises the ability to build accurate TWAS models for them (Morales et al. 2018; Martin et al. 2019). Thus, using data from the Geuvadis cohort, we evaluated the transcriptome prediction performance of our models and found out that MASHR models either significantly outperformed EN and MatrixeQTL models, or had similar performance.

Previous studies have shown that by borrowing information across different conditions, such as tissues or cell types, MASHR identifies shared- or condition-specific eQTLs, which can enhance causal gene identification, as well as improve effect size estimation accuracy (Urbut et al. 2019; Sheng et al. 2021; Barbeira et al. 2020). Similarly, by leveraging effect size estimates across multiple populations, MASHR improved cross-population transcriptome prediction without compromising population-matched prediction accuracy.

Discovery and replication of TWAS associations are also related to the ancestries of the transcriptome prediction model training dataset and ancestries of the TWAS sample dataset (Geoffroy, Gregga, and Wheeler 2020). Thus, we assessed the applicability of our models in TWAS using S-PrediXcan on PAGE and PanUKBB GWAS summary statistics and found out that across all tissues and populations, MASHR models yielded the highest number of total gene-trait pairs associations, with MASHR AFA reporting the highest number. In this manner, it seems that although MASHR improved gene expression prediction accuracy for all populations analyzed, using transcriptome prediction models that match the ancestries of the GWAS dataset still yields the highest number of TWAS discoveries, which is in agreement with many previous works (Geoffroy, Gregga, and Wheeler 2020; Schubert et al. 2022; Bhattacharya et al. 2021; 2020; Kachuri et al. 2021). Among the most significant gene-trait associations found, most have been previously reported in the GWAS Catalog. Examples include *MED24* and white blood cell count (PAGE effect size = -0.044, PanUKBB effect size = -0.221), who has been previously reported in GWAS conducted with the eMERGE and HCHS/SOL cohorts (Crosslin et al. 2012; Jain et al. 2017); *LEPR* and C-reactive protein levels (PAGE effect size = 0.506, PanUKBB effect size = 1.054), also identified in a large GWAS meta-analysis across over 80,000 individuals (Dehghan et al. 2011); and *DOCK7* and triglyceride levels (PAGE effect size = 3.379, PanUKBB effect size

= 8.034), also reported in a multiancestry GWAS meta-analysis across approximately 400,000 subjects (de Vries et al. 2019).

Furthermore, by investigating which associations had been previously reported in the GWAS Catalog, we saw that most new discoveries were found by MASHR models. In fact, one possible novel association reported by MASHR was the fifth most significant associations found across all gene-trait pair associations (*NRBF2* and platelet count, PAGE effect size = -12.119, p-value = 8.38e-22; PanUKBB effect size = -0.199, p-value = 7.054e-243). The same association was also reported by EN and MatrixeQTL, with the same direction of effects but not as significant. Some of these possible new discoveries are unique to MASHR models and have been corroborated previously, such as *YJEFN3* (also known as *AIBP2*) and triglycerides, whose low expression in zebrafish increases cellular unesterified cholesterol levels, consistent with our S-PrediXcan effect size directions (PAGE effect size = -0.522, p-value = 6.07e-16; PanUKBB effect size = -0.860, p-value = 7.12e-86) (Fang et al. 2013). Additionally, we also saw that MASHR models showed higher consistency than EN, which means that TWAS results are not as affected by the population model used as EN.

One limitation of our TWAS is that we used transcriptome prediction models trained in PBMCs, monocytes and T cells, and those tissues might not be the most appropriate for some phenotypes in PAGE or PanUKBB. Additionally, because of the smaller sample sizes for some populations in our training dataset, h^2 and eQTL effect sizes estimates have large standard errors, which may affect the ability of MASHR to adjust effect sizes across different conditions based on correlation patterns present in the data. Regardless of that, our results mainly demonstrate that we can implement cross-population effect size leveraging using a method first applied to do cross-tissue effect size leveraging - and improve cross-population transcriptome prediction accuracy in doing so. Thus, increasing sample size for underrepresented populations will improve current MASHR TWAS models' performances, as well as increase genetic diversity in the data. MASHR is most useful when population effects are shared, as demonstrated by the more consistent S-PrediXcan results, but population-specific effects are also relevant. For instance, a study in a large African American and Latino cohort discovered eQTLs only present at appreciable allele frequencies in African ancestry populations (Kachuri et al. 2021). Moreover, since our MASHR and MatrixeQTL models focus on the top SNPs, we might not be including enough eQTLs in the models, especially for those genes whose expression is genetically regulated by multiple eQTLs with small effects.

In conclusion, our results demonstrate the importance and the benefits of increasing ancestry diversity in the field of human genetics, especially regarding association studies. As shown, sample size is valuable for assessing gene expression heritability and for accurately estimating eQTL effect sizes, and thus some populations are negatively affected due to the lack of data. However, by making transcriptome prediction models that leverage effect size estimates across different populations using multivariate adaptive shrinkage, we were able to increase gene expression prediction performance for scenarios in which the training data and test data have distant ("cross-population") genetic distances with available data. Additionally, when applied to multi-ethnic TWAS, the MASHR models yielded more discoveries across all methods analyzed, even detecting well-known associations that were not detected by other methods. Thus, in order to further improve TWAS in multi-ethnic or underrepresented populations and possibly reduce health care disparities, it is necessary to use methods that consider shared and population-specific effect sizes, as well as increase available data of underrepresented populations.

REFERENCE LIST

Aguet, François, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, et al. 2017. "Genetic Effects on Gene Expression across Human Tissues." *Nature* 550 (7675): 204–13. https://doi.org/10.1038/nature24277.

Altshuler, David M., Richard A. Gibbs, Leena Peltonen, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, Emmanouil Dermitzakis, et al. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58. https://doi.org/10.1038/nature09298.

Barbeira, Alvaro N., Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, Eric S. Torstenson, et al. 2018. "Exploring the Phenotypic Consequences of Tissue Specific Gene Expression Variation Inferred from GWAS Summary Statistics." *Nature Communications* 9 (1): 1825. https://doi.org/10.1038/s41467-018-03621-1.

Barbeira, Alvaro N., Owen J. Melia, Yanyu Liang, Rodrigo Bonazzola, Gao Wang, Heather E. Wheeler, François Aguet, Kristin G. Ardlie, Xiaoquan Wen, and Hae K. Im. 2020. "Fine-Mapping and QTL Tissue-Sharing Information Improves the Reliability of Causal Gene Identification." *Genetic Epidemiology* 44 (8): 854–67. https://doi.org/10.1002/gepi.22346.

Barbeira, Alvaro N., Milton Pividori, Jiamao Zheng, Heather E. Wheeler, Dan L. Nicolae, and Hae Kyung Im. 2019. "Integrating Predicted Transcriptome from Multiple Tissues Improves Association Detection." *PLOS Genetics* 15 (1): e1007889. https://doi.org/10.1371/journal.pgen.1007889.

Bhattacharya, Arjun, Montserrat García-Closas, Andrew F. Olshan, Charles M. Perou, Melissa A. Troester, and Michael I. Love. 2020. "A Framework for Transcriptome-Wide Association Studies in Breast Cancer in Diverse Study Populations." *Genome Biology* 21 (1): 42. https://doi.org/10.1186/s13059-020-1942-6.

Bhattacharya, Arjun, Jibril B. Hirbo, Dan Zhou, Wei Zhou, Jie Zheng, Masahiro Kanai, the Global Biobank Meta-analysis Initiative, Bogdan Pasaniuc, Eric R. Gamazon, and Nancy J. Cox. 2021. "Best Practices for Multi-Ancestry, Meta-Analytic Transcriptome-Wide Association Studies: Lessons from the Global Biobank Meta-Analysis Initiative." Preprint. Genetic and Genomic Medicine. https://doi.org/10.1101/2021.11.24.21266825.

Bild, Diane E., David A. Bluemke, Gregory L. Burke, Robert Detrano, Ana V. Diez Roux, Aaron R. Folsom, Philip Greenland, et al. 2002. "Multi-Ethnic Study of

Atherosclerosis: Objectives and Design." *American Journal of Epidemiology* 156 (9): 871–81. https://doi.org/10.1093/aje/kwf113.

Brumpton, Ben M., and Manuel A. R. Ferreira. 2016. "Multivariate EQTL Mapping Uncovers Functional Variation on the X-Chromosome Associated with Complex Disease Traits." *Human Genetics* 135 (7): 827–39. https://doi.org/10.1007/s00439-016-1674-6.

Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–12. https://doi.org/10.1093/nar/gky1120.

Campbell, Michael C., and Sarah A. Tishkoff. 2008. "African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping." *Annual Review of Genomics and Human Genetics* 9 (1): 403–33. https://doi.org/10.1146/annurev.genom.9.081307.164258.

Cann, Rebecca L., Mark Stoneking, and Allan C. Wilson. 1987. "Mitochondrial DNA and Human Evolution." *Nature* 325 (6099): 31–36. https://doi.org/10.1038/325031a0.

Cano-Gamez, Eddie, and Gosia Trynka. 2020. "From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases." *Frontiers in Genetics* 11. https://www.frontiersin.org/articles/10.3389/fgene.2020.00424.

Chakravarti, Aravinda, and Peter Little. 2003. "Nature, Nurture and Human Disease." *Nature* 421 (6921): 412–14. https://doi.org/10.1038/nature01401.

Clarke, Laura, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, and Paul Flicek. 2017. "The International Genome Sample Resource (IGSR): A Worldwide Collection of Genome Variation Incorporating the 1000 Genomes Project Data." *Nucleic Acids Research* 45 (D1): D854–59. https://doi.org/10.1093/nar/gkw829.

Conomos, Matthew P., Michael B. Miller, and Timothy A. Thornton. 2015. "Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness." *Genetic Epidemiology* 39 (4): 276–93. https://doi.org/10.1002/gepi.21896.

Crosslin, David R., Andrew McDavid, Noah Weston, Sarah C. Nelson, Xiuwen Zheng, Eugene Hart, Mariza de Andrade, et al. 2012. "Genetic Variants Associated with the White Blood Cell Count in 13,923 Subjects in the EMERGE Network." *Human Genetics* 131 (4): 639–52. https://doi.org/10.1007/s00439-011-1103-9.

Dandine-Roulland, Claire, and Hervé Perdry. 2015. "Where Is the Causal Variant? On the Advantage of the Family Design over the Case–Control Design in Genetic Association Studies." *European Journal of Human Genetics* 23 (10): 1357–63. https://doi.org/10.1038/ejhg.2014.284.

Dehghan, Abbas, Josée Dupuis, Maja Barbalic, Joshua C. Bis, Gudny Eiriksdottir, Chen Lu, Niina Pellikka, et al. 2011. "Meta-Analysis of Genome-Wide Association Studies in >80 000 Subjects Identifies Multiple Loci for C-Reactive Protein Levels." *Circulation* 123 (7): 731–38. https://doi.org/10.1161/CIRCULATIONAHA.110.948570.

Derrida, Bernard, Susanna C. Manrubia, and Damián H. Zanette. 2000. "On the Genealogy of a Population of Biparental Individuals." *Journal of Theoretical Biology* 203 (3): 303–15. https://doi.org/10.1006/jtbi.2000.1095.

Fang, Longhou, Soo-Ho Choi, Ji Sun Baek, Chao Liu, Felicidad Almazan, Florian Ulrich, Philipp Wiesner, et al. 2013. "Control of Angiogenesis by AIBP-Mediated Cholesterol Efflux." *Nature* 498 (7452): 118–22. https://doi.org/10.1038/nature12166.

Fine, Michael J., Said A. Ibrahim, and Stephen B. Thomas. 2005. "The Role of Race and Genetics in Health Disparities Research." *American Journal of Public Health* 95 (12): 2125–28. https://doi.org/10.2105/AJPH.2005.076588.

Friedman, Jerome H., Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Fryett, James J., Andrew P. Morris, and Heather J. Cordell. 2020. "Investigation of Prediction Accuracy and the Impact of Sample Size, Ancestry, and Tissue in Transcriptome-Wide Association Studies." *Genetic Epidemiology* 44 (5): 425–41. https://doi.org/10.1002/gepi.22290.

Gamazon, Eric R., Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, et al. 2015. "A Gene-Based Association Method for Mapping Traits Using Reference Transcriptome Data." *Nature Genetics* 47 (9): 1091–98. https://doi.org/10.1038/ng.3367.

Gaziano, John Michael, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, et al. 2016. "Million Veteran Program: A Mega-Biobank to Study Genetic Influences on Health and Disease." *Journal of Clinical Epidemiology* 70 (February): 214–23. https://doi.org/10.1016/j.jclinepi.2015.09.016.

Geoffroy, Elyse, Isabelle Gregga, and Heather E. Wheeler. 2020. "Population-Matched Transcriptome Prediction Increases TWAS Discovery and Replication Rate." *IScience* 23 (12): 101850. https://doi.org/10.1016/j.isci.2020.101850.

Gibson, Greg. 2008. "The Environmental Contribution to Gene Expression Profiles." *Nature Reviews Genetics* 9 (8): 575–81. https://doi.org/10.1038/nrg2383.

Henn, Brenna M, Teresa E Steele, and Timothy D Weaver. 2018. "Clarifying Distinct Models of Modern Human Origins in Africa." *Current Opinion in Genetics & Development*, Genetics of Human Origins, 53 (December): 148–56. https://doi.org/10.1016/j.gde.2018.10.003.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45. https://doi.org/10.1038/nature03001.

Jain, Deepti, Chani J. Hodonsky, Ursula M. Schick, Jean V. Morrison, Sharon Minnerath, Lisa Brown, Claudia Schurmann, et al. 2017. “Genome-Wide Association of White Blood Cell Counts in Hispanic/Latino Americans: The Hispanic Community Health Study/Study of Latinos.” *Human Molecular Genetics* 26 (6): 1193–1204. https://doi.org/10.1093/hmg/ddx024.

Kachuri, Linda, Angel C.Y. Mak, Donglei Hu, Celeste Eng, Scott Huntsman, Jennifer R. Elhawary, Namrata Gupta, et al. 2021. “Gene Expression in African Americans and Latinos Reveals Ancestry-Specific Patterns of Genetic Architecture.” Preprint. Genetics. https://doi.org/10.1101/2021.08.19.456901.

Keys, Kevin L., Angel C. Y. Mak, Marquitta J. White, Walter L. Eckalbar, Andrew W. Dahl, Joel Mefford, Anna V. Mikhaylova, et al. 2020. “On the Cross-Population Generalizability of Gene Expression Prediction Models.” *PLOS Genetics* 16 (8): e1008927. https://doi.org/10.1371/journal.pgen.1008927.

Kichaev, Gleb, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. 2014. “Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies.” *PLoS Genetics* 10 (10): e1004722. https://doi.org/10.1371/journal.pgen.1004722.

Kirsten, Holger, Hoor Al-Hasani, Lesca Holdt, Arnd Gross, Frank Beutner, Knut Krohn, Katrin Horn, et al. 2015. “Dissecting the Genetics of the Human Transcriptome Identifies Novel Trait-Related Trans-EQTLs and Corroborates the Regulatory Relevance of Non-Protein Coding Loci.” *Human Molecular Genetics* 24 (16): 4746–63. https://doi.org/10.1093/hmg/ddv194.

Klein, Richard G. 2008. “Out of Africa and the Evolution of Human Behavior.” *Evolutionary Anthropology: Issues, News, and Reviews* 17 (6): 267–81. https://doi.org/10.1002/evan.20181.

Kukurba, Kimberly R., Princy Parsana, Brunilda Balliu, Kevin S. Smith, Zachary Zappala, David A. Knowles, Marie-Julie Favé, et al. 2016. “Impact of the X Chromosome and Sex on Regulatory Variation.” *Genome Research* 26 (6): 768–77. https://doi.org/10.1101/gr.197897.115.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ‘t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468): 506–11. https://doi.org/10.1038/nature12531.

Lewis, Anna C. F., Santiago J. Molina, Paul S. Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M. Fullerton, et al. 2022. “Getting Genetic Ancestry Right for Science and Society.” *Science* 376 (6590): 250–52. https://doi.org/10.1126/science.abm7530.

Li, Binglan, Shefali S. Verma, Yogasudha C. Veturi, Anurag Verma, Yuki Bradford, David W. Haas, and Marylyn D. Ritchie. 2018. "Evaluation of PrediXcan for Prioritizing GWAS Associations and Predicting Gene Expression." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 23: 448–59.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. "The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog)." *Nucleic Acids Research* 45 (D1): D896–901. https://doi.org/10.1093/nar/gkw1133.

Mancuso, Nicholas, Simon Gayther, Alexander Gusev, Wei Zheng, Kathryn L. Penney, Zsofia Kote-Jarai, Rosalind Eeles, Matthew Freedman, Christopher Haiman, and Bogdan Pasaniuc. 2018. "Large-Scale Transcriptome-Wide Association Study Identifies New Prostate Cancer Risk Regions." *Nature Communications* 9 (1): 4079. https://doi.org/10.1038/s41467-018-06302-1.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics* 51 (4): 584–91. https://doi.org/10.1038/s41588-019-0379-x.

Mikhaylova, Anna V., and Timothy A. Thornton. 2019. "Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations." *Frontiers in Genetics* 10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00261.

Mizuno, Akira, and Yukinori Okada. 2019. "Biological Characterization of Expression Quantitative Trait Loci (EQTLs) Showing Tissue-Specific Opposite Directional Effects." *European Journal of Human Genetics* 27 (11): 1745–56. https://doi.org/10.1038/s41431-019-0468-4.

Mogil, Lauren S., Angela Andaleon, Alexa Badalamenti, Scott P. Dickinson, Xiuqing Guo, Jerome I. Rotter, W. Craig Johnson, Hae Kyung Im, Yongmei Liu, and Heather E. Wheeler. 2018. "Genetic Architecture of Gene Expression Traits across Diverse Populations." *PLOS Genetics* 14 (8): e1007586. https://doi.org/10.1371/journal.pgen.1007586.

Morales, Joannella, Danielle Welter, Emily H. Bowler, Maria Cerezo, Laura W. Harris, Aoife C. McMahon, Peggy Hall, et al. 2018. "A Standardized Framework for Representation of Ancestry Data in Genomics Studies, with Application to the NHGRI-EBI GWAS Catalog." *Genome Biology* 19 (1): 21. https://doi.org/10.1186/s13059-018-1396-2.

Mulford, Ashley J, Claudia Wing, M Eileen Dolan, and Heather E Wheeler. 2021. "Genetically Regulated Expression Underlies Cellular Sensitivity to Chemotherapy in Diverse Populations." *Human Molecular Genetics* 30 (3–4): 305–17. https://doi.org/10.1093/hmg/ddab029.

Nica, Alexandra C., and Emmanouil T. Dermitzakis. 2013. “Expression Quantitative Trait Loci: Present and Future.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (1620): 20120362. https://doi.org/10.1098/rstb.2012.0362.

Okoro, Paul C., Ryan Schubert, Xiuqing Guo, W. Craig Johnson, Jerome I. Rotter, Ina Hoeschele, Yongmei Liu, et al. 2021. “Transcriptome Prediction Performance across Machine Learning Models and Diverse Ancestries.” *Human Genetics and Genomics Advances* 2 (2): 100019. https://doi.org/10.1016/j.xhgg.2020.100019.

Pan UKBB Team. 2022. “Pan UKBB.” 2022. https://pan.ukbb.broadinstitute.org/.

Petersen, Ashley, Carolina Alvarez, Scott DeClaire, and Nathan L. Tintle. 2013. “Assessing Methods for Assigning SNPs to Genes in Gene-Based Tests of Association Using Common Variants.” *PLoS ONE* 8 (5): e62161. https://doi.org/10.1371/journal.pone.0062161.

Popejoy, Alice B., and Stephanie M. Fullerton. 2016. “Genomics Is Failing on Diversity.” *Nature* 538 (7624): 161–64. https://doi.org/10.1038/538161a.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75. https://doi.org/10.1086/519795.

Qiao, Jiahao, Zhonghe Shao, Yuxuan Wu, Ping Zeng, and Ting Wang. 2022. “Detecting Associated Genes for Complex Traits Shared across East Asian and European Populations under the Framework of Composite Null Hypothesis Testing.” *Journal of Translational Medicine* 20 (1): 424. https://doi.org/10.1186/s12967-022-03637-8.

Rutherford, Adam. 2020. *How to Argue with a Racist: What Our Genes Do (and Don’t) Say about Human Difference*. New York: The Experiment.

Schubert, Ryan, Elyse Geoffroy, Isabelle Gregga, Ashley J. Mulford, Francois Aguet, Kristin Ardlie, Robert Gerszten, et al. 2022. “Protein Prediction for Trait Mapping in Diverse Populations.” *PLOS ONE* 17 (2): e0264341. https://doi.org/10.1371/journal.pone.0264341.

Shabalin, Andrey A. 2012. “Matrix EQTL: Ultra Fast EQTL Analysis via Large Matrix Operations.” *Bioinformatics* 28 (10): 1353–58. https://doi.org/10.1093/bioinformatics/bts163.

Shastry, B. S. 2002. “SNP Alleles in Human Disease and Evolution.” *Journal of Human Genetics* 47 (11): 0561–66. https://doi.org/10.1007/s100380200086.

Sheng, Xin, Yuting Guan, Ziyuan Ma, Junnan Wu, Hongbo Liu, Chengxiang Qiu, Steven Vitale, et al. 2021. “Mapping the Genetic Architecture of Human Traits to Cell Types in the Kidney Identifies Mechanisms of Disease and Potential Treatments.” *Nature Genetics* 53 (9): 1322–33. https://doi.org/10.1038/s41588-021-00909-9.

Sollis, Elliot, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, et al. 2023. “The NHGRI-EBI GWAS Catalog: Knowledgebase and Deposition Resource.” *Nucleic Acids Research* 51 (D1): D977–85. https://doi.org/10.1093/nar/gkac1010.

Stringer, Chris. 2014. “Why We Are Not All Multiregionalists Now.” *Trends in Ecology & Evolution* 29 (5): 248–51. https://doi.org/10.1016/j.tree.2014.03.001.

Sun, Lei, Zhong Wang, Tianyuan Lu, Teri A. Manolio, and Andrew D. Paterson. 2023. “EXclusionarY: Ten Years Later, Where Are the Sex Chromosomes in GWAS?” Preprint. Genetics. https://doi.org/10.1101/2023.02.03.526992.

Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, et al. 2021. “Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program.” *Nature* 590 (7845): 290–99. https://doi.org/10.1038/s41586-021-03205-y.

The 1000 Genomes Project Consortium. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422): 56–65. https://doi.org/10.1038/nature11632.

The All of Us Research Program Investigators. 2019. “The ‘All of Us’ Research Program.” *New England Journal of Medicine* 381 (7): 668–76. https://doi.org/10.1056/NEJMsr1809937.

The GTEx Consortium, Kristin G. Ardlie, David S. Deluca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, et al. 2015. “The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348 (6235): 648–60. https://doi.org/10.1126/science.1262110.

The H3Africa Consortium, Enock Matovu, Bruno Bucheton, John Chisi, John Enyaru, Christiane Hertz-Fowler, Mathurin Koffi, et al. 2014. “Enabling the Genomic Revolution in Africa.” *Science* 344 (6190): 1346–48. https://doi.org/10.1126/science.1251546.

Tishkoff, Sarah A., and Scott M. Williams. 2002. “Genetic Analysis of African Populations: Human Evolution and Complex Disease.” *Nature Reviews Genetics* 3 (8): 611–21. https://doi.org/10.1038/nrg865.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. “Genome-Wide Association Studies.” *Nature Reviews Methods Primers* 1 (1): 1–21. https://doi.org/10.1038/s43586-021-00056-9.

Urbut, Sarah M., Gao Wang, Peter Carbonetto, and Matthew Stephens. 2019. “Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions.” *Nature Genetics* 51 (1): 187–95. https://doi.org/10.1038/s41588-018-0268-8.

Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. “Heritability in the Genomics Era — Concepts and Misconceptions.” *Nature Reviews Genetics* 9 (4): 255–66. https://doi.org/10.1038/nrg2322.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. “10 Years of GWAS Discovery: Biology, Function, and Translation.” *The American Journal of Human Genetics* 101 (1): 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. “Five Years of GWAS Discovery.” *The American Journal of Human Genetics* 90 (1): 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029.

Vries, Paul S de, Michael R Brown, Amy R Bentley, Yun J Sung, Thomas W Winkler, Ioanna Ntalla, Karen Schwander, et al. 2019. “Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions.” *American Journal of Epidemiology* 188 (6): 1033–54. https://doi.org/10.1093/aje/kwz005.

Wainberg, Michael, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, et al. 2019. “Opportunities and Challenges for Transcriptome-Wide Association Studies.” *Nature Genetics* 51 (4): 592–99. https://doi.org/10.1038/s41588-019-0385-z.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, et al. 2014. “The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations.” *Nucleic Acids Research* 42 (D1): D1001–6. https://doi.org/10.1093/nar/gkt1229.

Westra, Harm-Jan, and Lude Franke. 2014. “From Genome to Function by Studying EQTLs.” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, From genome to function, 1842 (10): 1896–1902. https://doi.org/10.1016/j.bbadis.2014.04.024.

Willer, Cristen J, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, et al. 2013. “Discovery and Refinement of Loci Associated with Lipid Levels.” *Nature Genetics* 45 (11): 1274–83. https://doi.org/10.1038/ng.2797.

Wise, Anastasia L., Lin Gyi, and Teri A. Manolio. 2013. “EXclusion: Toward Integrating the X Chromosome in Genome-Wide Association Analyses.” *The American Journal of Human Genetics* 92 (5): 643–47. https://doi.org/10.1016/j.ajhg.2013.03.017.

Wojcik, Genevieve L., Mariaelisa Graff, Katherine K. Nishimura, Ran Tao, Jeffrey Haessler, Christopher R. Gignoux, Heather M. Highland, et al. 2019. “Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits.” *Nature* 570 (7762): 514–18. https://doi.org/10.1038/s41586-019-1310-4.

Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, et al. 2010. “Common SNPs Explain a Large Proportion of the Heritability for Human Height.” *Nature Genetics* 42 (7): 565–69. https://doi.org/10.1038/ng.608.

Zhu, Jinhang, Geng Chen, Sibo Zhu, Suqing Li, Zhuo Wen, Bin Li, Yuanting Zheng, and Leming Shi. 2016. “Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-Seq.” *Scientific Reports* 6 (1): 28400. https://doi.org/10.1038/srep28400.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

VITA

Daniel S. Araujo was born in Contagem, Minas Gerais, Brazil. After obtaining his Bachelor of Science degree in Biological Sciences at the Federal University of Minas Gerais, Brazil, in August of 2020, he sought to continue his training in the field of genetics. Araujo was granted the Graduate Research Fellowship at Loyola University Chicago and joined the Master of Science Bioinformatics program in August 2020 to work under Dr. Heather Wheeler's supervision. He completed the program in May, 2023. Going forward, Araujo will pursue a Doctor of Philosophy degree in Human Genetics at The University of Chicago, starting Fall 2023.