TITLE

On the cross-population portability of gene expression prediction models

AUTHORS

Kevin L. Keys[1,2,*], Angel C.Y. Mak[1], Marquitta J. White[1], Walter L. Eckalbar[1], Andy Dahl[1], Joel Mefford[1], Anna V. Mikaylova[3], María G. Contreras[1,4], Jennifer R. Elhawary[1], Celeste Eng[1], Donglei Hu[1], Scott Huntsman[1], Sam S. Oh[1], Sandra Salazar[1], Michael A. Lenoir[5], Jimmie C. Ye[6,7], Timothy A. Thornton[3], Noah Zaitlen[8], Esteban G. Burchard[1,7,+], and Christopher R. Gignoux[9,10,+*].

[1] Department of Medicine, University of California, San Francisco, CA, USA
[2] Berkeley Institute for Data Science, University of California, Berkeley, California, USA
[3] Department of Biostatistics, University of Washington, Seattle, WA, USA
[4] San Francisco State University, San Francisco, CA, USA
[5] Bay Area Pediatrics, Oakland, CA, USA
[6] Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA
[7] Department of Bioengineering and Therapeutic Biosciences, University of California, San Francisco, CA, USA
[8] Department of Neurology, University of California, Los Angeles, CA, USA
[9] Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA
[10] Department of Biostatistics and Informatics, School of Public Health, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA

[+] Shared senior authorship.
[*] Correspondence should be addressed to Kevin L. Keys (klkeys@g.ucla.edu) and Christopher R. Gignoux (chris.gignoux@ucdenver.edu)

ABSTRACT

The genetic control of gene expression is a core component of human physiology. For the past several years, transcriptome-wide association studies have leveraged large datasets of linked genotype and RNA sequencing information to create a powerful gene-based test of association that has been used in dozens of studies. While numerous discoveries have been made, the populations in the training data are overwhelmingly of European descent, and little is known about the portability of these models to other populations. Here, we test for cross-population portability of gene expression prediction models using a dataset of African American individuals with RNA-Seq data in whole blood. We find that the default models trained in large datasets such as GTEx and DGN fare poorly in African Americans, with a notable reduction in prediction accuracy when compared to European Americans. We replicate these limitations in cross-population portability using the five populations in the GEUVADIS dataset. Via simulations of both populations and gene expression, we show that accurate cross-population portability of transcriptome imputation only arises when eQTL architecture is substantially shared across

populations. In contrast, models with non-identical eQTL showed patterns similar to real-world data. Therefore, generating RNA-Seq data in diverse populations is a critical step towards multi-ethnic utility of gene expression imputation.

KEYWORDS

TWAS, gene expression, admixed populations, GTEx, PrediXcan

MANUSCRIPT TEXT

In the last decade, large-scale genome-wide genotyping projects have enabled a revolution in our understanding of complex traits.[1–4] This explosion of genome sequencing data has spurred the development of new methods that integrate large genotype sets with additional molecular measurements such as gene expression. A recently popular integrative approach to genetic association analyses, known as a transcriptome-wide association study (TWAS)[5,6], leverages reference datasets such as the Genotype-Tissue Expression (GTEx) repository[7] or the Depression and Genes Network (DGN)[8] to link associated genetic variants with a molecular trait like gene expression. The general TWAS framework requires previously estimated *cis*-eQTL for all genes in a dataset with both genotype and gene expression measurements. The resulting eQTL effect sizes build a predictive model that can impute gene expression in an independently genotyped population. A TWAS is similar in spirit to the widely-known genome-wide association study (GWAS) but suffers less of a multiple testing burden and can potentially detect more associations as a result.[5,6]

Unlike a normal GWAS, where phenotypes are regressed onto genotypes, in TWAS the phenotype is regressed onto the imputed gene expression values, thus constituting a new gene-based association test. TWAS can also link phenotypes to variation in gene expression and provide researchers with additional biological and functional insights over those afforded by GWAS alone. While these models are imperfect predictors, imputing gene expression allows researchers to test phenotype associations to expression levels in existing GWAS datasets without measuring gene expression directly. In particular, these methods enable analysis of predicted gene expression in very large cohorts (~10^4 – 10^6 individuals) rather than typical gene expression studies that measure expression directly (~10^2 – 10^3 individuals). Several methods have been recently developed to perform TWAS in existing genotyped datasets. PrediXcan[6] uses eQTL precomputed from paired genotype-expression data, such as those in GTEx, in conjunction with a new genotype set to predict gene expression. These gene expression prediction models are freely available online (PredictDB). Related TWAS approaches, such as FUSION[5], MetaXcan[9], or SMR[10], leverage eQTL with GWAS summary statistics instead of requiring the availability of raw individual-level genotype data.

As evidenced by application to numerous disease domains, the TWAS framework is capable of uncovering new genic associations.[11–17] However, the power of TWAS is inherently limited by the data used for eQTL discovery. For example, since gene expression varies by tissue type, researchers must ensure that the prediction weights are estimated using RNA from a tissue related to their phenotype, whether that be the direct tissue of interest or one with sufficiently

correlated gene expression.[18] Furthermore, the ability of predictive models to impute gene expression from genotypes is limited by the heritability in the cis region around the gene.[6] Consequently, genes with little or no measurable genetically regulated effect on their expression in the discovery data would not be good candidates for TWAS.

A subtler but more troubling issue arises from the lack of genetic diversity present in the datasets used for predictive model training: most paired genotype-expression datasets consist almost entirely of data from European-descent individuals. The European overrepresentation in genetic studies is well documented[19–21] and has severe negative consequences for equity as well as for gene discovery[22], fine mapping[23–25], and applications in personalized medicine.[26–34] Genetic architecture and genotype frequencies can vary across populations, which presents a potential problem for the application of predictive models with genotype predictors across multiple populations.

The training data for most models in PredictDB are highly biased toward European ancestry: GTEx version v6p subjects are over 85% European, while the GTEx v7 and DGN subjects are entirely of European descent. The lack of suitable genotype-expression datasets in non-European individuals leads to scenarios in which PredictDB models trained in Europeans are used to impute into non-European or admixed populations. As shown previously in the context of polygenic risk scores[35], multi-SNP prediction models trained in one population can suffer from unpredictable bias and poor prediction accuracy that impair their cross-population portability. Recent analyses of genotype-expression data from the Multi-Ethnic Study of Atherosclerosis (MESA)[36–38] explore cross-population transcriptome imputation and conclude that predictive accuracy is highest when training and testing populations match in ancestry. These results dovetail with our experience analyzing diverse populations, but offer little insight into the mechanisms underlying the cross-population portability of transcriptome prediction models, particularly when eQTL architecture is known.

Here, we investigate the cross-population portability of gene expression models using paired genotype and gene expression data and using simulations derived from real genotypic data and realistic models of gene expression. We analyze prediction quality from currently available PrediXcan prediction weights using a pilot subset of paired genotype and whole blood transcriptome data from the Study of African Americans, Asthma, Genes, and Environment (SAGE).[39–42] SAGE is a pediatric cohort study of childhood-onset asthma and pulmonary phenotypes in African American subjects of 8 to 21 years of age. To tease apart cross-population prediction quality, we turn to GEUVADIS and the 1000 Genomes Project datasets.[4,43] The GEUVADIS dataset has been used extensively to validate PrediXcan models.[6,38] However, recent analyses suggest that GTEx and DGN PrediXcan models behave differently on the constituent populations in GEUVADIS.[44] To our knowledge, nobody has investigated cross-population portability within GEUVADIS. GEUVADIS provides us an opportunity to investigate predictive models with an experimentally homogeneous dataset: the GEUVADIS RNA-Seq data were produced in the same environment under the same protocol, from lymphoblastoid cell lines (LCLs) derived from similar sampling efforts, providing a high degree of technical harmonization. We train, test, and validate predictive models wholly within GEUVADIS with a nested cross-

validation scheme. Finally, to understand the consequences of eQTL architecture on TWAS, we use existing 1000 Genomes data to simulate two ancestral populations and an admixed population and then apply the same "train-test-validate" scheme with various simulated eQTL models to study cross-population prediction efficacy when a gold standard is known.

We compared transcriptome imputation accuracy in SAGE whole blood RNA using three PredictDB prediction weight sets for whole blood RNA: GTEx v6p, GTEx v7, and DGN. We also evaluated expression prediction with all four MESA monocyte weight sets: MESA_ALL (populations combined), MESA_AFA (African Americans), MESA_AFHI (combined African Americans and Hispanic Americans), and MESA_CAU (Caucasians). For each gene where both measured RNA-Seq gene expression and predictions are available in SAGE, we compute both the coefficient of determination (R^2) and Spearman correlation to analyze the direction of prediction. As we are primarily interested in describing the relationship between predicted outcome and real outcome, we prefer Spearman's ρ to describe correlations, while for determining prediction accuracy, we use the standard regression R^2, corresponding to the squared Pearson correlation, to facilitate comparisons to prior work. We then benchmark these against the out-of-sample R^2 and correlations from GTEx v7 and MESA as found in PredictDB. Prediction results in SAGE were available for 11,545 genes with a predictive model from at least one weight set. Not all sets derived models at the same genes: the prediction results across all weight sets overlapped at 273 genes, of which 39 genes had predictions with positive correlation to measurements. Since the estimation of these prediction models requires both high quality expression data and inferred eQTL, each weight set may have a different number of gene models. Therefore, intersecting seven different weight sets reduces the overall number of models available for comparison. This small number of genes in common is largely driven by MESA_AFA, the repository with the smallest number of predictive models. MESA_AFA contains the models that should best reflect the genetic ancestry in SAGE (Supplementary Table 1). We note that MESA_AFA also has the smallest training sample size among our weight sets (N = 233)[38], so the small number of predicted genes from MESA_AFA probably results from the small training sample size and not from any feature of the underlying MESA_AFA training data.

The concordance between predicted and measured gene expression over the 273 genes in common to all seven weight sets, with corresponding training metrics from PredictDB as benchmarks, shows worse performance than expected for R^2 (Figure 1) and correlations (Figure 2). The highest mean R^2 of 0.0336 was observed in DGN. Here, we highlight the intersection of genes across model sets for investigation, but the overall patterns for all genes are similar; results for the 11,545 total genes (Supplementary Figure 1) and the 39 genes with positive correlations (Supplementary Figure 2) showed little appreciable deviation from R^2 shown in Figure 1. Because SAGE is an independent validation set for the training populations, we would expect to observe some deterioration in imputation R^2 due to differences in population structure and linkage disequilibrium. However, Figure 1 shows a marked difference in model performance.

More noteworthy is the substantial proportion of predictions in SAGE with negative correlations to the real data. All seven weight sets produced negative mean correlations. The least negative mean correlation (-0.0044) was observed with GTEx_v6p, while the most negative mean

correlation (-0.020) was observed with MESA_AFA (Supplementary Table 1). The fact that correlations to SAGE measurements are negative on average suggests that some large R^2 values in Figure 1 may result from gene models with incorrect direction of prediction. While there are some fluctuations in prediction accuracy, no prediction weight set produces practically meaningfully better correlations to data than the others (-0.020 to -0.044). In contrast, the published PredictDB models for these genes show positive correlations to their training data, indicating no obvious incapacity for accurate prediction. However, available predictions into SAGE from otherwise valid prediction models are uniformly limited in power to capture true genotype-expression relationships.

To analyze genes with ostensibly high imputation R^2, we focus on genes in GTEx v7 with cross-validated $R^2 > 0.2$ in the reference population. Figure 3 compares PredictDB testing R^2 against the empirical R^2 from regressing predictions onto observations in SAGE. In this case, even the better-imputed gene models derived from PredictDB have limited ability to capture gene expression accurately in SAGE (mean R^2 0.031, IQR [0.0027, 0.037]).

It is important to note that real-world comparisons of RNA-Seq datasets can be subject to numerous sources of heterogeneity besides differential ancestry. Possible confounders include technical differences in sequencing protocols, differences in the age of participants[45], and the postmortem interval to tissue collection (for GTEx).[46–48] To investigate cross-population portability in an experimentally homogeneous context, we turn to GEUVADIS.[43] The GEUVADIS data include two continental population groups from the 1000 Genomes Project: the Europeans (EUR373), composed of 373 unrelated individuals from four subpopulations (Utahns (CEU), Finns (FIN), British (GBR), Toscani (TSI)), and the Africans (AFR) composed of 89 unrelated Yoruba (YRI) individuals. In light of the known bottleneck in Finnish population history , we analyze EUR373 both as one population and as two independent subgroups: the 95 Finnish individuals (FIN) and the 278 non-Finnish Europeans (EUR278). We used matched RNA-Seq, generated and harmonized together by the GEUVADIS Consortium and whole-genome genotype data in the resulting four populations (EUR373, EUR278, FIN, and AFR) to train predictive models for gene expression in a nested cross-validation scheme[6] and perform cross-population tests of imputation accuracy.

Table 1 shows R^2 from three training sets (EUR373, EUR278 and AFR) into the four testing populations (EUR373, EUR278, FIN, and AFR) for genes with positive correlation between prediction and measurement. While the number of genes with applicable models including genetic data varies in each train-test scenario (see Supplementary Table 3), we note that not all predictive models are trained on equal sample sizes, so the resulting R^2 only provide a general idea of how well one population imputes into another. Analyses within a population use out-of-sample imputation R^2 to avoid overfitting across train-test scenarios. Predicting from a population into itself yields R^2 ranging from 0.079 – 0.098 (Table 1) consistent with the smaller sample sizes in GEUVADIS versus GTEx and DGN. In contrast, predicting across populations yields more variable predictions, with R^2 ranging from 0.029 – 0.087. At the lower range of R^2 (0.029 – 0.039) are predictions from AFR into European testing groups (EUR373, EUR278, and FIN). Alternatively, when predicting from European training groups into AFR, the R^2 are noticeably

higher (0.051 – 0.054). Prediction from EUR278 into FIN (R^2 = 0.087) is better than prediction from EUR278 into AFR (R^2 = 0.051), suggesting that imputation R^2 may deteriorate with increased genetic distance. A comparison of the 564 genes in common across all train-test scenarios (Table 2) yields a more equal basis of comparison between populations, albeit from a subset of genes with potentially more consistent gene expression levels. In this case involving better-predicted genes, we see that imputation quality between the European groups improves noticeably, with R^2 ranging between 0.183 to 0.216, while R^2 between Europeans and Africans ranges from 0.095 to 0.147. In general, populations seem to predict better when imputing into themselves, and less well when imputing into other populations.

Combining all European subpopulations obscures population structure and can complicate analysis of cross-population imputation performance. To that end, we divide the GEUVADIS data into its five constituent populations and randomly subsample each of them to the smallest population size (n = 89). We then estimate models from each subpopulation and predict into all five subpopulations. Table 3 shows average R^2 from each population into itself and others. The populations consistently impute well into themselves, with imputation R^2 ranging from 0.104 – 0.136. However, a notable difference exists between the EUR subpopulations and YRI. The cross-population R^2 between CEU, TSI, GBR, and FIN ranges from 0.103 to 0.137, while cross-population R^2 from these populations into YRI ranges from 0.062 to 0.084. Imputation between YRI and the EUR populations taken together is consistently lower than within the EUR populations (Supplementary Figure 3) and statistically significant (p-value < 1.36 x 10^{-4}, Dunn test; see Supplementary Table 6). The cross-population differences remain for the 142 genes with positive correlation in all train-test scenarios (Table 4), where R^2 for imputation into YRI ranges from 0.166 to 0.244, while imputation within EUR populations ranges from 0.239 to 0.331. These results clearly suggest problems for prediction models that impute gene expression across populations, in similar regimes to those tested with linear predictive models and datasets of size consistent with current references. In addition, since AFR is genetically more distant from the EUR subpopulations than they are to each other, we interpret these results to imply that structure in populations can potentially exacerbate cross-population imputation quality (Supplementary Figure 4).

The unresolved question is the extent to which these results hold with oracle knowledge of eQTL architecture, something impossible to investigate in real data when the causal links between eQTL and gene expression can only be estimated. To investigate genomic architectures giving rise to gene expression, and in particular to investigate behavior in admixed populations, we simulate haplotypes from HapMap3[50] CEU and YRI using HAPGEN2[51] and then sample haplotypes in proportions consistent with observed admixture proportions (80% YRI, 20% CEU)[52] to construct a simulated African-American (AA) admixed population. We simulate eQTL architectures under an additive model of size k causal alleles (k = 1, 5, 10, and 20) and a phenotype with *cis*-heritability h^2 = 0.15 (recapitulating average h^2 in GTEx) using the genomic background of genic regions on chromosome 22, thus testing various model sizes and LD patterns. To tease apart the effect of shared eQTL architecture, we allow the populations to share eQTL with fixed effects in various proportions (0%, 10%, 20%, …, 100%). With these simulations providing known architectures for comparison, we then apply the train-test-validate scheme as before.

Figure 4 shows the cross-population Spearman correlations between predicted and simulated phenotypes in our simulated AA, CEU, and YRI, partitioned by proportion of shared eQTL, for *k* = 10 causal eQTL, with k = 5 and k = 20 showing similar effects (Supplementary Figure 5 and Supplementary Figure 6). Imputation within a population produced similar correlations in all cases, ranging from 0.323 to 0.329 (Supplementary Table 4). Secondly, the case of 100% shared eQTL architecture (where eQTL positions and effects are exactly the same across populations) models provide predictions with no loss in cross-population portability, with correlations ranging from 0.299 to 0.336 even when imputing across populations (Supplementary Table 5). This case suggests that eQTL that are causal in all populations can impute gene expression reliably regardless of the population in which they were ascertained, provided that the eQTL can be correctly mapped and genotyped in all populations, that the eQTL effects are identical across populations, and that a linear model of eQTL is assumed. For cases where eQTL architecture is not fully shared across populations, we see that imputation from each population into the other improves as the proportion of shared eQTL increases (Figure 4). The cross-population correlation between predicted gene expression versus measurement is highest between YRI and AA (0.037 to 0.088), intermediate between CEU and AA (0.0083 to 0.0245), and lowest between CEU and YRI (0.0017 to 0.0290). When imputing between two populations, the choice of which population is used to train predictive models produces no obvious difference in imputation quality. More explicitly, imputation quality between AA to CEU and CEU to AA is not significantly different (p-value ~ 1, Dunn test). The same applies between AA to YRI and YRI to AA (p-value ~ 1) and between CEU to YRI and YRI to CEU (p-value < 0.12). All other train/test scenarios are significantly different from each other as expected (Supplementary Table 7). The results for k = 5, 10, and 20 eQTL are consistent with the higher overall ancestral similarity of AA to YRI versus AA to CEU or CEU to YRI (k = 10, Figure 4, similar plots in Supplementary Figure 5 and Supplementary Figure 6). Although less realistic for most genes[5,6,18], we also analyzed models with a single causal eQTL. Trends for single-eQTL models are difficult to analyze due to simplicity in architecture (Supplementary Figure 7) and binary inference as to whether the causal is identified or not.

Overall, these results highlight two points: firstly, since prediction within populations is better than prediction between populations, our results reaffirm prior investigations[38] that population matching matters for optimally imputing gene expression. This is consistent with our results of impaired transcriptome imputation performance in SAGE with currently available resources. Secondly, despite decreased prediction accuracy when imputing between different populations, the populations that are more closely genetically related demonstrate better cross-population prediction. Imputation results from both GTEx and DGN into SAGE suggest that current predictive models, even for genes with greater heritability, perform worse than expected despite matching tissue types. Focusing on imputation R^2, as previous studies have done, may hide the observation of a substantial proportion of negative correlations between predictions and gene expression measurements in cross-population scenarios. Our investigation into cross-population imputation accuracy with GEUVADIS data replicates this lack of cross-population portability as observed with current GTEx and DGN predictive models. Since transcriptome prediction models use multivariate genotype predictors trained on a specific outcome, the impaired cross-population

application can be viewed as an analogous observation to that seen previously in polygenic scores.[35]

It is important to note that our observations do not reflect shortcomings of either the initial PrediXcan or TWAS frameworks. Nor do our findings affect the positive discoveries made using these frameworks over the past several years. These methods fully rely on the data used as input for training, and the most commonly used datasets for model training are overwhelmingly of European descent. Here we note that the current models fail to capture the complexity of the cross-population genomic architecture of gene expression for populations of non-European descent. Failing to account for this could lead researchers to draw incorrect conclusions from their genetic data, particularly as these models would lead to false negatives.

To this end, our simulations strongly suggest that imputing gene expression in a target population is improved by using predictive models constructed in a genetically similar training population. If populations share the exact same eQTL architecture, then they are essentially interchangeable for the purposes of gene expression imputation so long as eQTL are genotyped and accurately estimated, which remains a technological and statistical challenge. As the proportion of shared eQTL architecture decreases between two populations, the cross-population imputation quality decreases as well, and often dramatically. In both SAGE and GEUVADIS, we observe cross-population patterns consistent with an imperfect overlap of eQTL across populations. Ensuring representative eQTL architecture for all populations in genotype-expression repositories will require a solid understanding of true cross-population and population-specific eQTL. However, expanding the amount of global genetic architecture represented in genotype-expression repositories, which can be accomplished by sampling more populations, provides the most desirable course for improving gene expression prediction models. Additionally, this presents an opportunity for future research in methods that could improve cross-population portability, particularly when one population is over-represented in reference data. Tools from transfer learning could facilitate porting TWAS eQTL models from reference populations to target populations using little or no RNA-Seq data.

In light of the surging interest in gene expression imputation, we see a pressing need for freely distributed predictive models of gene expression estimated from coupled transcriptome-genome data sampled in a variety of populations and tissues. The recently published predictive models with multi-ethnic MESA data constitute a crucial first step in this direction for researchers working with admixed populations. However, the clinical and biomedical research communities must push for more diverse genotype-expression resources to ensure that the fruits of genomic studies benefit all populations.

345 **Online Resources**
346
347 PredictDB: http://predictdb.org/
348 GTEx: http://gtexportal.org/
349 DGN: http://dags.stanford.edu/dgn/
350 GEUVADIS: https://www.ebi.ac.uk/Tools/geuvadis-das/
351 Source code: https://github.com/asthmacollaboratory/sage-geuvadis-predixcan
352 Results and simulation data: https://ucsf.box.com/v/sage-geuvadis-predixcan
353

SUPPLEMENTAL MATERIALS AND METHODS

Genotype and RNA-Seq data

RNA-Seq (RNA sequencing) data generation and cleaning protocols for 39 SAGE subjects analyzed here were initially described in (Mak, White, Eckalbar, et al. 2018).[39] Genotypes were generated on the Affymetrix Axiom array as described previously.[53] Genotypes were then imputed on the Michigan Imputation Server[54] with EAGLE v2.3[55] and the 1000 Genomes panel phase 3 v5[56] and then subjected to the following filters: <5% missing sample, <5% missing genotypes, >1% MAF, >1e-4 HWE, and >0.3 imputation R^2. The choice of the 1000 Genomes panel follows GTEx protocol, though GTEx used the smaller 1000 Genomes phase 1 panel.[4] Gene expression counts were processed through the GTEx v6p eQTL quality control pipeline and as described previously.[18] This filtering process kept 20,985 genes with Ensembl identifiers for analysis, of which 20,268 were autosomal genes. We then quantile normalized the remaining gene expression values across samples as our gene expression measurements.

GEUVADIS genotype VCF files and normalized gene expression data (filename `GD462.GeneQuantRPKM.50FN.samplename.resk10.txt.gz`) were downloaded directly from the EMBL-EBI GEUVADIS Data Browser. Genotypes were filtered similarly to SAGE subjects. No manipulation was performed on expression data. This process yielded 23,722 genes for analysis.

Running PrediXcan models

We ran PrediXcan on SAGE subjects using PredictDB prediction weights from three paired genotype-expression datasets from PredictDB: GTEx, DGN, and MESA.[6,9,38,57] For GTEx, we used both GTEx v6p and GTEx v7 weights. For MESA, we used all weight sets from the freeze dated 2018-05-30: African Americans (MESA_AFA), African Americans and Hispanics (MESA_AFHI), Caucasians (MESA_CAU), and all MESA samples (MESA_ALL). Overall, the analysis included 10,161 genes, of which only 273 had *both* normalized RNA-Seq measures and predictions from *all* weight sets. Of these, 126 had positive correlation between prediction and measurement. We assessed imputation quality by comparing PrediXcan predictions to normalized gene expression from SAGE using linear regression and correlation tests.

Building prediction models

We trained prediction models in GEUVADIS on genotypes in a 500Kb window around each of 23,723 genes with measured and normalized gene expression. GEUVADIS subjects were partitioned into various groups: the Europeans (EUR373), the non-Finnish Europeans (EUR278), the Yoruba (AFR), and the constituent 1000 Genomes populations (CEU, GBR, TSI, FIN, and YRI). For each training set, we performed nested cross-validation. The external cross-validation for all populations used leave-one-out cross-validation (LOOCV). The internal cross-validation used 10-fold cross-validation for EUR373 and EUR278 and LOOCV for the five constituent GEUVADIS populations in order to fully utilize the smaller sample size (n = 89) compared to EUR278 (n = 278)

and EUR373 (n = 373). Internal cross-validation used elastic net regression with mixing parameter α = 0.5 as implemented in the `glmnet` package in R. The nonzero weights for each SNP from each LOOCV were compiled and averaged for each gene, yielding a single set of prediction weights for each gene. Predictions were computed by parsing genotype dosages from the target population corresponding to the nonzero SNP predictors, and then multiplying dosages against the prediction weights. The resulting predictions were compared to normalized gene expression measurements downloaded from the GEUVADIS data portal. The comparison of predictive models cannot easily differentiate predictions of 0 (no gene expression) and NA (missing expression). We addressed this with two additional filters. Firstly, we removed genes that did not have any eQTL in their predictive models. Secondly, genes where fewer than half of the individuals had nonmissing predictions were removed from further analysis. Coefficients of determination (R^2) were computed with the `lm` function in R. Spearman correlations were computed with the `cor.test` function in R.

Simulation

We downloaded a sample of 20,085 HapMap 3 SNPs[50] from each of CEU and YRI on chromosome 22 as provided by HAPGEN2.[51] The data include 234 phased haplotypes for CEU and 230 phased haplotypes for YRI. We forward-simulated from these haplotypes to obtain two populations of n = 1000 individuals each. We then sampled haplotypes in proportions of 80% YRI and 20% CEU to obtain a mixture of CEU and YRI where the ancestry patterns roughly mimic those of African Americans. For computational simplicity, and in keeping with the high ancestry LD present in African Americans[58,59], for each gene we assumed local ancestry was constant for each haplotype. For each of the three simulated populations, we applied the same train-test-validate scheme used for cross-population analysis in GEUVADIS. Genetic data for model simulation were downloaded from Ensembl 89 and included the largest 100 genes from chromosome 22. We defined each gene as the start and end positions corresponding to the canonical transcript, plus 1 megabase in each direction. Two genes, PPP6R2 and MOV10L1, spanned no polymorphic markers in our simulated data, resulting in 98 gene models used for analysis. To simulate predictive eQTL models, we tested multiple parameter configurations for each gene: we varied the number of causal eQTL (k = 1, 5, 10, and 20), the positions (all same or not all same), and the proportion of shared positions (p = 0.0, 0.1, 0.2, …, 0.9, 1). Each model included a simulated gene expression phenotype with *cis*-heritability set to 0.15. For each parameter configuration, we ran 100 different random instantiations of the model simulations.

Analysis tools

Analyses used GNU `parallel`[60] and the `tidyverse` bundle of R packages.[61] All plots were generated with `ggplot2`.[62]

CONFLICT OF INTEREST

C.R.G. owns stock in 23andMe, Inc. The remaining authors declare no potential conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported in part by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, the Harry Wm. and Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, the National Heart, Lung, and Blood Institute (NHLBI) R01HL117004, R01HL128439, R01HL135156, X01HL134589, R01HL141992, and R01HL104608, the National Human Genome Research Institute (NHGRI) U01HG007419, National Institute of Environmental Health Sciences R01ES015794, R21ES24844, the National Institute on Minority Health and Health Disparities P60MD006902, R01MD010443, RL5GM118984 and the Tobacco-Related Disease Research Program under Award Numbers 24RT-0025, 27IR-0030. Research reported in this article was funded by the National Institutes of Health Common Fund and Office of Scientific Workforce Diversity under three linked awards RL5GM118984, TL4GM118986, 1UL1GM118985 administered by the National Institute of General Medical Sciences.

The authors wish to acknowledge the following SAGE co-investigators for subject recruitment, sample processing and quality control: Luisa N. Borrell, DDS, PhD, Emerita Brigino-Buenaventura, MD, Adam Davis, MA, MPH, Michael A. LeNoir, MD, Kelley Meade, MD, Fred Lurmann, MS and Harold J. Farber, MD, MSPH. The authors also wish to thank the staff and participants who contributed to the SAGE study.

K.L.K was additionally supported by a diversity supplement of NHLBI R01HL135156, the UCSF Bakar Computational Health Sciences Institute, the Gordon and Betty More Foundation grant GBMF3834, and the Alfred P. Sloan Foundation grant 2013-10-27 to UC Berkeley through the Moore-Sloan Data Sciences Environment initiative at the Berkeley Institute for Data Science (BIDS). The logistical space, technical support, administrative assistance, and indefatigable good humor of the members and staff at BIDS is gratefully acknowledged.

M.J.W. was additionally supported by a diversity supplement of NHLBI R01HL117004, an Institutional Research and Academic Career Development Award K12GM081266, and an NHLBI Research Career Development (K) Award K01HL140218.

M.G.C was additionally supported by NIH MARC U-STAR grant T34GM008574 at San Francisco State University.

C.R.G. was additionally supported by grants R56HG010297 and T32HG00044.

476 **REFERENCES**
477
478 1. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P.,
479 Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the
480 Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. *12*,.

481 2. NHLBI Trans-Omics for Precision Medicine.

482 3. NHGRI Genome Sequencing Program (GSP).

483 4. The 1000 Genomes Consortium An integrated map of genetic variation from 1,092 human
484 genomes | Nature.

485 5. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus,
486 E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale
487 transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

488 6. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll,
489 R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., et al. (2015). A gene-based association
490 method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.

491 7. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*,
492 580–585.

493 8. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C.,
494 Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic
495 basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*,
496 14–24.

497 9. Barbeira, A.N., Dickinson, S.P., Torres, J.M., Bonazzola, R., Zheng, J., Torstenson, E.S.,
498 Wheeler, H.E., Shah, K.P., Edwards, T., Garcia, T., et al. (2017). Exploring the phenotypic
499 consequences of tissue specific gene expression variation inferred from GWAS summary
500 statistics. BioRxiv 045260.

501 10. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W.,
502 Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from
503 GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.

504 11. Mostafavi, S., Gaiteri, C., Sullivan, S.E., White, C.C., Tasaki, S., Xu, J., Taga, M., Klein, H.-
505 U., Patrick, E., Komashko, V., et al. (2018). A molecular network of the aging human brain
506 provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat. Neurosci.
507 *21*, 811.

508 12. Ferreira, M.A.R., Jansen, R., Willemsen, G., Penninx, B., Bain, L.M., Vicente, C.T., Revez,
509 J.A., Matheson, M.C., Hui, J., Tung, J.Y., et al. (2017). Gene-based analysis of regulatory
510 variants identifies four putative novel asthma risk genes related to nucleotide synthesis and
511 signaling. J. Allergy Clin. Immunol. *139*, 1148–1157.

13. Lamontagne, M., Bérubé, J.-C., Obeidat, M., Cho, M.H., Hobbs, B.D., Sakornsakolpat, P., de Jong, K., Boezen, H.M., International COPD Genetics Consortium, Nickle, D., et al. (2018). Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. Hum. Mol. Genet. *27*, 1819–1829.

14. Thériault, S., Gaudreault, N., Lamontagne, M., Rosa, M., Boulanger, M.-C., Messika-Zeitoun, D., Clavel, M.-A., Capoulade, R., Dagenais, F., Pibarot, P., et al. (2018). A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. Nat. Commun. *9*, 988.

15. Porcu, E., Rüeger, S., Consortium, eQTLGen, Santoni, F.A., Reymond, A., and Kutalik, Z. (2018). Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. BioRxiv 377267.

16. Gusev, A., Lawrenson, K., Segato, F., Fonseca, M., Kar, S., Lee, J., Pejovic, T., Consortium, O.C.A., Karlan, B., Freedman, M., et al. (2018). Multi-Tissue Transcriptome-Wide Association Studies Identify 21 Novel Candidate Susceptibility Genes for High Grade Serous Epithelial Ovarian Cancer. BioRxiv 330613.

17. Huckins, L.M., Dobbyn, A., Ruderfer, D., Hoffman, G., Wang, W., Pardinas, A.F., Rajagopal, V.M., Als, T.D., Hoang, H.T., Girdhar, K., et al. (2017). Gene expression imputation across multiple brain regions reveals schizophrenia risk throughout development. BioRxiv 222596.

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

19. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.

20. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.

21. Bentley, A.R., Callier, S., and Rotimi, C.N. (2017). Diversity and inclusion in genomic research: why the uneven progress? J. Community Genet. *8*, 255–266.

22. Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E.C., Hutter, C.M., Manolio, T.A., and Green, E.D. (2018). Prioritizing diversity in human genomics research. Nat. Rev. Genet. *19*, 175–185.

23. Asimit, J.L., Hatzikotoulas, K., McCarthy, M., Morris, A.P., and Zeggini, E. (2016). Trans-ethnic study design approaches for fine-mapping. Eur. J. Hum. Genet. *24*, 1330–1336.

24. Wang, X., Cheng, C.-Y., Liao, J., Sim, X., Liu, J., Chia, K.-S., Tai, E.-S., Little, P., Khor, C.-C., Aung, T., et al. (2016). Evaluation of transethnic fine mapping with population-specific and cosmopolitan imputation reference panels in diverse Asian populations. Eur. J. Hum. Genet. *24*, 592–599.

548 25. Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies:
549 advantages and challenges of mapping in diverse populations. Genome Med. *6*, 91.

550 26. Kumar, R., Seibold, M.A., Aldrich, M.C., Williams, L.K., Reiner, A.P., Colangelo, L.,
551 Galanter, J., Gignoux, C., Hu, D., Sen, S., et al. (2010). Genetic ancestry in lung-function
552 predictions. N. Engl. J. Med. *363*, 321–330.

553 27. Yang, J.J., Cheng, C., Devidas, M., Cao, X., Fan, Y., Campana, D., Yang, W., Neale, G.,
554 Cox, N.J., Scheet, P., et al. (2011). Ancestry and pharmacogenomics of relapse in acute
555 lymphoblastic leukemia. Nat. Genet. *43*, 237–241.

556 28. Acuña-Alonzo, V., Flores-Dorantes, T., Kruit, J.K., Villarreal-Molina, T., Arellano-Campos,
557 O., Hünemeier, T., Moreno-Estrada, A., Ortiz-López, M.G., Villamil-Ramírez, H., León-Mimila,
558 P., et al. (2010). A functional ABCA1 gene variant is associated with low HDL-cholesterol
559 levels and shows evidence of positive selection in Native Americans. Hum. Mol. Genet. *19*,
560 2877–2885.

561 29. Adeyemo, A., and Rotimi, C. (2010). Genetic variants associated with complex human
562 diseases show wide variation across multiple populations. Public Health Genomics *13*, 72–79.

563 30. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P.,
564 Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic Misdiagnoses and the Potential
565 for Health Disparities. N. Engl. J. Med. *375*, 655–665.

566 31. Petrovski, S., and Goldstein, D.B. (2016). Unequal representation of genetic variation across
567 ancestry groups creates healthcare inequality in the application of precision medicine. Genome
568 Biol. *17*, 157.

569 32. Oh, S.S., White, M.J., Gignoux, C.R., and Burchard, E.G. (2016). Making Precision
570 Medicine Socially Precise. Take a Deep Breath. Am. J. Respir. Crit. Care Med. *193*, 348–350.

571 33. Oh, S.S., Galanter, J., Thakur, N., Pino-Yanes, M., Barcelo, N.E., White, M.J., de Bruin,
572 D.M., Greenblatt, R.M., Bibbins-Domingo, K., Wu, A.H.B., et al. (2015). Diversity in Clinical
573 and Biomedical Research: A Promise Yet to Be Fulfilled. PLoS Med. *12*,.

574 34. Belbin, G.M., Nieves-Colón, M.A., Kenny, E.E., Moreno-Estrada, A., and Gignoux, C.R.
575 (2018). Genetic diversity in populations across Latin America: implications for population and
576 medical genetic studies. Curr. Opin. Genet. Dev. *53*, 98–104.

577 35. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly,
578 M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic
579 Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.

580 36. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R.,
581 Greenland, P., Jacob, D.R., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of
582 Atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881.

583 37. Liu, Y., Ding, J., Reynolds, L.M., Lohman, K., Register, T.C., De La Fuente, A., Howard,
584 T.D., Hawkins, G.A., Cui, W., Morris, J., et al. (2013). Methylomics of gene expression in
585 human monocytes. Hum. Mol. Genet. *22*, 5065–5074.

586 38. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson,
587 W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene expression traits
588 across diverse populations. PLOS Genet. *14*, e1007586.

589 39. Mak, A.C.Y., White, M.J., Eckalbar, W.L., Szpiech, Z.A., Oh, S.S., Pino-Yanes, M., Hu, D.,
590 Goddard, P., Huntsman, S., Galanter, J., et al. (2018). Whole-Genome Sequencing of
591 Pharmacogenetic Drug Response in Racially Diverse Children with Asthma. Am. J. Respir. Crit.
592 Care Med. *197*, 1552–1564.

593 40. Thakur, N., Oh, S.S., Nguyen, E.A., Martin, M., Roth, L.A., Galanter, J., Gignoux, C.R.,
594 Eng, C., Davis, A., Meade, K., et al. (2013). Socioeconomic status and childhood asthma in
595 urban minority youths. The GALA II and SAGE II studies. Am. J. Respir. Crit. Care Med. *188*,
596 1202–1209.

597 41. Borrell, L.N., Nguyen, E.A., Roth, L.A., Oh, S.S., Tcheurekdjian, H., Sen, S., Davis, A.,
598 Farber, H.J., Avila, P.C., Brigino-Buenaventura, E., et al. (2013). Childhood Obesity and Asthma
599 Control in the GALA II and SAGE II Studies. Am. J. Respir. Crit. Care Med. *187*, 697–702.

600 42. Nishimura, K.K., Galanter, J.M., Roth, L.A., Oh, S.S., Thakur, N., Nguyen, E.A., Thyne, S.,
601 Farber, H.J., Serebrisky, D., Kumar, R., et al. (2013). Early-life air pollution and asthma risk in
602 minority children. The GALA II and SAGE II studies. Am. J. Respir. Crit. Care Med. *188*, 309–
603 318.

604 43. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A.,
605 Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and
606 genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

607 44. Mikhaylova, A.V., and Thornton, T.A. (2019). Accuracy of gene expression prediction from
608 genotype data with PrediXcan varies across diverse populations. BioRxiv 524728.

609 45. Viñuela, A., Brown, A.A., Buil, A., Tsai, P.-C., Davies, M.N., Bell, J.T., Dermitzakis, E.T.,
610 Spector, T.D., and Small, K.S. (2018). Age-dependent changes in mean and variance of gene
611 expression across tissues in a twin cohort. Hum. Mol. Genet. *27*, 732–741.

612 46. McCall, M.N., Illei, P.B., and Halushka, M.K. (2016). Complex Sources of Variation in
613 Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. Am. J. Hum. Genet. *99*,
614 624–635.

615 47. Zhu, Y., Wang, L., Yin, Y., and Yang, E. (2017). Systematic analysis of gene expression
616 patterns associated with postmortem interval in human tissues. Sci. Rep. *7*, 5435.

617 48. Ferreira, P.G., Muñoz-Aguirre, M., Reverter, F., Godinho, C.P.S., Sousa, A., Amadoz, A.,
618 Sodaei, R., Hidalgo, M.R., Pervouchine, D., Carbonell-Caballero, J., et al. (2018). The effects of
619 death and post-mortem cold ischemia on human tissue transcriptomes. Nat. Commun. *9*, 490.

620 49. Martin, A.R., Karczewski, K.J., Kerminen, S., Kurki, M.I., Sarin, A.-P., Artomov, M.,
621 Eriksson, J.G., Esko, T., Genovese, G., Havulinna, A.S., et al. (2018). Haplotype Sharing
622 Provides Insights into Fine-Scale Population History and Disease in Finland. Am. J. Hum. Genet.
623 *102*, 760–775.

624 50. The International HapMap 3 Consortium (2010). Integrating common and rare genetic
625 variation in diverse human populations. Nature *467*, 52–58.

626 51. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease
627 SNPs. Bioinformatics *27*, 2304–2305.

628 52. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J.,
629 Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., et al. (2016). The Great
630 Migration and African-American Genomic Diversity. PLOS Genet. *12*, e1006059.

631 53. Hoffmann, T.J., Zhan, Y., Kvale, M.N., Hesselson, S.E., Gollub, J., Iribarren, C., Lu, Y.,
632 Mei, G., Purdy, M.M., Quesenberry, C., et al. (2011). Design and coverage of high throughput
633 genotyping arrays optimized for individuals of East Asian, African American, and Latino
634 race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics *98*, 422–
635 430.

636 54. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew,
637 E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and
638 methods. Nat. Genet. *48*, 1284–1287.

639 55. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K.,
640 Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing
641 using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.

642 56. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J.,
643 Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural
644 variation in 2,504 human genomes. Nature *526*, 75–81.

645 57. Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., Consortium, Gte.,
646 Cox, N.J., Nicolae, D.L., and Im, H.K. (2016). Survey of the Heritability and Sparse Architecture
647 of Gene Expression Traits across Human Tissues. PLOS Genet. *12*, e1006423.

648 58. Gravel, S. (2012). Population genetics models of local ancestry. Genetics *191*, 607–619.

649 59. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H.,
650 Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of
651 distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.

652 60. Tange, O. (2018). GNU Parallel 2018 (Ole Tange).

653 61. Wickham, Hadley, and Grolemund, Garrett (2017). R for Data Science (O'Reilly Media,
654 Inc.).

655 62. Wickham, Hadley (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag
656 New York).

657 63. The 1000 Genomes Project Consortium (2010). A map of human genome variation from
658 population-scale sequencing. Nature *467*, 1061–1073.

659

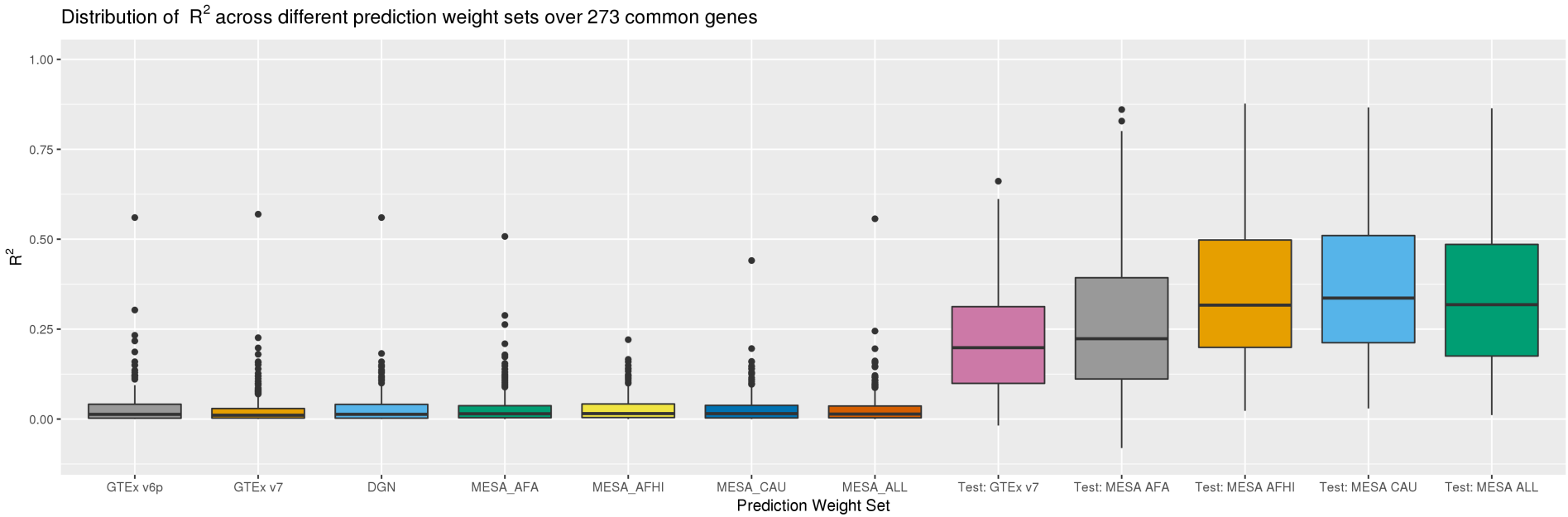


Figure 1: R^2 of measured gene expression versus predictions from PrediXcan. The prediction weights used here are, from left to right: GTEx v6p, GTEx v7, DGN, MESA African Americans, MESA African Americans and Hispanics, MESA Caucasians, and all MESA subjects. Test R^2 from model training in GTEx 7 and MESA appear on the right and provide a performance baseline.

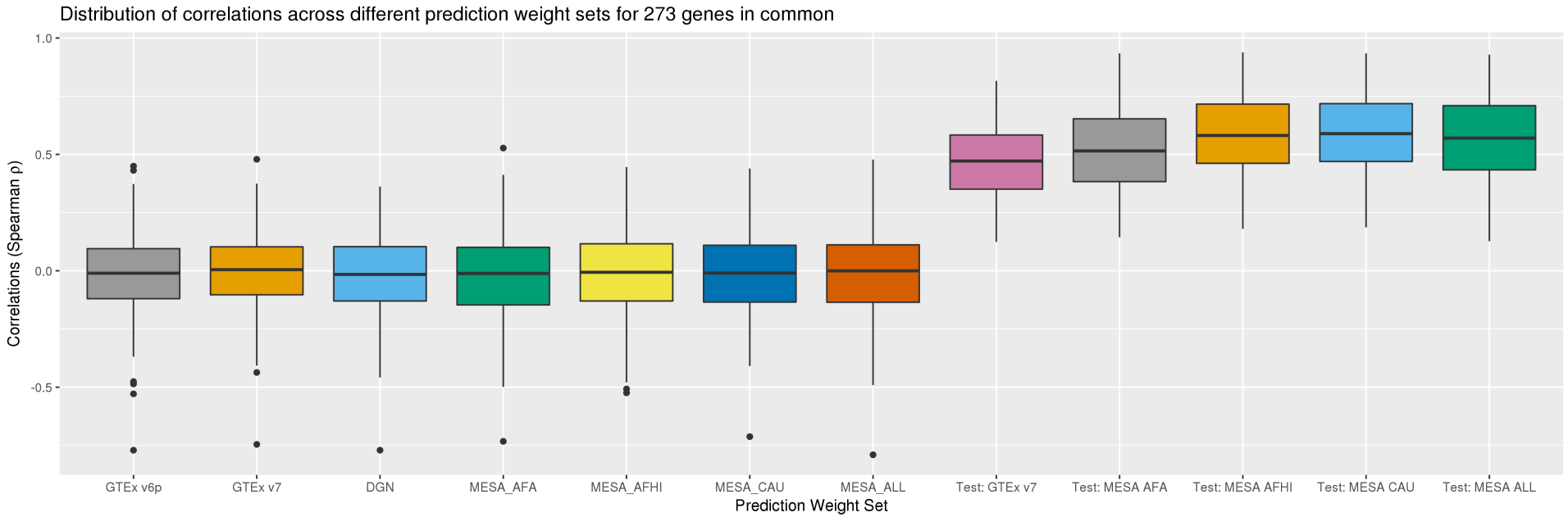
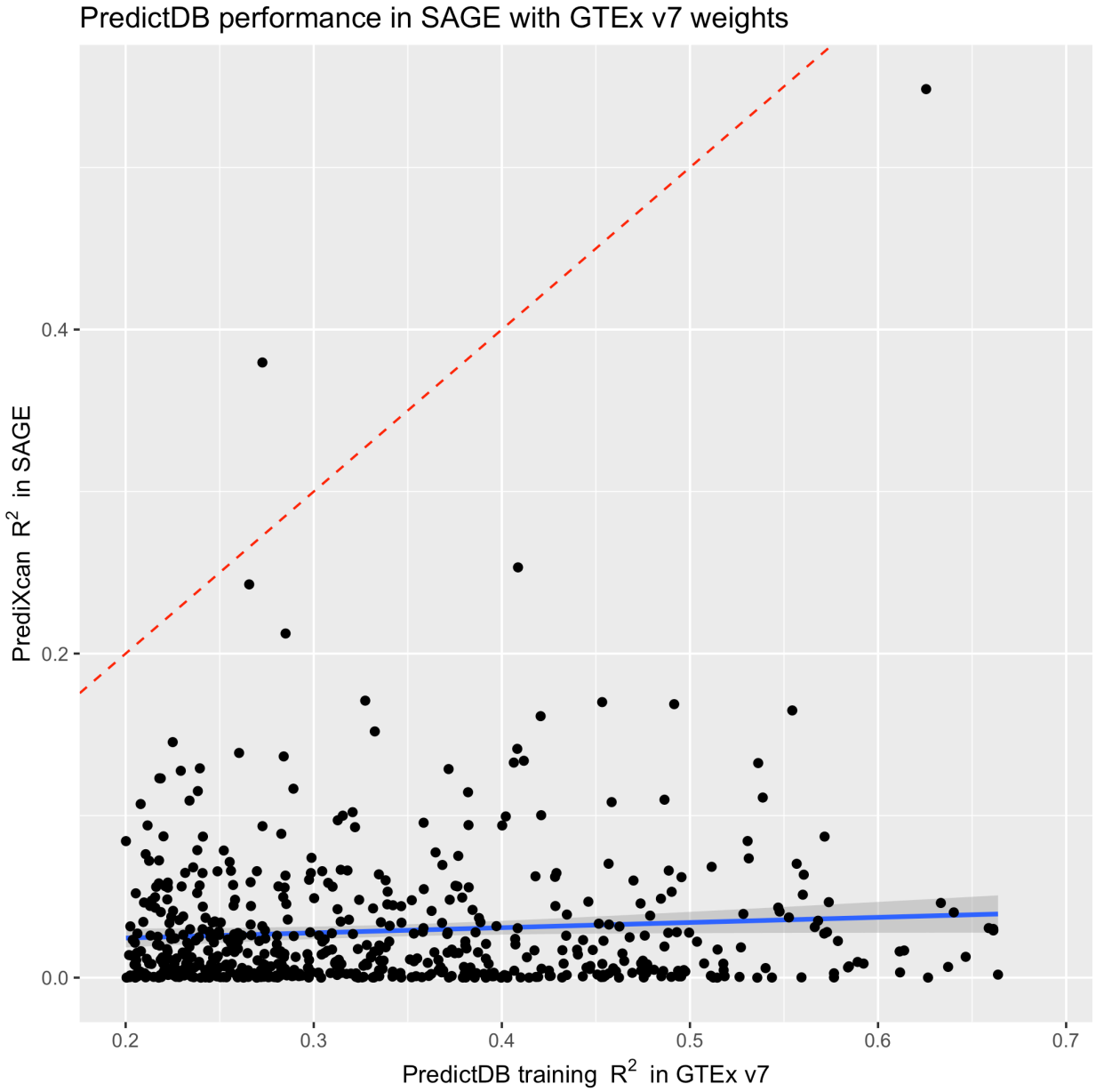


Figure 2: Spearman correlations of measured gene expression versus predicted expression from PrediXcan. The order of the weight sets matches Figure 1.

670



671

672 *Figure 3: A comparison of R^2 from SAGE and GTEx v7 training diagnostics. The SAGE R^2 are*
673 *computed from regressing PrediXcan predictions onto gene expression measurements. The GTEx*
674 *v7 R^2 are taken from PredictDB. The red dotted line marks where R^2 between the two groups*
675 *match, while the blue line denotes the best linear fit.*

R^2		Train Pop		
		EUR373	**EUR278**	**AFR**
Test Pop	**EUR373**	0.098	n/a	0.029
	EUR278	n/a	0.096	0.030
	FIN	n/a	0.087	0.039
	AFR	0.054	0.051	0.079

Table 1: Imputation R^2 between populations in GEUVADIS for genes with positive correlation between predictions and measurements. Scenarios where the training sample is contained in the testing sample cannot be accurately tested and are marked with "n/a". EUR373 includes all Europeans, EUR278 includes only non-Finnish Europeans, FIN includes only the Finnish, and AFR includes only the Yoruba.

R^2		Train Pop		
		EUR373	**EUR278**	**AFR**
Test Pop	**EUR373**	0.201	n/a	0.096
	EUR278	n/a	0.183	0.095
	FIN	n/a	0.216	0.111
	AFR	0.147	0.141	0.130

Table 2: Imputation R^2 between populations in GEUVADIS for 564 gene models that show positive correlation between prediction and measurement in all 9 train-test scenarios that were analyzed. Scenarios that were not tested are marked with "n/a". As before, EUR373 includes all Europeans, EUR278 includes only non-Finnish Europeans, FIN includes only the Finnish, and AFR includes only the Yoruba.

R2 Mean (Std Err)		Training population				
		CEU	TSI	GBR	FIN	YRI
Testing Pop	CEU	0.115 (0.139)	0.106 (0.139)	0.107 (0.134)	0.103 (0.133)	0.069 (0.116)
	TSI	0.124 (0.158)	0.121 (0.151)	0.124 (0.149)	0.118 (0.145)	0.083 (0.13)
	GBR	0.132 (0.16)	0.137 (0.155)	0.136 (0.156)	0.133 (0.155)	0.087 (0.132)
	FIN	0.128 (0.158)	0.130 (0.155)	0.130 (0.153)	0.130 (0.152)	0.084 (0.134)
	YRI	0.065 (0.108)	0.069 (0.112)	0.063 (0.1)	0.062 (0.102)	0.104 (0.138)

Table 3: Cross-population prediction performance across all five constituent GEUVADIS populations over genes with positive correlation between predictions and measurements. All populations were subsampled to N = 89 individuals. The number of genes represented varies by training sample (CEU: N = 1029, FIN: N = 1320, GBR: 1436, TSI: 1250, YRI: 914).

R2 Mean (Std Err)		Training population				
		CEU	TSI	GBR	FIN	YRI
Testing Pop	CEU	0.239 (0.18)	0.269 (0.177)	0.291 (0.166)	0.297 (0.168)	0.201 (0.164)
	TSI	0.307 (0.188)	0.294 (0.21)	0.331 (0.182)	0.322 (0.185)	0.227 (0.185)
	GBR	0.320 (0.175)	0.326 (0.181)	0.318 (0.191)	0.350 (0.178)	0.235 (0.183)
	FIN	0.318 (0.191)	0.320 (0.198)	0.343 (0.182)	0.323 (0.201)	0.244 (0.192)
	YRI	0.166 (0.164)	0.205 (0.163)	0.195 (0.157)	0.189 (0.156)	0.213 (0.177)

Table 4: Cross-population prediction performance across all five subsampled GEUVADIS populations over the 142 genes with positive correlation between prediction and measurement in all 25 train-test scenarios.

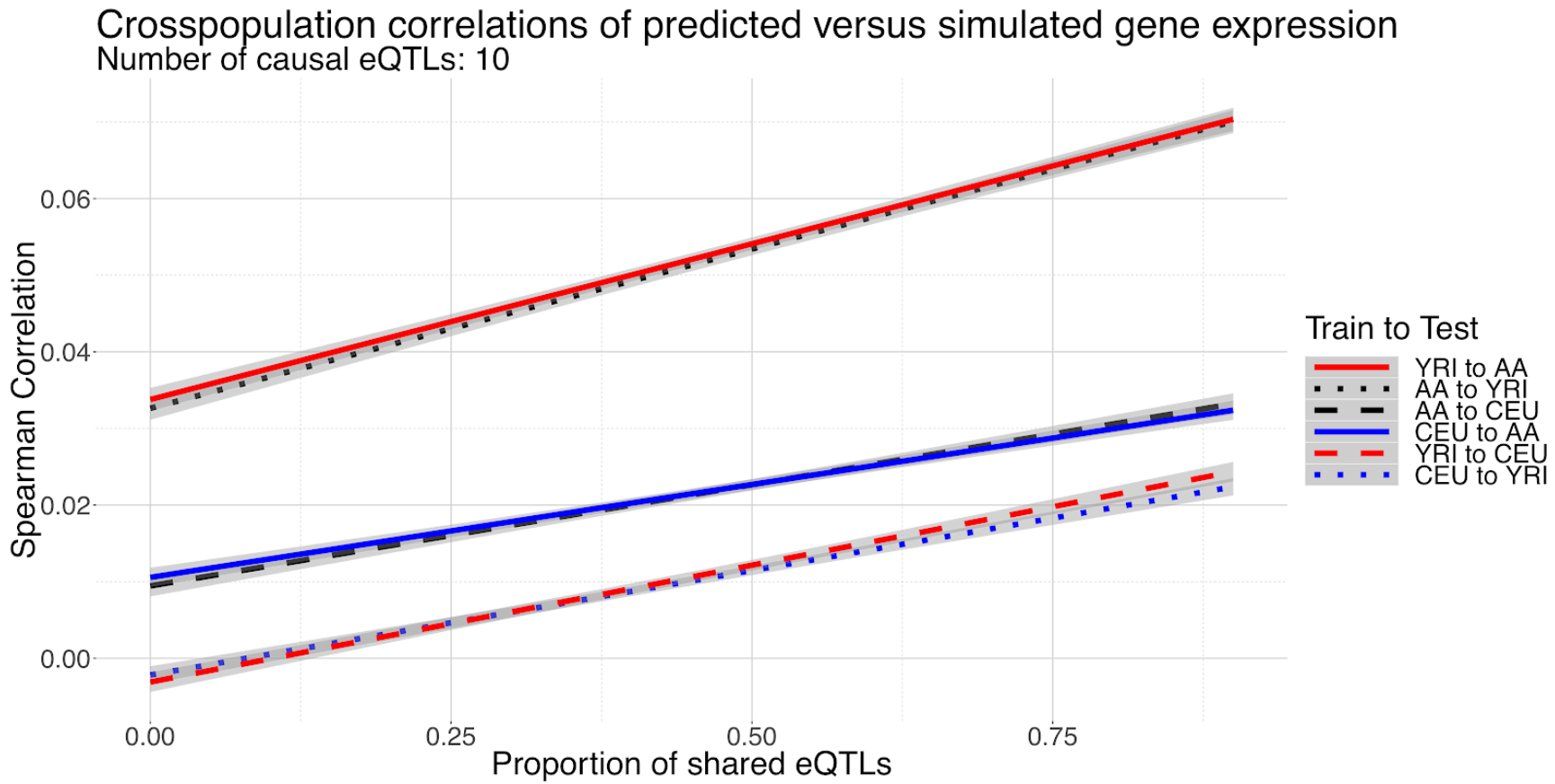


Figure 4: Correlations between predictions and simulated gene expression measurements from simulated populations across various proportions of shared eQTL architecture with 10 causal cis-eQTL. Here YRI is simulated from the 1000 Genomes Yoruba, CEU is simulated from the Utahns, and AA is constructed from YRI and CEU. Each trend line represents a linear interpolation of correlation versus shared eQTL proportion. Gray areas denote 95% confidence regions of LOESS-smoothed mean correlations conditional on the proportion of shared eQTL.

Weight set	Gene models	Genes predicted in SAGE	Genes both predicted and measured	Genes with positively correlated predictions and measurements	Mean Correlation (273 common genes)
GTEx v6p	6588	5773	5348	2730	-0.0044
GTEx v7	6297	2742	2570	1319	-0.0113
DGN	13171	4033	3678	1819	-0.0124
MESA_AFA	3551	995	982	497	-0.0204
MESA_AFHI	5556	1889	1862	969	-0.0049
MESA_CAU	4674	1654	1633	837	-0.0082
MESA_ALL	6217	2443	2408	1201	-0.0107

Supplementary Table 1: Summary statistics for analyzing gene expression prediction in SAGE for all seven weight sets in PredictDB. SAGE has measurements for 20,985 genes, of which 20,268 are autosomal. The intersection of genes with both predictions and measurements in SAGE across all seven weight sets is 273, of which 39 produce predictions positively correlated to data in all comparisons.

703
704

Pop	Measured genes	Predictive Models	With >50% samples predicted	Analyzed prediction v. measurement	Positive correlation
EUR373	23723	20418	11917	11914	5586
EUR278	23723	20182	11043	11043	4817
YRI89	23723	20699	11180	11179	4867

705

Supplementary Table 2: Summary statistics for each filtering step in the analysis of gene expression models from GEUVADIS for the 3 training populations EUR373, EUR278, and AFR. The analysis of prediction vs. measurement contains 5038 genes in common between all three populations. Of these genes, 1476 genes demonstrate positive correlation between predictions and measurements.

Training Pop	Testing Pop	R^2	Correlation	Transcripts
AFR	AFR	0.079	0.2329	2562
AFR	EUR278	0.030	0.1122	2996
AFR	EUR373	0.029	0.1072	3043
AFR	FIN	0.039	0.1377	2908
EUR278	AFR	0.051	0.1632	3079
EUR278	EUR278	0.096	0.2291	2857
EUR278	FIN	0.087	0.2171	3994
EUR373	AFR	0.054	0.1683	3105
EUR373	EUR373	0.098	0.2325	3132

706 *Supplementary Table 3: Summary statistics from training and testing results with continental*
707 *GEUVADIS populations for gene models with positive correlations. The R^2 correspond to Table 1.*
708 *The column "Correlation" lists the Spearman correlations for each scenario, while "Transcripts"*
709 *gives the number of gene models used to compute the R^2 and correlation summaries.*

710

Training Pop	Testing Pop	Shared eQTL Proportion	Correlation (Mean)	Correlation (StdErr)
AA	AA	0	0.323	0.0350
AA	AA	0.1	0.323	0.0357
AA	AA	0.2	0.323	0.0356
AA	AA	0.3	0.324	0.0355
AA	AA	0.4	0.323	0.0361
AA	AA	0.5	0.323	0.0355
AA	AA	0.6	0.323	0.0358
AA	AA	0.7	0.321	0.0364
AA	AA	0.8	0.323	0.0361
AA	AA	0.9	0.322	0.0358
CEU	CEU	0	0.329	0.0345
CEU	CEU	0.1	0.329	0.0345
CEU	CEU	0.2	0.329	0.0345
CEU	CEU	0.3	0.329	0.0345
CEU	CEU	0.4	0.329	0.0345
CEU	CEU	0.5	0.329	0.0345
CEU	CEU	0.6	0.329	0.0345
CEU	CEU	0.7	0.329	0.0346
CEU	CEU	0.8	0.329	0.0345
CEU	CEU	0.9	0.329	0.0345
YRI	YRI	0	0.325	0.0354
YRI	YRI	0.1	0.325	0.0355
YRI	YRI	0.2	0.324	0.0351
YRI	YRI	0.3	0.324	0.0354
YRI	YRI	0.4	0.325	0.0354
YRI	YRI	0.5	0.325	0.0352
YRI	YRI	0.6	0.324	0.0351
YRI	YRI	0.7	0.322	0.0354
YRI	YRI	0.8	0.324	0.0352
YRI	YRI	0.9	0.324	0.0350

Supplementary Table 4: Spearman correlations between prediction versus simulated measurement from simulated populations to themselves across various shared eQTL proportions for k = 10 causal eQTL.

Correlation Mean (Std Err)		Train-test direction		
		AA	**CEU**	**YRI**
Training Pop	**AA**	0.321 (0.0071)	0.308 (0.0058)	0.336 (0.0052)
	CEU	0.334 (0.006)	0.329 (0.0069)	0.326 (0.0063)
	YRI	0.336 (0.0051)	0.299 (0.007)	0.325 (0.0063)

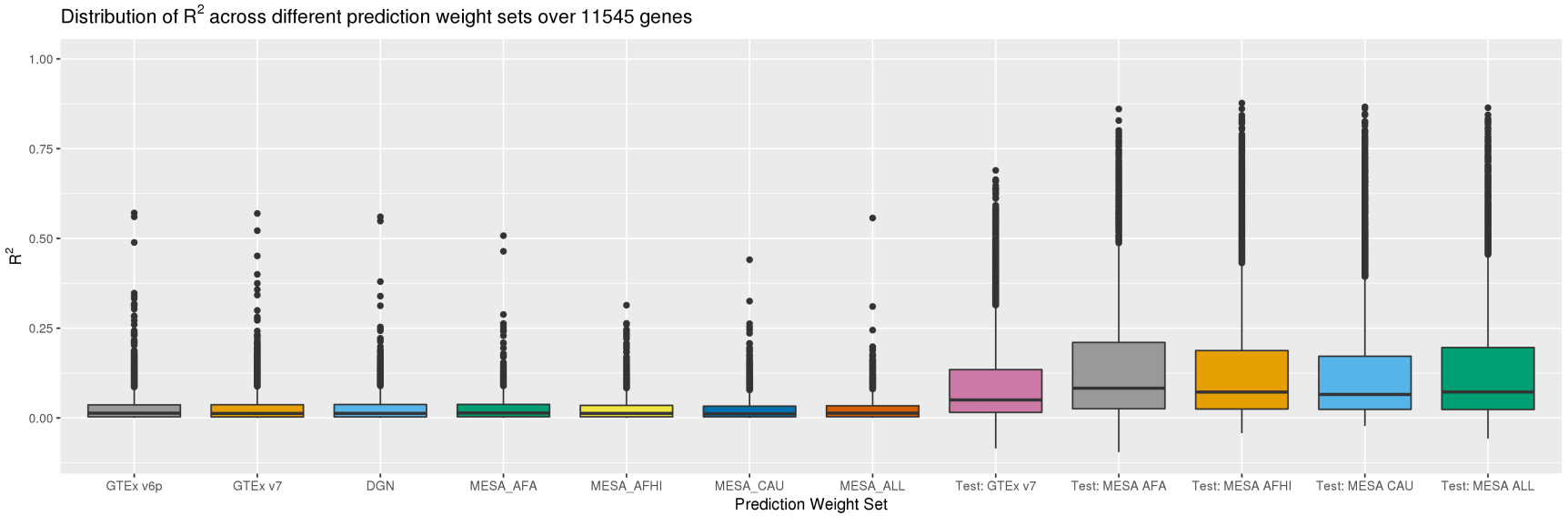
Supplementary Table 5: Prediction performance under fully shared eQTL architecture for k = 10 eQTL yields reliable cross-population gene expression imputation. Results for other sizes of eQTL models are similar.

R^2	AFR to AFR	AFR to EUR	EUR to AFR
AFR to EUR	1.222×10^{-12}		
EUR to AFR	1.705×10^{-24}	6.636×10^{-06}	
EUR to EUR	1.357×10^{-04}	1.487×10^{-112}	1.753×10^{-228}

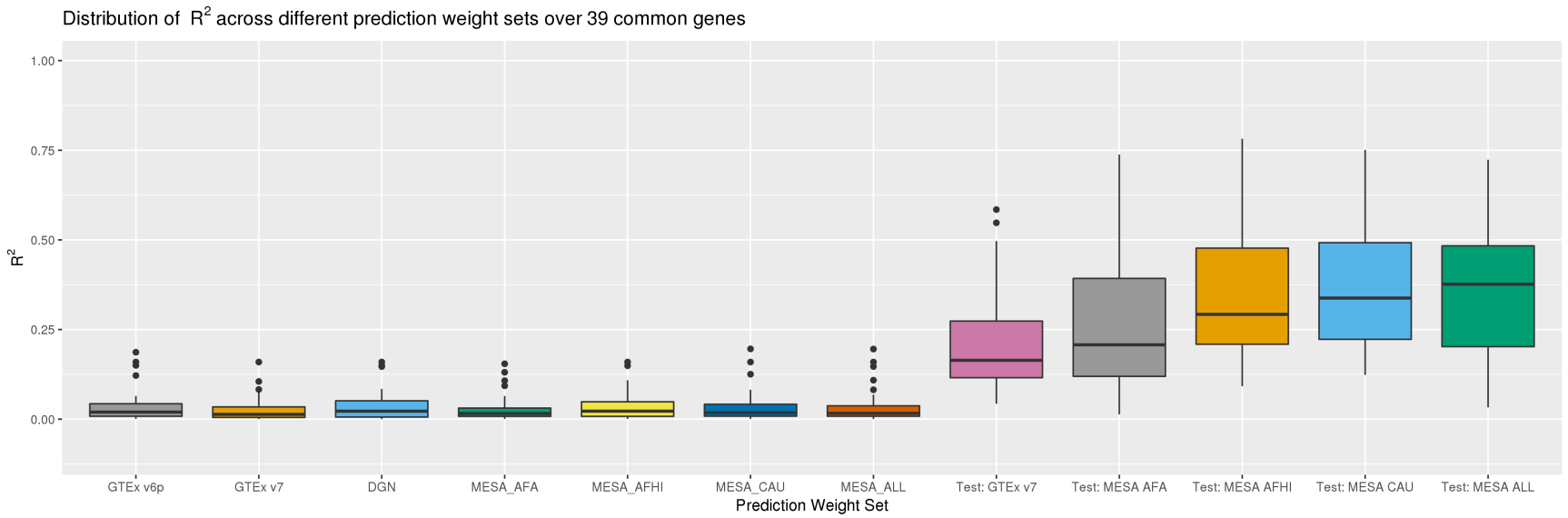
Supplementary Table 6: A Dunn test shows statistically significant differences when imputing between AFR and EUR populations versus imputing between EUR populations.

Z-score (*p*-value)		Train-test direction				
		AA to CEU	**AA to YRI**	**CEU to AA**	**CEU to YRI**	**YRI to AA**
Train-test direction	**AA to YRI**	-28.029 ($p < 5.6 \times 10^{-172}$)	n/a n/a	n/a n/a	n/a n/a	n/a n/a
	CEU to AA	-0.244 (p ~ 1)	27.784 ($p < 5.0 \times 10^{-169}$)	n/a n/a	n/a n/a	n/a n/a
	CEU to YRI	13.373 ($p < 6.5 \times 10^{-40}$)	41.403 (p ~ 0)	13.618 ($p < 2.3 \times 10^{-41}$)	n/a n/a	n/a n/a
	YRI to AA	-28.725 ($p < 1.4 \times 10^{-180}$)	-0.695 (p ~ 1)	-28.480 ($p < 1.5 \times 10^{-177}$)	-42.099 (p ~ 0)	n/a n/a
	YRI to CEU	12.508 (p ~ 0)	40.538 (p ~ 0)	12.753 (p ~ 0)	-0.865 (p ~ 0)	41.234 (p ~ 0)

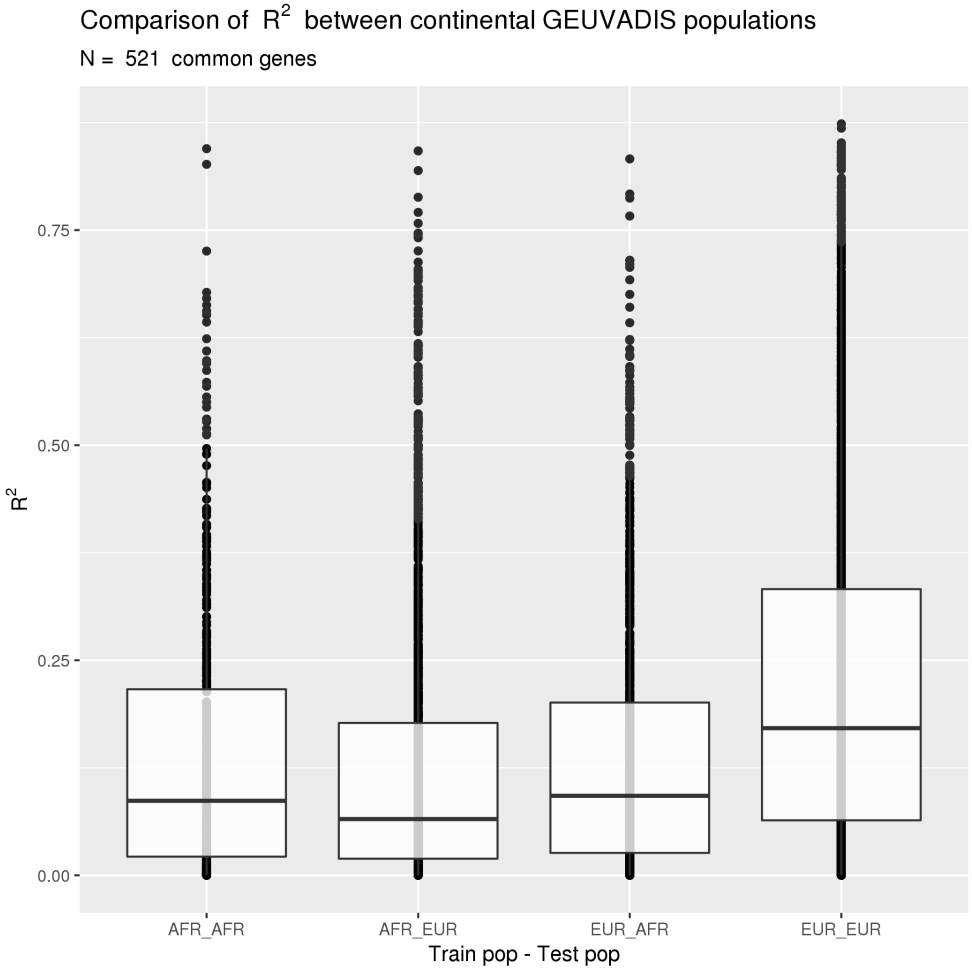
Supplementary Table 7: Differences in cross-population imputation performance are statistically significant, with a few notable exceptions. Imputation between AA and YRI, between AA and CEU, and between CEU and YRI is essentially the same, indicating that the direction of imputation does not matter.



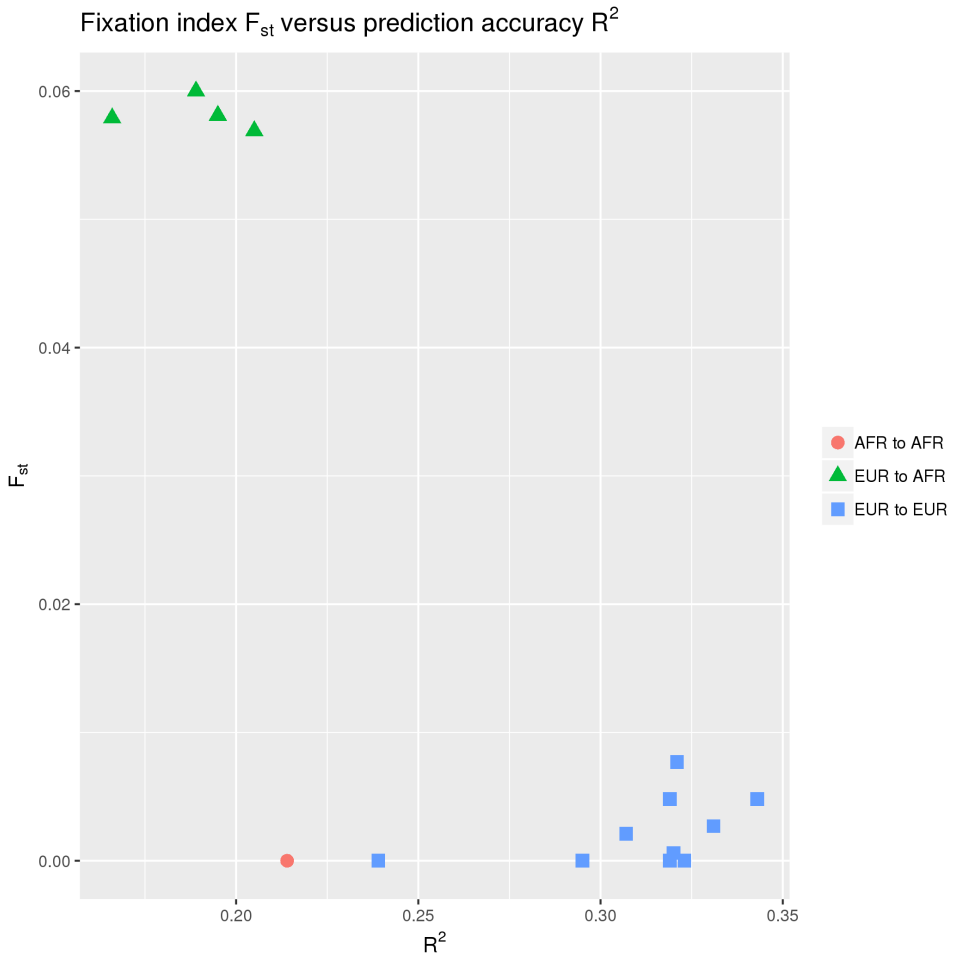
Supplementary Figure 1: A comparison of R^2 between prediction and measurement in SAGE, with PredictDB test metrics as benchmarks, for 11,545 genes total. The number of genes per weight set varies; see Supplementary Table 1.



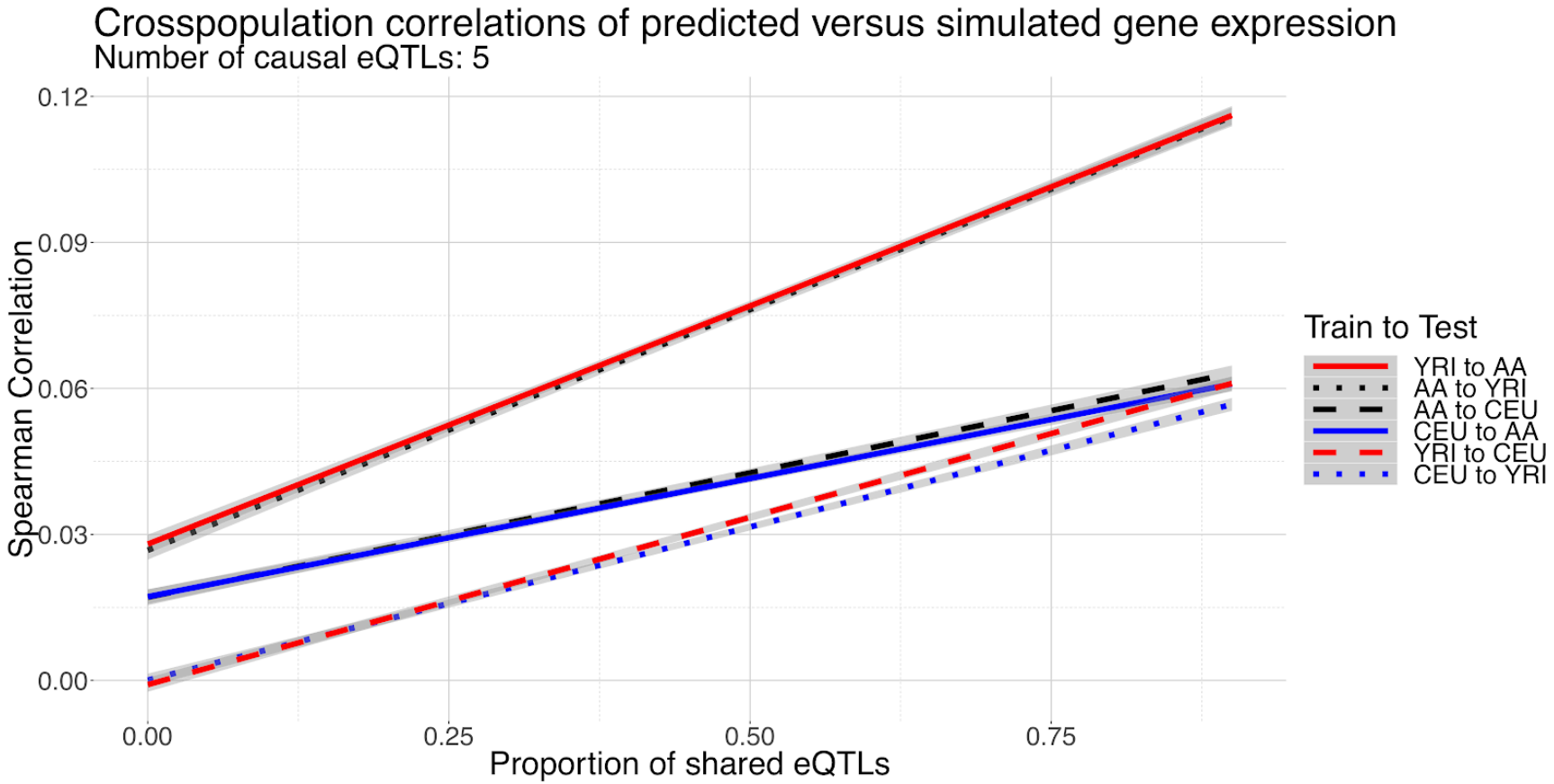
Supplementary Figure 2: R^2 between prediction and measurement in SAGE only using the 39 genes with positive correlation between prediction and measurement in all weight sets and benchmarks.



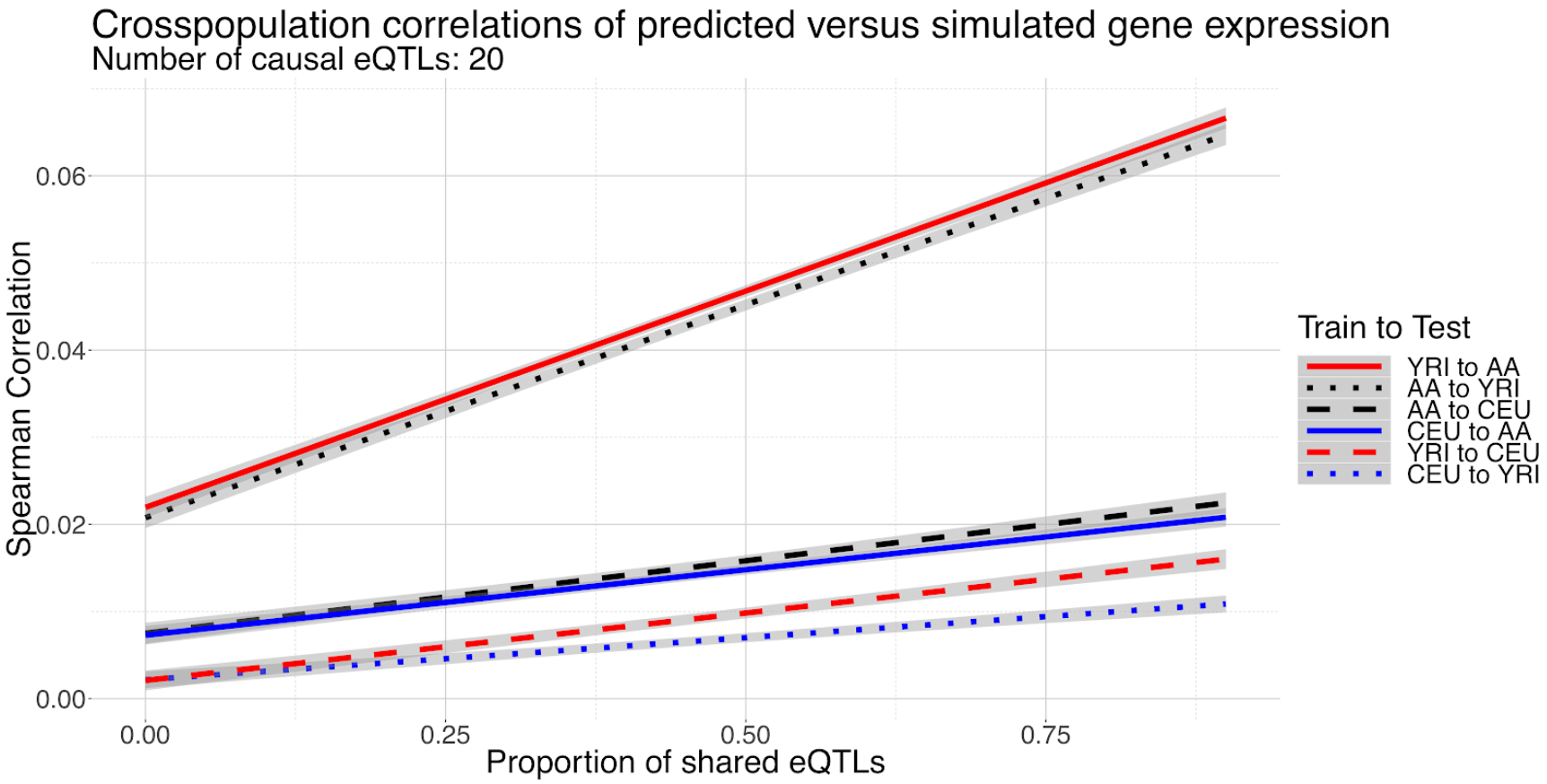
Supplementary Figure 3: Imputation R^2 between AFR (YRI) and EUR (CEU, TSI, GBR, and FIN). Imputing into and from AFR produces consistently lower R^2 than imputing within EUR, suggesting a potential decrease in prediction accuracy when imputing across continental population groups.



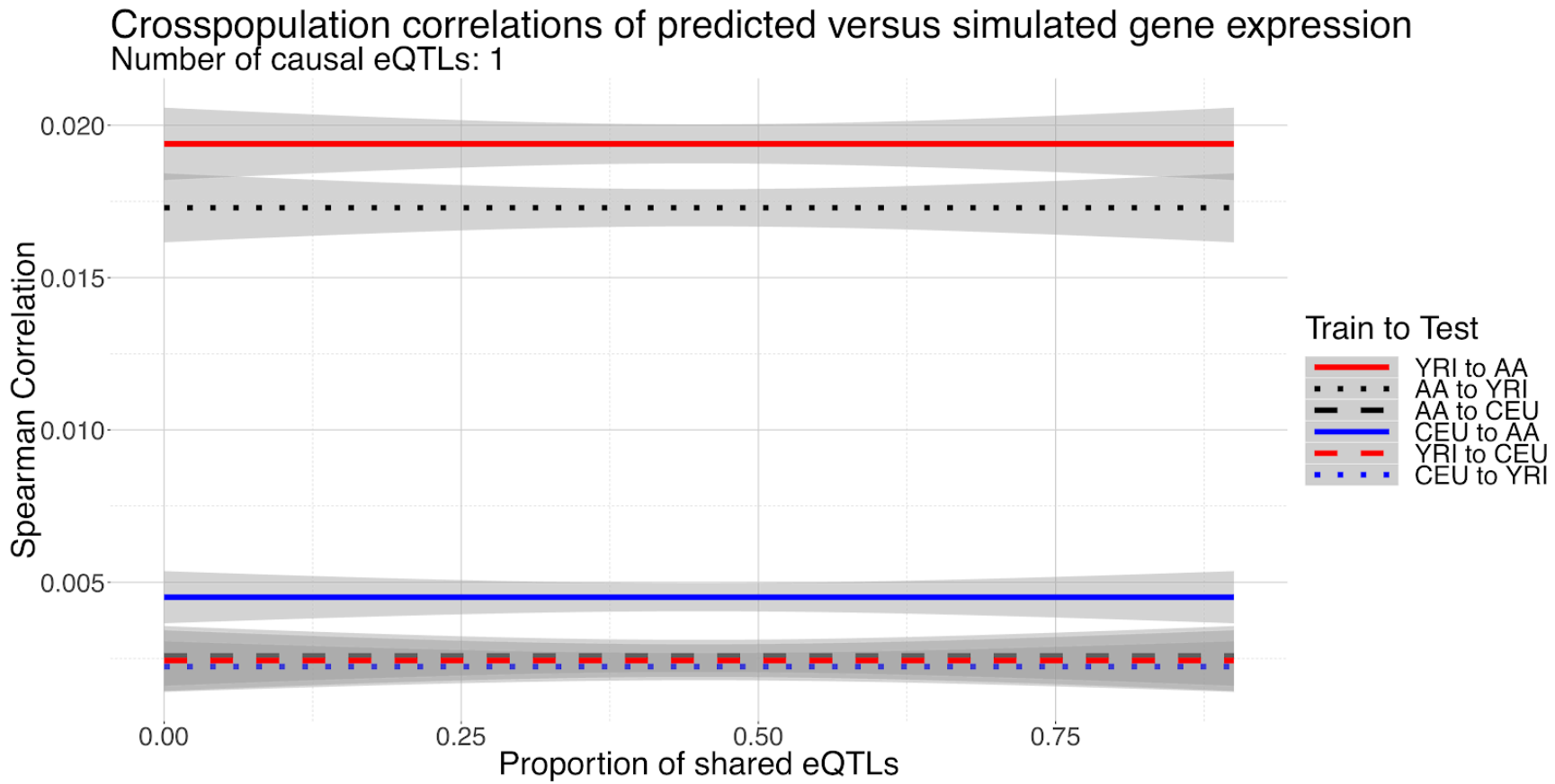
Supplementary Figure 4: Genetic distance versus imputation accuracy over 142 genes with positive correlation across all train-test scenarios. Here the GEUVADIS populations are arranged into three groups. AFR to AFR includes imputation from YRI into itself; EUR to AFR includes imputation into YRI from CEU, GBR, TSI, and FIN; and EUR to EUR includes imputation within and between all European populations in GEUVADIS. Clustering by genetic distance separates imputation between European populations from imputation between European populations and AFR. F_{ST} are taken from the 1000 Genomes Project (Table S11).[63]



Supplementary Figure 5: Correlations between predictions and simulated gene expression measurements from simulated populations across various proportions of shared eQTL architecture with 5 causal cis-eQTL.



Supplementary Figure 6: Correlations between predictions and simulated gene expression measurements from simulated populations across various proportions of shared eQTL architecture with 20 causal cis-eQTL.



Supplementary Figure 7: Correlations between predictions and simulated gene expression measurements from simulated populations across various proportions of shared eQTL architecture with a single causal cis-eQTL.