Genetic architecture of gene expression traits across diverse populations

Lauren S. Mogil[1], Angela Andaleon[1,2], Alexa Badalamenti[2], Scott P. Dickinson[3], Xiuqing Guo[4], Jerome I. Rotter[4], W. Craig Johnson[5], Hae Kyung Im[3], Yongmei Liu[6], Heather E. Wheeler[*1,2,7,8]

1 Department of Biology, Loyola University Chicago, Chicago, IL, USA
2 Program in Bioinformatics, Loyola University Chicago, Chicago, IL, USA
3 Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA
4 Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute and Department of Pediatrics at Harbor-UCLA Medical Center, Torrance, CA.
5 Department of Biostatistics, University of Washington, Seattle, WA, USA
6 Department of Epidemiology & Prevention, Wake Forest School of Medicine, Winston-Salem, NC, USA
7 Department of Computer Science, Loyola University Chicago, Chicago, IL, USA
8 Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, IL, USA

* hwheeler1@luc.edu

Abstract

For many complex traits, gene regulation is likely to play a crucial mechanistic role. How the genetic architectures of complex traits vary between populations and subsequent effects on genetic prediction are not well understood, in part due to the historical paucity of GWAS in populations of non-European ancestry. We used data from the MESA (Multi-Ethnic Study of Atherosclerosis) cohort to characterize the genetic architecture of gene expression within and between diverse populations. Genotype and monocyte gene expression were available in individuals with African American (AFA, n=233), Hispanic (HIS, n=352), and European (CAU, n=578) ancestry. We performed expression quantitative trait loci (eQTL) mapping in each population and show genetic correlation of gene expression depends on share ancestry proportions. Using elastic net modeling with cross validation to optimize genotypic predictors of gene expression in each population, we show the genetic architecture of gene expression is sparse across populations. We found the best predicted gene, *HLA-DRB5*, was the same across populations with $R^2 > 0.81$ in each population. However, there were 1094 (11.3%) well predicted genes in AFA and 372 (3.8%) well predicted genes in HIS that were poorly predicted in CAU. Using genotype weights trained in MESA to predict gene expression in 1000 Genomes populations showed that a training set with ancestry similar to the test set is better at predicting gene expression in test populations, demonstrating an urgent need for diverse population sampling in genomics. Our predictive models in diverse cohorts are made publicly available for use in transcriptome mapping methods at `http://predictdb.hakyimlab.org/`.

Author summary

Most genome-wide association studies (GWAS) have been conducted in populations of European ancestry leading to a disparity in understanding the genetics of complex traits between populations. For many complex traits, gene regulation is likely to play a critical mechanistic role given the consistent enrichment of regulatory variants among trait-associated variants. However, it is still unknown how the effects of these key variants differ across populations. We used data from MESA to study the underlying genetic architecture of gene expression by optimizing gene expression prediction within and across diverse populations. The populations with genotype and gene expression data available are from individuals with African American (AFA, n=233), Hispanic (HIS, n=352), and European (CAU, n=578) ancestry. After calculating the prediction performance, we found that there are many genes that were well predicted in AFA and HIS that were poorly predicted in CAU. We further showed that a training set with ancestry similar to the test set resulted in better gene expression predictions, demonstrating the need to incorporate diverse populations in genomic studies. Our gene expression prediction models are publicly available to facilitate future transcriptome mapping studies in diverse populations.

Introduction

For over a decade, genome-wide association studies (GWAS) have facilitated the discovery of thousands of genetic variants associated with complex traits and new insights into the biology of these traits [1]. Most of these studies involved individuals of primarily European descent, which can lead to disparities when attempting to apply this information across populations [2–4]. Continued increases in GWAS sample sizes and new integrative methods will lead to more clinically relevant and applicable results. Non-European populations need to be included in these studies to avoid further contribution to health care disparities [5]. A recent study shows that the lack of diversity in large GWAS skew the prediction accuracy across non-European populations [6]. This discrepancy in predictive accuracy demonstrates that adding ethnically diverse populations is critical for the success of precision medicine, genetic research, and understanding the biology behind genetic variation [6–8].

Gene regulation is likely to play a critical role for many complex traits as trait-associated variants are enriched in regulatory, not protein-coding, regions [9–13]. Numerous expression quantitative trait loci (eQTL) studies have provided insight into how genetic variation affects gene expression [14–17]. While eQTL can act at a great distance, or in *trans*, the largest effect sizes are consistently found near the transcription start sites of genes [14–17]. Because gene expression shows a more sparse genetic architecture than many other complex traits, gene expression is amenable to genetic prediction with relatively modest sample sizes [18, 19]. This has led to new mechanistic methods for gene mapping that integrate transcriptome prediction, including PrediXcan [20] and TWAS [21]. These methods have provided useful tools for understanding the genetics of complex traits; however, most of the models have been built using predominantly European populations.

How the key variants involved in gene regulation differ among populations has not been fully explored. While the vast majority of eQTL mapping studies have been performed in populations of European descent, increasing numbers of transcriptome studies in non-European populations make the necessary comparisons between populations feasible [14, 22, 23]. An eQTL study across eight diverse HapMap populations (∼100 individuals/population) showed that the directions of effect sizes were usually consistent when an eQTL was present in two populations [14]. However, the impact of a particular genetic variant on population gene expression differentiation is also dependent on allele frequencies, which often vary between populations. A better understanding of the degree of transferability of gene expression prediction models across populations is essential for broad application of methods like PrediXcan in the study of the genetic architecture of complex diseases and traits in diverse populations.

Here, in order to better define the genetic architecture of gene expression across populations, we combine genotype [24] and monocyte gene expression [25] data from the Multi-Ethnic Study of Atherosclerosis (MESA) for the first time. We perform eQTL mapping and optimize multi-SNP predictors of gene expression in three diverse populations. The MESA populations studied herein comprise 233 African American (AFA), 352 Hispanic (HIS), and 578 European (CAU) self-reported ancestry individuals. Using elastic net regularization and Bayesian sparse linear mixed modeling, we show sparse models outperform polygenic models in each population. We show the genetic correlation of SNP effects and the predictive performance correlation is highest between populations with the most overlapping admixture proportions. We found a subset of genes that are well predicted in the AFA and/or HIS cohorts that are poorly predicted, if predicted at all, in the CAU cohort. We also test our predictive models trained in MESA cohorts in independent cohorts from the HapMap Project [14] and show the correlation between predicted and observed gene expression is highest when the ancestry of the test set is similar to that of the training set. By diversifying our model-building populations, new genes may be implicated in complex trait mapping studies that were not previously interrogated. Models built here have been added to PredictDB `http://predictdb.hakyimlab.org/` for use in PrediXcan [20] and other studies.

Results

Common and unique eQTLs across populations in MESA

We surveyed each MESA population (AFA, HIS, CAU) and two combined populations (AFHI, ALL) for cis-eQTLs. SNPs within 1Mb of each of 10,143 genes were tested for association with monocyte gene expression levels using a linear additive model. We used 10 genotype principal components in each model (Fig. 1) and compared models that included a range of PEER factors (0, 10, 20, 30, 50, 100) to adjust for hidden confounders in the expression data [26]. As expected, the sample size of the data influences the number of eQTLs mapped (Fig. 2A). We found that using at least 20 PEER factors was best at finding the optimal number of eQTLs with a FDR < 0.05 for each population (Fig. 2A). For the remainder of this work, all models were adjusted for 10 genotype principal components and 20 PEER factors. Hundreds of thousands to millions of SNPs were found to associate with gene expression (eSNPs) and most genes had at least one associated variant (eGenes) at FDR < 0.05 (Table 1). We quantified the number of eSNPS and eGenes as well as the percentage of common and unique eSNPs found for each population. Common eSNPs met FDR $< .05$ in all three self-identified populations (AFA, HIS, CAU) or, in the case of the combined AFHI population, common eSNPs met FDR < 0.05 in both AFHI and CAU. Unique eSNPs met FDR < 0.05 in only the designated population. While the AFA population has a sample size of less than half of the CAU population, the two populations have a similar proportion of unique eSNPs (Table 1). SNPs discovered in the CAU population were less likely to be replicated in the other populations than those discovered in the AFA population (Fig. 2B).

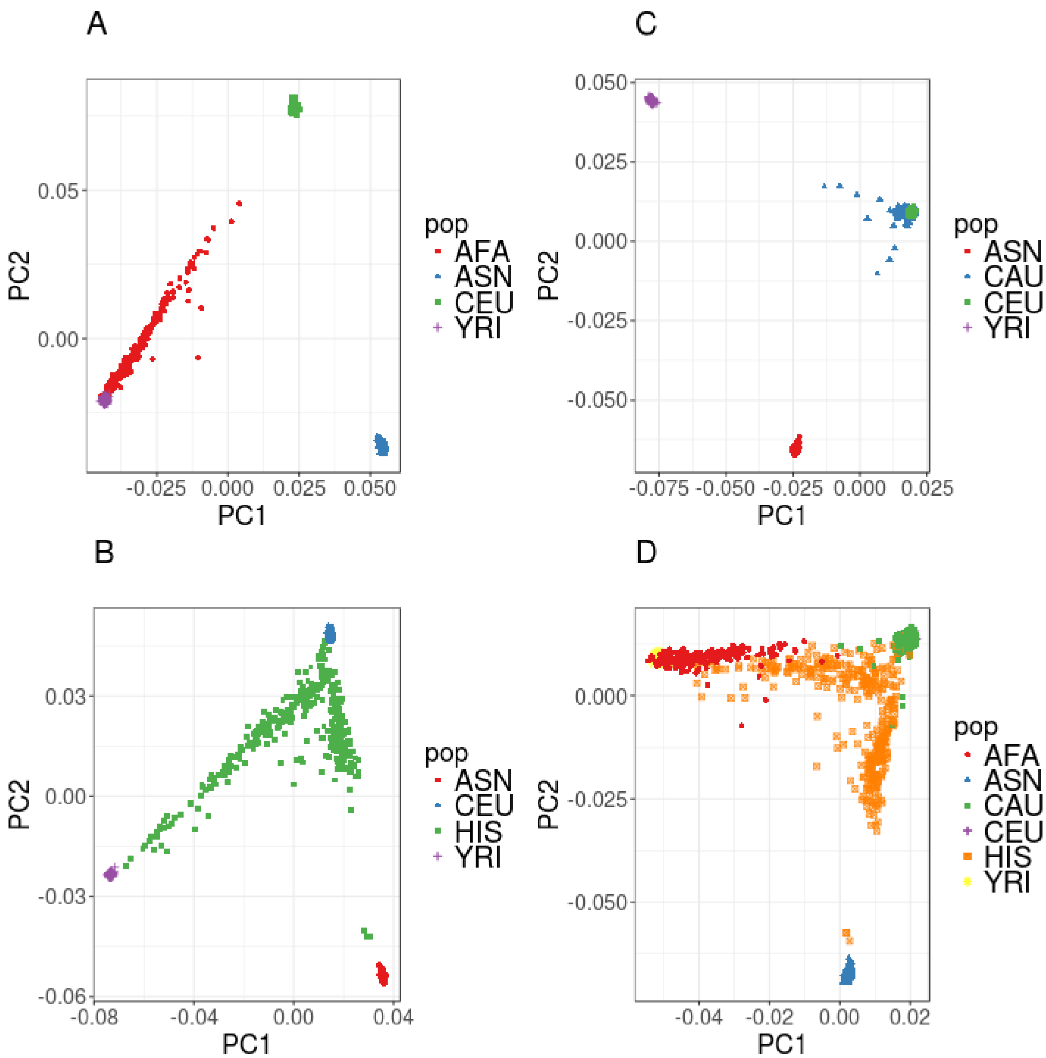


Fig 1. Genotype principal component (PC) analysis of MESA populations. PC1 vs. PC2 plots of each MESA population when analyzed with HapMap populations show varying degrees of admixture. The HapMap populations are defined by the following abbreviations: Yoruba from Ibadan, Nigeria (YRI), European ancestry from Utah (CEU), East Asians from Beijing, China and Tokyo, Japan (ASN). (**A**) MESA AFA population (red), (**B**) MESA HIS population (green), (**C**) MESA CAU population (blue), (**D**) all MESA populations combined.

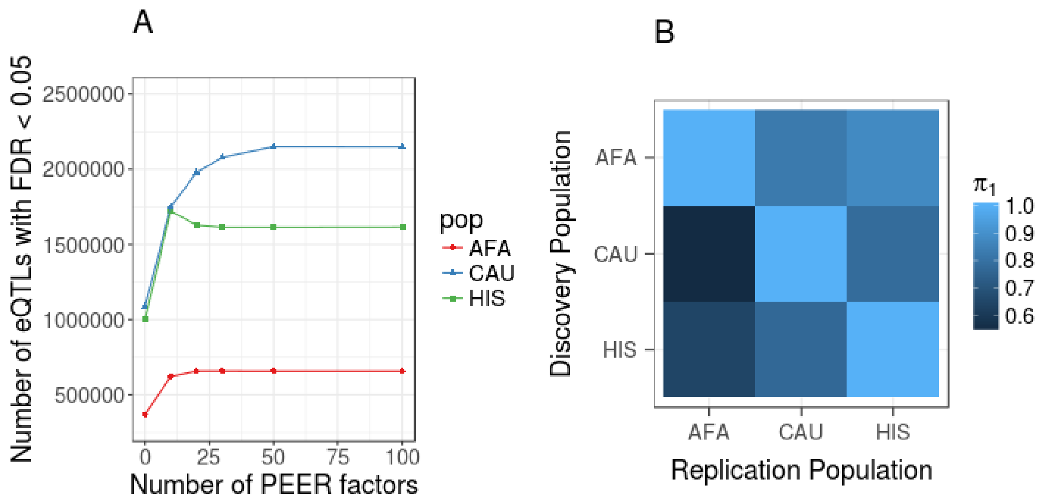


Fig 2. Summary of eQTL analyses in MESA populations (**A**) The number of eQTLs with FDR < 0.05 increases when accounting for at least 20 PEER factors in each population.(**B**) π_1 statistics [27] for cis-eQTLs are reported for all pairwise combinations of discovery (y-axis) and replication (x-axis) populations. Higher π_1 values indicate a stronger replication signal. π_1 is calculated when the SNP from the discovery population is present in the replication population.

Table 1. cis-eQTL (FDR $<$ 0.05) characteristics across MESA populations

Population	number eSNPs	number eGenes	common SNPs	unique SNPs
AFA (n=233)	657,185	7559	41%	38%
HIS (n=352)	1,628,344	8621	26%	33%
CAU (n=578)	1,977,647	8602	25%	39%
AFHI (n=585)	2,008,900	9074	35%	22%
ALL (n=1163)	3,051,709	9393	NA	NA

Linear additive models were adjusted for 10 genotype principal components and 20 PEER factors. FDR = Benjamini-Hochberg false discovery rate. AFA = African American, HIS = Hispanic, CAU = European American, AFHI = AFA and HIS, ALL = AFA, HIS, and CAU.

Pairwise comparison between populations show CAU and HIS are the most correlated

We estimated the local heritability (h^2) for each gene and the genetic correlation (rG) between genes in each MESA population using GCTA [28]. The sample sizes are not large enough to estimate genetic correlation for individual genes, but since there are a large number of genes, we can estimate the mean rG across genes [29]. The population pair with the highest mean rG was CAU and HIS, followed by AFA and HIS, and the least correlated pair was AFA and CAU (Table 2, Fig. 3). As the heritability threshold within a population increase, the mean rG between populations also increases (Fig. 3B).

Sparse models outperform polygenic models for gene expression

We examined the prediction performance of a range of models using elastic net regularization [30] to characterize the genetic architecture of gene expression in each population. The mixing parameter that gives the largest prediction performance R^2

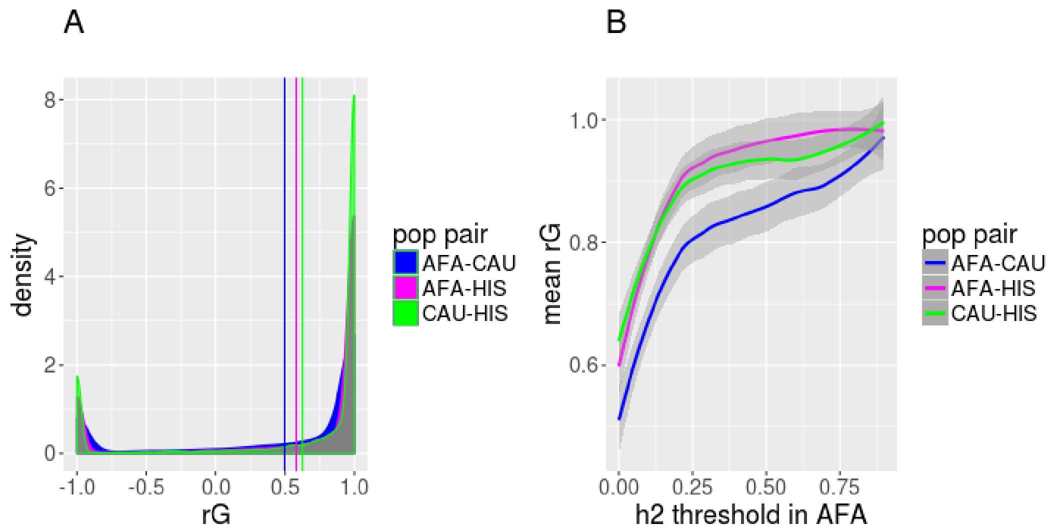


Fig 3. Genetic correlation (rG) of gene expression between MESA populations.
(**A**) Distribution of genetic correlation (rG) between populations. The vertical lines represent the mean rG across genes for the population pair. The most correlated populations are CAU and HIS and the least correlated populations are AFA and CAU. (**B**) Comparison of the genetic correlation between pairwise MESA populations and the subset of genes with heritability (h^2) greater than a given threshold in the AFA population.

Table 2. Genetic correlation (rG) between MESA populations

pop pair	mean rG	SE rG	genes that converged
AFA-CAU	0.48	0.0080	9227
AFA-HIS	0.57	0.0076	9269
CAU-HIS	0.62	0.0071	9480

rG was estimated using a bivariate restricted maximum likelihood (REML) model implemented in GCTA.

indicates the degree of sparsity or polygenicity of the gene expression trait. If the highest R^2 occurs when $\alpha = 0.05$, then the gene expression trait exhibits a more polygenic architecture. However, if the optimal R^2 occurs when $\alpha = 1$ then the trait has a sparse architecture [18]. We performed 10-fold cross-validation across three mixing parameters ($\alpha = 0.05, 0.5, 1$). We found that the highest R^2 predictive performance occurred when $\alpha = 0.5$ or $\alpha = 1$, whereas the R^2 was smaller when $\alpha = 0.05$, indicating that the sparse model outperformed the polygenic model. Figure 4 shows that models with 0.5 and 1 had similar predictive power while an $\alpha = 0.05$ was suboptimal for gene expression prediction in each of the populations. The number of genes that converged when $\alpha = 0.5$ was 9695 for each population.

In addition to elastic net, we also used Bayesian Sparse Linear Mixed Modeling (BSLMM) [31] to estimate if the local genetic contribution to gene expression is more polygenic or sparse. This approach models the genetic contribution of the trait as the sum of a sparse component and a polygenic component. The parameter PGE represents the proportion of the genetic variance explained by sparse effects. We also estimated heritability (h^2) using GCTA, a linear mixed model approach [28]. The PVE is the BSLMM equivalent of h^2 that is estimated from GCTA. We found that BSLMM PVE,

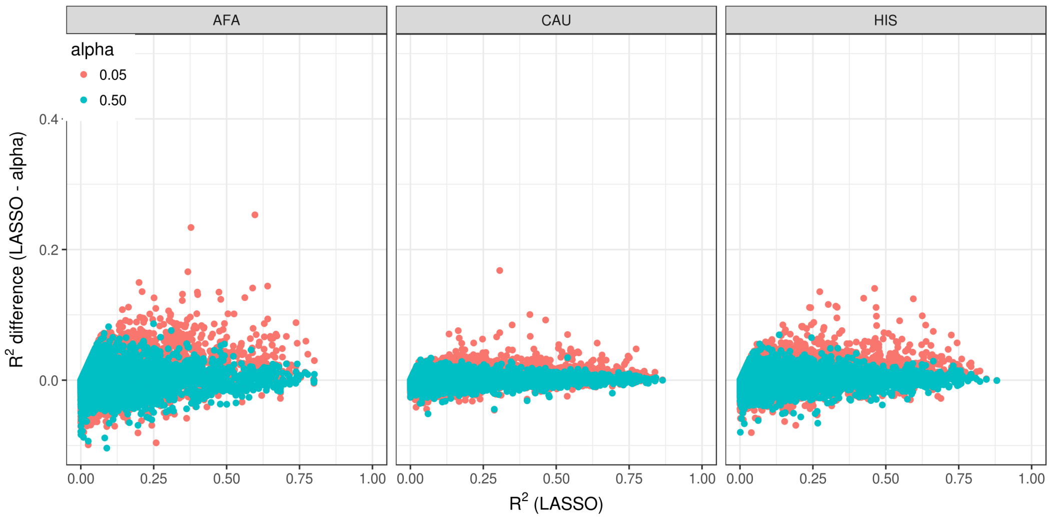


Fig 4. MESA cross-validated predictive performance across a range of elastic net mixing parameters. The difference between the 10-fold cross-validated R^2 of LASSO and elastic net mixing parameters 0.05 or 0.5 is compared to the LASSO R^2 across genes in MESA populations AFA, HIS, and CAU. The R^2 difference values with a mixing parameter $\alpha = 0.5$ are close to zero indicating that they perform similarly to the LASSO model. The values with a mixing parameter $\alpha = 0.05$ are above zero indicating that they perform worse than the LASSO model.

GCTA h^2, and elastic net R^2 are highly correlated in each population (S1 Fig). Using BLSMM, we also found that for highly heritable genes, the sparse component (PGE) is large; however, for genes with low PVE, we are unable to determine whether the sparse or polygenic component is predominant (S1 Fig).

A subset of well-predicted genes in AFA and HIS were missed in CAU

We then compared each population's gene expression predictive performance. Higher correlation values indicate similar accuracy in prediction performance of gene expression models between two populations. The correlation between CAU and HIS is highest (R^2=0.853) followed by AFA and HIS (R^2=0.702) and the lowest correlation between two populations was AFA and CAU with R^2=0.678 (Fig. 5A-C). These correlation relationships mirror the European and African admixture proportions in the MESA HIS and AFA cohorts (Fig. 1). There are many genes that are well predicted in both populations and there are some that are poorly predicted between populations. We found the best predicted gene, *HLA-DRB5*, was the same across each population with an $R^2 > 0.81$ in each population. On the other hand, there are some genes that are well predicted in one population, but poorly predicted in the other and vice versa (Fig. 5D-E). There were 1094 (11.3%) well predicted genes in AFA that were poorly predicted in CAU with an R^2 difference greater than 0.2 between AFA and CAU (Table 3). When comparing HIS and CAU, there were 372 (3.8%) well predicted genes in HIS and poorly predicted in CAU with an R^2 difference greater than 0.2. In contrast, a much smaller proportion of genes were well predicted in CAU and poorly predicted in AFA or HIS, 2.8% and 0.61%, respectively (Table 3).

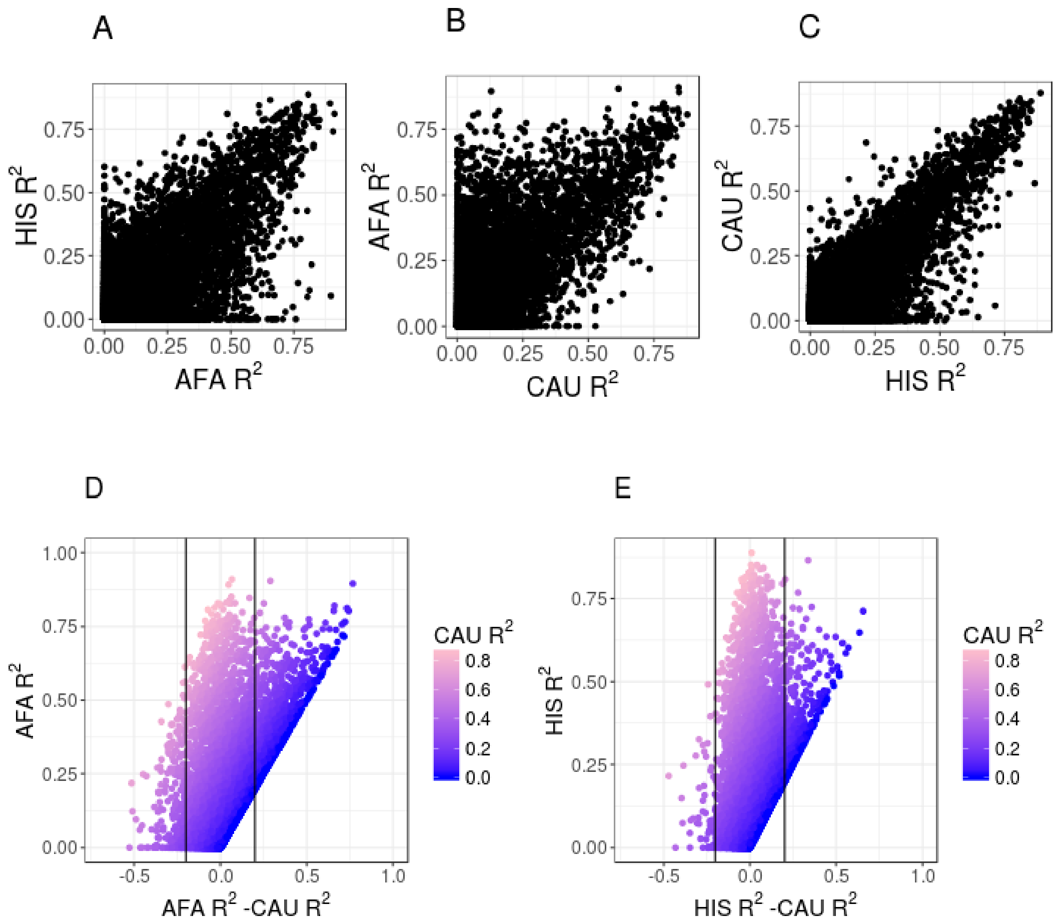


Fig 5. Comparison of predictive performance between populations. The correlation of predictive performance between HIS and AFA (**A**, $R^2 = 0.702$), AFA and CAU (**B**, $R^2 = 0.678$), and CAU and HIS (**C**, $R^2 = 0.853$). The most correlated populations are HIS and CAU and least correlated populations are AFA and CAU. The difference in predictive performance of AFA (**D**) and HIS (**E**) population compared to CAU. Note that there are more genes that are better predicted in AFA and HIS that are not present or poorly predicted in CAU than genes better predicted in CAU that are poorly predicted in the AFA and HIS populations.

Table 3. Comparison of gene expression prediction performance in AFA and HIS compared to CAU

pop pair difference in R^2	diff > 0.2	diff < -0.2	-0.2 < diff < 0.2	total
AFA R^2 − CAU R^2	1094 (11.3%)	276 (2.8%)	8325 (85%)	9695
HIS R^2 − CAU R^2	372 (3.8%)	60 (0.61%)	9263 (95%)	9695

Predictive performance improves when training set has similar ancestry to test set

In order to further compare the predictive performance between populations, using each of the MESA populations as training sets, we predicted gene expression in two populations, Mexican ancestry individuals in Los Angeles (MXL) and Yoruba individuals in Ibadan, Nigeria (YRI), from the HapMap and 1000 Genomes Projects (Table 4, Fig. 6). The mean predicted vs. observed Pearson correlation (R) for YRI was 0.081 when using the AFA population as a training set, while mean R = 0.051 when using the CAU training set (Table 4). The MXL population had a mean R = 0.092 using the HIS population as a training set, whereas the mean R was 0.090 when CAU was the training set (Table 4). The AFA training set is suboptimal across models with varying predictive performance R^2 when tested in MXL (Fig. 6A). Similarly, the CAU training set is suboptimal across models when used to predict expression in YRI (Fig. 6B). When using the currently available DGN training set that consists of 922 European individuals [20], both YRI and MXL are more poorly predicted than when the MESA training sets are used (Table 4). After combining the AFA and HIS population (AFHI), we see that the predicted expression for YRI does better than HIS or AFA alone (Table 4). When all of the MESA populations are combined, the MXL and YRI mean predicted vs. observed correlation is optimized across models (Fig. 6). This demonstrates that when comparing predicted expression levels to the observed, a balance of the training population with ancestry most similar to the test population and total sample size leads to optimal predicted gene expression.

Table 4. Mean predictive performance in independent test cohorts across training models.

Population	AFA (n=233)	HIS (n=352)	CAU (n=578)	AFHI (n=585)	ALL (n=1163)	DGN (n=922)
YRI (n=107)	0.081	0.070	0.051	0.084	0.079	0.032
MXL (n=45)	0.073	0.092	0.090	0.091	0.094	0.053

The mean Pearson correlation (R) of the predicted vs. observed gene expression using MESA (AFA = African American, HIS = Hispanic, CAU = European American, AFHI = AFA and HIS, ALL = AFA, HIS, and CAU) and DGN (Depression Genes and Networks, all European ancestry) as training sets to predict gene expression in HapMap/1000 Genomes populations YRI (Yoruba in Ibadan, Nigeria) and MXL (Mexican ancestry in Los Angeles).

Discussion

We used three MESA populations (AFA, HIS, and CAU) to better understand the genetic architecture of gene expression in diverse populations. We optimized predictors of gene expression using elastic net regularization and found that sparse models outperform polygenic models. The genetic correlation of gene expression is highest when continental ancestry overlaps between populations. We identified genes that are better predicted in the AFA and/or HIS models that are either absent or poorly predicted in the CAU model. We tested our predictors developed in MESA in

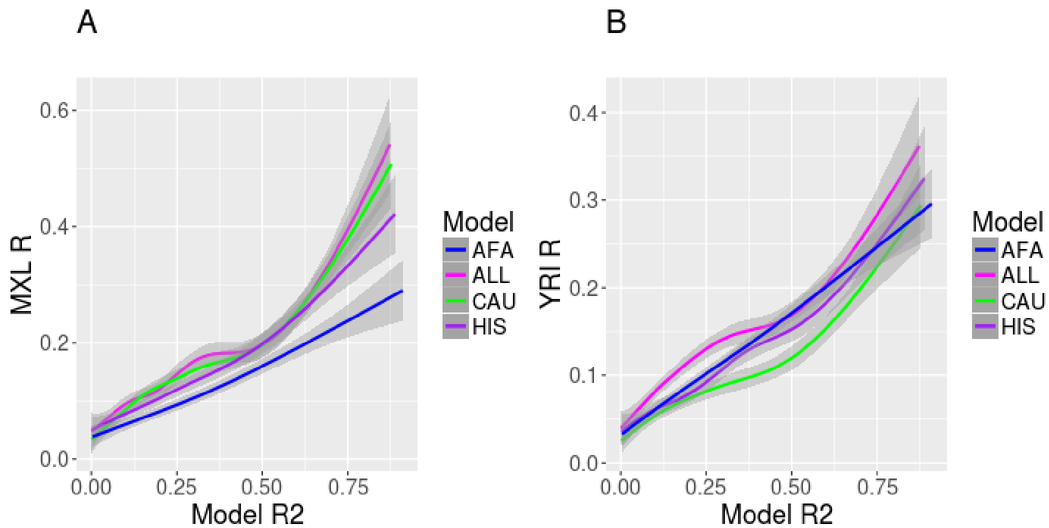


Fig 6. Predictive performance in independent test cohorts across MESA population models.
Loess smoothing lines of the predicted vs. observed gene expression correlation (R) in test HapMap/1000 Genomes cohorts MXL (**A**) and YRI (**B**) compared to the cross-validated predictive performance (R^2) of each prediction model built in the MESA populations.

independent cohorts and found that the best prediction of gene expression occurred when the training set included individuals with similar ancestry to the test set.

As seen in other studies [18, 21, 32], we show sparse models outperform polygenic models for gene expression prediction across diverse populations. Thus, the genetic architecture of gene expression for well predicted genes has a substantial sparse component. Larger sample sizes may reveal an additional polygenic component that may improve prediction for some genes.

We estimated the genetic correlation between each population pair for each gene. Populations with more shared ancestry as defined by clustering of genotypic principal components showed higher mean correlation across genes (Fig. 1, Table 2). As estimated heritability of genes increase, the mean genetic correlation between populations also increases (Fig. 3B), which indicates the genetic architecture underlying gene expression is similar for the most heritable genes. However, even though prediction across populations is possible for some of the most heritable genes, we define a class of genes where predictive performance drops substantially between populations.

There were several genes with high predictive performance ($R^2 > 0.2$) in AFA or HIS that were poorly predicted or not predicted at all in the CAU population (Fig. 5, S1 Table, S2 Table). Of the 372 genes found that were better predicted in HIS, there were 153 genes that overlapped with the 1094 gene found for AFA (S3 Table). Almost all of these well predicted genes in AFA and HIS populations also had biological implications in at least one study in the GWAS Catalog (S4 Table). Examples of such genes include *COMMD1* (ENSG00000147905.13), which has been associated with blood cell volume and elevated iron levels and *ZCCHC7* (ENSG00000173163.6), which has been linked to HIV susceptibility [33–35].

We tested our predictive gene expression models built in the MESA cohorts in two HapMap/1000 Genomes data sets (MXL and YRI) [14, 36] using the MESA population predictors we generated. As expected, the YRI gene expression prediction was best when using the AFA, AFHI, or ALL training sets, which each include individuals with African-ancestry admixture (Table 4, Fig. 6). The best gene expression prediction for

MXL was with the ALL training set, which indicates that admixed populations like MXL benefit from a pooled training set containing individuals of diverse ancestries. Thus, increasing the sample sizes of non-European populations in genomic studies will not only benefit the source population, but will also increase predictive power in admixed populations.

Predictive models of gene expression developed in this study are made publicly available at `http://predictdb.hakyimlab.org/` for use in future studies of complex trait genetics across diverse populations. Inclusion of diverse populations in complex trait genetics is crucial for equitable implementation of precision medicine.

Materials and methods

The Loyola University Chicago Institutional Review Board (IRB) reviewed our application for confirmation of exemption (IRB project number 2014). The IRB determined that this human subject research project is exempt from the IRB oversight requirements according to 45 CFR 46.101.

Genomic and transcriptomic data

The Multi-Ethnic Study of Atherosclerosis (MESA)

MESA includes 6814 individuals consisting of 53% females and 47% males between the ages of 45-84 [24]. The individuals were recruited from 6 sites across the US (Baltimore, MD; Chicago, IL; Forsyth County, NC; Los Angeles County, CA; northern Manhattan, NY; St.Paul, MN). MESA cohort population demographics were 39% Caucasian (CAU), 22% Hispanic (HIS), 28% African American (AFA), and 12% Chinese (CHN). Of those individuals, RNA was collected from CD14+ monocytes from 1264 individuals across three populations (AFA, HIS, CAU) and quantified on the Illumina Ref-8 BeadChip [25]. Individuals with both genotype (dbGaP: phs000209.v13.p3) and expression data (GEO: GSE56045) included 234 AFA, 386 HIS, and 582 CAU. Illumina IDs were converted to Ensembl IDs using the RefSeq IDs from MESA and gencode.v18 (gtf and metadata files) to match Illumina IDs to Ensembl IDs. If there were multiple Illumina IDs corresponding to an Ensembl ID, the average of those values was used as the expression level.

HapMap and 1000 Genomes data

We obtained genotype data from the 1000 Genomes Project [36] for populations of interest where lymphoblastoid cell line (LCL) gene expression data were also available [14]. Transcriptome data from Stranger et al. [14] included 45 Mexican ancestry individuals in Los Angeles, CA, USA (MXL) and 107 Yoruba individuals in Ibadan, Nigeria (YRI).

Quality control of genomic and transcriptomic data

MESA populations were previously imputed using IMPUTE 2.2.2 using the 1000 Genomes Phase I variant set and NCBI build 37/hg 19 for a final SNP count of at least 39 million variants [24,37,38]. Quality control and cleaning of the genotype data was done using PLINK (`https://www.cog-genomics.org/plink2`). SNPs were filtered by call rates less than 99%. Prior to IBD and principal component analysis (PCA), SNPs were LD pruned by removing 1 SNP in a 50 SNP window if $r^2 > 0.3$. One of a pair of related individuals (IBD > 0.05) were removed. Pruned genotypes were merged with HapMap populations and EIGENSTRAT [39] was used to perform PCA (Fig. 1). Final sample sizes for each population post quality control are AFA = 233, HIS = 352, and CAU = 578 . We used 5-7 million non-LD pruned SNPs per population post quality control. PEER factor analysis was performed on the expression data using the peer R package in order to correct for potential batch effects and experimental confounders [40]. A range of PEER factors (0, 10, 20, 30, 50, and 100) were calculated after 10 genotypic PC adjustment in each population to determine how many were required to maximize eQTL discovery. HapMap genotypes in individuals not sequenced through the 1000 Genomes Project were imputed using the Michigan Imputation Server for a total of 6-13 million SNPs per population, after undergoing PLINK quality control [41]. These imputed samples were then merged back with the individuals that were previously sequenced, filtering the SNPs (imputation $\mathrm{R}^2 > 0.8$, MAF > 0.01, HWE p $>$ 1e-06). HapMap expression data sets were adjusted by ten PEER factors.

eQTL analysis

We used Matrix eQTL [42] to perform a genome-wide cis-eQTL analysis in each population separately (AFA, HIS, CAU), in the AFA and HIS combined (AFHI), and in all three populations combined (ALL). We used SNPs with MAF > 0.05 and defined cis-acting as SNPs within 1 Mb of the transcription start site (TSS). The linear regression models included 10 genotype principal component covariates and a range of PEER factors (0, 10, 20, 30, 50, or 100) [26]. The false discovery rate (FDR) for each SNP was calculated using the Benjamini-Hochberg procedure. Similar to the approach recently taken by the GTEx Project Consortium to compare tissues, we estimate the pairwise population eQTL replication rates with π_1 statistics ($\pi_1 = 1 - \pi_0$; π_0 is the proportion of false positives) using the qvalue method [17, 27].

Genetic correlation analysis

eQTL effect size comparisons between populations were performed using Genome-wide Complex Trait Analysis (GCTA) software [28]. We performed a bivariate restricted maximum likelihood (REML) analysis to estimate the genetic correlation (rG) between each pair of MESA cohorts for each gene [43]. We also used GCTA to estimate the proportion of variance explained by all cis-region SNPs (local h^2) for each gene in each population using restricted maximum likelihood (REML).

Prediction model optimization

We used the glmnet R package [30] to fit an elastic net model to predict gene expression from cis-region SNP genotypes. The elastic net regularization penalty is controlled by the mixing parameter alpha, which can vary between ridge regression ($\alpha = 0$) and LASSO ($\alpha = 1$, default). We quantified the predictive performance of each model via 10-fold cross-validated Pearson R^2 (predicted vs. observed gene expression). A gene with the optimal predictive performance when $\alpha = 0$ has a polygenic architecture, whereas a gene with optimal performance when $\alpha = 1$ has a sparse genetic architecture. In the MESA cohort we tested three values of the mixing parameter (0.05, 0.5, and 1) for optimal prediction of gene expression of 10,143 genes for each population alone, AFA and HIS combined, and all three populations combined. We used the PredictDB pipeline developed by the Im lab to preprocess, train, and compile elastic net results into database files to use as weights for gene expression prediction. See `https://github.com/hakyimlab/PredictDBPipeline` and `https://github.com/lmogil/run_PredictDB_with_pops`.

We also used the software GEMMA [44] to implement Bayesian Sparse Linear Mixed Modeling (BSLMM) [31] for each gene with 100K sampling steps per gene. BSLMM estimates the PVE (the proportion of variance in phenotype explained by the additive genetic model, analogous to the heritability estimated in GCTA) and PGE (the proportion of genetic variance explained by the sparse effects terms where 0 means that genetic effect is purely polygenic and 1 means that the effect is purely sparse). From the second half of the sampling iterations for each gene, we report the median and the 95% credible sets of the PVE and PGE.

References

1. Welter D, Macarthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations;doi:10.1093/nar/gkt1229.

2. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies;doi:10.1016/j.tig.2009.09.012.

3. Bustamante CD, Burchard EG, De FM, Vega L. Genomics for the world: Medical genomics has focused almost entirely on those of European descent. Other ethnic groups must be studied to ensure that more people benefit, say;doi:10.1038/475163a.

4. Popejoy AB. Genomics is failing on diversity;doi:10.1038/538161a.

5. Oh SS, Galanter J, Thakur N, Pino-Yanes M, Barcelo NE, White MJ, et al. Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. PLOS Medicine. 2015;12(12):e1001918. doi:10.1371/journal.pmed.1001918.

6. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. The American Journal of Human Genetics. 2017;100:635–649. doi:10.1016/j.ajhg.2017.03.004.

7. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations The 1000 Genomes Project e;doi:10.1073/pnas.1019276108.

8. Carlson CS, Matise TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, et al. Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study;doi:10.1371/journal.pbio.1001661.

9. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genetics. 2010;6(4). doi:10.1371/journal.pgen.1000895.

10. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. AJHG. 2014;95:535–552. doi:10.1016/j.ajhg.2014.10.004.

11. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N, et al. Cross-Tissue and Tissue-Specific eQTLs: Partitioning the Heritability of a Complex Trait. The American Journal of Human Genetics. 2014;95(5):521–534. doi:10.1016/j.ajhg.2014.10.001.

12. Davis LK, Yu D, Keenan CL, Gamazon ER, Konkashbaev AI, Derks EM, et al. Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. PLoS genetics. 2013;9(10):e1003864. doi:10.1371/journal.pgen.1003864.
13. Li YI, Van De Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease;doi:10.1126/science.aad9417.
14. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of Cis Regulatory Variation in Diverse Human Populations. PLoS Genet. 2012;8(4). doi:10.1371/journal.pgen.1002639.
15. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome research. 2014;24(1):14–24. doi:10.1101/gr.155192.113.
16. Kelly DE, Hansen MEB, Tishkoff SA. Global variation in gene expression and the value of diverse sampling. Current Opinion in Systems Biology. 2017;doi:10.1016/j.coisb.2016.12.018.
17. Consortium G, Analysts: L, Laboratory DA&CCL, program Management: NIH, Collection: B, Pathology:, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204–213.
18. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Cox NJ, et al. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLOS Genetics. 2016;12(11):e1006423. doi:10.1371/journal.pgen.1006423.
19. Manor O, Segal E. Robust prediction of expression differences among human individuals using only genotype information. PLoS genetics. 2013;9(3):e1003396. doi:10.1371/journal.pgen.1003396.
20. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015;47(9):1091–1098. doi:10.1038/ng.3367.
21. Gusev A, Ko A, Shi H, Bhatia G, Chung W, J H Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies HHS Public Access. Nat Genet. 2016;48(3):245–252. doi:10.1038/ng.3506.
22. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics. 2015;doi:10.1038/nrg3891.
23. Sajuthi SP, Sharma NK, Chou JW, Palmer ND, McWilliams DR, Beal J, et al. Mapping adipose and muscle tissue expression quantitative trait loci in African Americans to identify genes for type 2 diabetes and obesity. Human genetics. 2016;135(8):869–80. doi:10.1007/s00439-016-1680-8.
24. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. American journal of epidemiology. 2002;156(9):871–81. doi:10.1093/AJE/KWF113.
25. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De la Fuente A, et al. Methylomics of gene expression in human monocytes. Human Molecular Genetics. 2013;22(24):5065–5074. doi:10.1093/hmg/ddt356.

26. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses;doi:10.1038/nprot.2011.457.

27. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003;100(16):9440–9445. doi:10.1073/pnas.1530509100.

28. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. 2011;doi:10.1016/j.ajhg.2010.11.011.

29. Brown BC, Jimmie Ye C, Price AL, Zaitlen N. Transethnic Genetic-Correlation Estimates from Summary Statistics. 2016;doi:10.1016/j.ajhg.2016.05.001.

30. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1–22. doi:10.18637/jss.v033.i01.

31. Zhou X, Carbonetto P, Stephens M, Visscher PM. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet. 2013;9(2).

32. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics. 2016;doi:10.1038/ng.3538.

33. Raffield LM, Louie T, Sofer T, Jain D, Ipp E, Taylor KD, et al. Genome-wide association study of iron traits and relation to diabetes in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL): potential genomic intersection of iron and glucose regulation? Human Molecular Genetics. 2017;26(10):1966–1978. doi:10.1093/hmg/ddx082.

34. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167(5):1415–1429.e19. doi:10.1016/j.cell.2016.10.042.

35. Johnson EO, Hancock DB, Gaddis NC, Levy JL, Page G, Novak SP, et al. Novel genetic locus implicated for HIV-1 acquisition with putative regulatory links to HIV replication and infectivity: a genome-wide association study. PloS one. 2015;10(3):e0118149. doi:10.1371/journal.pone.0118149.

36. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393.

37. Loth DW, Soler Artigas M, Gharib SA, Wain LV, Franceschini N, Koch B, et al. Genome-wide association analysis identifies six new loci associated with forced vital capacity. Nature genetics. 2014;46(7):669–677. doi:10.1038/ng.3011.

38. Wang Z, Manichukal A, Goff DC, Mora S, Ordovas JM, Pajewski NM, et al. Genetic associations with lipoprotein subfraction measures differ by ethnicity in the multi-ethnic study of atherosclerosis (MESA). Human Genetics. 2017;136(6):715–726. doi:10.1007/s00439-017-1782-y.

39. Price AL, j patterson N, m plenge R, e weinblatt M, a shadick N. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–909. doi:10.1038/ng1847.

40. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature protocols. 2012;7(3):500–7. doi:10.1038/nprot.2011.457.

41. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016;48(10):1284–1287. doi:10.1038/ng.3656.

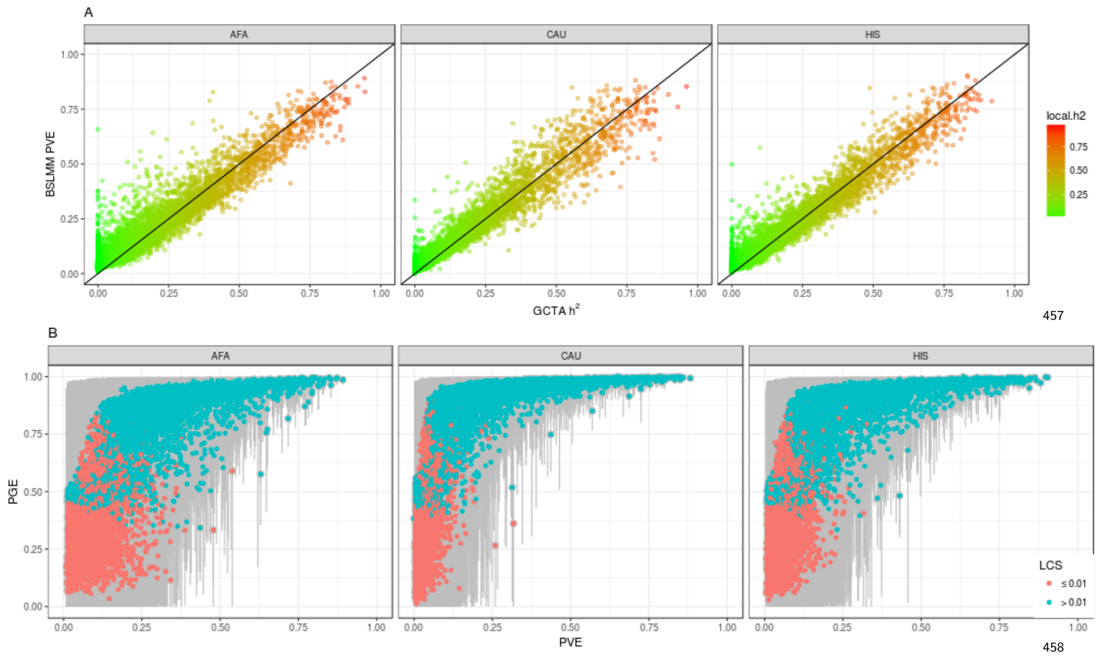
42. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics (Oxford, England). 2012;28(10):1353–8. doi:10.1093/bioinformatics/bts163.

43. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR, Barrett J. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. 2012;28(19):2540–2542. doi:10.1093/bioinformatics/bts474.

44. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature genetics. 2012;44(7):821–4. doi:10.1038/ng.2310.

Supporting information

S1 Fig. Sparsity estimates using Bayesian Sparse Linear Mixed Models in MESA populations.(**A**) This shows the heritability estimate of BSLMM vs GCTA in AFA, HIS, CAU respectively. The majority of genes have similar heritability estimates and are colored by the elastic net R^2. The genes with high heritability tend to have the highest prediction performance R^2. There are some genes that have better heritiability estimates using BSLMM. (**B**) This shows the sparsity of gene expression traits examining the PGE parameter of BSLMM approach of AFA, HIS, and CAU respectively. PGE is the parameter that represents the proportion of the sparse component of the total variance explained by genetic variance and PVE is the BSLMM equivalent of h^2. The highly heritable genes have a sparse component thats close to 1 and therefore the local genetic architecture is sparse. There is not enough evidence to determine if the lower heritablility genes are more sparse or polygenic.



S1 Table. Genes better predicted in AFA than CAU.

S2 Table. Genes better predicted in HIS than CAU.

S3 Table. Overlap genes better predicted in both AFA and HIS than CAU.

S4 Table. GWAS catalog information on genes better predicted in AFA and HIS than CAU.

Acknowledgments

This work is supported by the NIH National Human Genome Research Institute Academic Research Enhancement Award R15 HG009569 (HEW), start-up funds from Loyola University Chicago (HEW), the Loyola Carbon Undergraduate Research Fellowship (AA), the Loyola Biology Summer Research Fellowship (AB), and the Loyola Mulcahy Scholars Program (AB). MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. The MESA Epigenomics & Transcriptomics Study was funded by NIA grant 1R01HL101250-01 to Wake Forest University Health Sciences (YL). DGN gene expression prediction models were obtained from PredictDB at `http://predictdb.hakyimlab.org/`.