Title

On the cross-population generalizability of gene expression prediction models

Authors

Kevin L. Keys[1,2,*], Angel C.Y. Mak[1], Marquitta J. White[1], Walter L. Eckalbar[1], Andrew W. Dahl[1], Joel Mefford[1], Anna V. Mikhaylova[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

* Corresponding authors

Email: klkeys@g.ucla.edu (KLK) and chris.gignoux@ucdenver.edu (CRG)

¶ These authors share senior authorship.

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 generalizability of these models to other populations. Here, we test for cross-population generalizability 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 generalizability using the five populations in the GEUVADIS dataset. Via realistic simulations of both populations and gene expression, we show that accurate cross-population generalizability of transcriptome prediction only arises when eQTL architecture is substantially shared across populations. In contrast, models with non-identical eQTLs 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 prediction.

Keywords

TWAS, gene expression, admixed populations, GTEx, PrediXcan

Author summary

Advances in RNA sequencing technology have reduced the cost of measuring gene expression at a genome-wide level. However, sequencing enough human RNA samples for adequately-powered disease association studies remains prohibitively costly. To this end, modern transcriptome-wide association analysis tools leverage existing paired genotype-expression datasets by creating models to predict gene expression using genotypes. These predictive models enable researchers to perform cost-effective association tests with gene expression in independently genotyped samples. However, most of these models use European reference data, and the extent to which gene expression prediction models work across populations is not fully resolved. We observe that these models predict gene expression worse than expected in a dataset of African-Americans when derived from European-descent individuals. Using simulations, we show that gene expression predictive model performance depends on both the amount of shared genotype predictors as well as the genetic relatedness between populations. Our findings suggest a need to carefully select reference populations for prediction and point to a pressing need for more genetically diverse genotype-expression datasets.

Introduction

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*-eQTLs 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, predicted 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

84 expression in very large cohorts (~10^4 – 10^6 individuals) rather than typical gene expression
85 studies that measure expression directly (~10^2 – 10^3 individuals). Several methods have been
86 recently developed to perform TWAS in existing genotyped datasets. PrediXcan[6] uses eQTLs
87 precomputed from paired genotype-expression data, such as those in GTEx, in conjunction with
88 a new genotype set to predict gene expression. These gene expression prediction models are
89 freely available online (PredictDB), creating resources for external researchers. Related TWAS
90 approaches, such as FUSION[5], MetaXcan[9], or SMR[10], leverage eQTL information with GWAS
91 summary statistics instead, thus circumventing the need for raw individual-level genotype data.
92
93 As evidenced by application to numerous disease domains, the TWAS framework is capable of
94 uncovering new genic associations.[11–17] However, the power of TWAS is inherently limited by
95 the data used for eQTL discovery. For example, since gene expression varies by tissue type,
96 researchers must ensure that the prediction weights are estimated using RNA from a tissue
97 related to their phenotype, whether that be the direct tissue of interest or one with sufficiently
98 correlated gene expression.[18] Furthermore, the ability of predictive models to impute gene
99 expression from genotypes is limited by the heritability in the *cis* region around the gene.[6]
100 Consequently, genes with little or no measurable genetically regulated effect on their expression
101 in the discovery data are poor candidates for TWAS.
102
103 A subtler but more troubling issue arises from the lack of genetic diversity present in the datasets
104 used for predictive model training: most paired genotype-expression datasets consist almost
105 entirely of data from European-descent individuals.[8,18] The European overrepresentation in

106 genetic studies is well documented[19–21] and has severe negative consequences for equity as
107 well as for gene discovery[22], fine mapping[23–25], and applications in personalized
108 medicine.[26–34] Genetic architecture, linkage disequilibrium, and genotype frequencies can
109 vary across populations, which presents a potential problem for the application of predictive
110 models with genotype predictors across multiple populations.
111
112 The training data for most models in the models derived from PrediXcan weights in PredictDB
113 (predictdb.org) are highly biased toward European ancestry: GTEx version v6p subjects are over
114 85% European, while the GTEx v7 and DGN subjects are entirely of European descent. The lack
115 of suitable genotype-expression datasets in non-European individuals leads to scenarios in which
116 PredictDB models trained in Europeans are used to predict into non-European or admixed
117 populations. As shown previously in the context of polygenic risk scores[35], multi-SNP prediction
118 models trained in one population can suffer from unpredictable bias and poor prediction
119 accuracy that impair their cross-population generalizability. Recent analyses of genotype-
120 expression data from the Multi-Ethnic Study of Atherosclerosis (MESA)[36–38], which includes
121 non-European individuals, explore cross-population transcriptome prediction and conclude that
122 predictive accuracy is highest when training and testing populations match in ancestry. These
123 results are consistent with our experience analyzing admixed populations, but offer little insight
124 into the mechanisms underlying the cross-population generalizability of transcriptome prediction
125 models, particularly when eQTL architecture is known.
126

Here we investigate the cross-population generalizability 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.[4,43,44] 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.[45] To our knowledge, nobody has investigated cross-population generalizability of new prediction models generated 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) that, despite some variation in when cells were collected[46], are derived from similar sampling efforts and treatments, thereby 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.

Results

Concordance of measured gene expression and PrediXcan predictions is lower than expected

We compared transcriptome prediction 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: since the estimation of these prediction models requires both high quality expression data and inferred eQTLs, 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. After applying the recommended filters, the prediction results across all seven weight sets overlapped at 273 genes, of which 39 genes had predictions with positive correlation to measurements. This small number of genes in common is largely driven by MESA_AFA, the repository with the smallest number of predictive models. However

MESA_AFA contains the models that should best reflect the genetic ancestry of African Americans 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.

Here, we highlight the union of genes across model sets for investigation. The concordance between predicted and measured gene expression over the union of 11,545 from 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 0.0298 was observed in MESA_AFA. We note the intersection of all prediction models is limited, but reflects a similar pattern: results for the 273 common genes (Supplementary Figure 1) and the 39 genes with positive correlations (Supplementary Figure 2) showed little difference in the 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 prediction R^2 due to out-of-sample estimation. 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 gene expression predictions with negative correlations, but average performance across genes varied. The least negative mean correlation (0.00707) was observed with MESA_AFHI (MESA African Americans and Hispanics), while the

most negative mean correlation (-0.00415) was observed with MESA_AFA (MESA African Americans, Supplementary Table 1). The observation that correlations to SAGE measurements are sometimes negative on average suggests that some large R^2 values seen in Figure *1* may result from gene models with incorrect direction of prediction, thereby limiting interpretability of results. While there are some fluctuations in prediction accuracy, no prediction weight set produces practically meaningfully better correlations to data than the others (-0.00415 to 0.00707). In contrast, the published models for these genes show positive correlations to their training data, ranging from 0.308 in GTEx_v7 to 0.379 in MESA_AFA, indicating no obvious incapacity for accurate prediction, even with out-of-sample data. 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 high prediction R^2 in the original experiment, 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]).

Cross-population prediction quality declines with increasing genetic distance

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[47] or cell lines[46], and the postmortem interval to tissue collection (for GTEx).[48–50] To investigate cross-population

generalizability 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[51], 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 expression data, generated and harmonized together by the GEUVADIS Consortium, with matched 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 prediction 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 prediction 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 prediction 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 subset of genes with potentially more consistent gene expression levels. In this case involving better-predicted genes, we see that prediction quality between the European groups improves noticeably (*p*-value ~ 0, Dunn test) with R^2 ranging between 0.183 to 0.216, while R^2 between Europeans and Africans ranges from 0.095 to 0.147, a significant improvement (*p*-value < 7.07 x 10^{-22}, Dunn test) that nonetheless highlights a continental gap in prediction performance. In general, populations seem to predict better into themselves, and less well into other populations.

Combining all European subpopulations obscures population structure and can complicate analysis of cross-population prediction 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 prediction R^2 from each population into itself and others. The populations consistently predict well into themselves, with prediction R^2 ranging from 0.104 – 0.136. We observe that prediction quality using models trained in CEU shows a miniscule decline relative to other EUR subpopulations. This observation is potentially due to the older age of CEU LCLs[45,52,53], but did not appreciably change our results. In contrast, a more 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. Prediction 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 prediction into YRI ranges from 0.166 to 0.244, while R^2 within EUR populations ranges from 0.239 to 0.331. These results clearly suggest problems for prediction models that predict 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 prediction quality (Supplementary Figure 4).

Admixture influences cross-population gene expression prediction quality under known eQTL architecture

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 eQTLs 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[54] CEU and YRI using HAPGEN2[55] and then sample haplotypes in proportions consistent with realistic admixture proportions (80% YRI, 20% CEU)[56] to construct a simulated African-American (AA) admixed population. We simulate eQTL architectures under an additive model of size k causal alleles (k = 1, 10, 20, and 40) and an

281 expression phenotype with *cis*-heritability h^2 = 0.15 (recapitulating average h^2 in GTEx) using the
282 genomic background of genic regions on chromosome 22, thus testing various model sizes and
283 LD patterns. To tease apart the effect of shared eQTL architecture, we allow the two ancestral
284 populations CEU and YRI to share eQTLs with fixed effects in various proportions (0%, 10%, 20%,
285 …, 100%) to test a range of eQTL architectures. The admixed population AA always inherited all
286 eQTLs from the two ancestral populations, which yielded different numbers of eQTLs per gene
287 depending on how many eQTLs were shared by CEU and YRI. For example, for eQTL model size *k*
288 = 10, when CEU and YRI shared all 10 eQTLs, then all three populations had the exact same 10
289 eQTLs. When CEU and YRI shared half of their eQTLs with each other, then each one had 5
290 population-specific eQTLs, and AA inherited 15 total eQTLs (5 unique to CEU, 5 unique to YRI, and
291 5 shared). If CEU and YRI shared no eQTLs, then all eQTLs were population-specific, and AA
292 inherited 20 eQTLs (10 from CEU and 10 from YRI; see Supplementary Figure 5 for an illustration).
293 With these simulations providing known architectures for comparison, we then apply the train-
294 test-validate scheme as before.
295
296 Figure 4 shows the cross-population Spearman correlations between predicted and simulated
297 phenotypes in our simulated AA, CEU, and YRI, partitioned by proportion of shared eQTLs, for *k*
298 = 10 causal eQTLs. Scenarios with k = 20 and k = 40 causal eQTLs show similar trends
299 (Supplementary Figure 6 and Supplementary Figure 7). Prediction within a population produced
300 similar correlations in all cases, ranging from 0.310 to 0.338 (Supplementary Table 4). The case
301 of models with 100% shared eQTL architecture – where eQTL positions and effects are exactly
302 the same between the ancestral populations – yields predictions with no loss in cross-population

generalizability, with correlations ranging from 0.299 to 0.336 even when predicting across populations (Supplementary Table 5). This case suggests that eQTLs that are causal in all populations can impute gene expression reliably regardless of the population in which they were ascertained, provided that the eQTLs can be correctly mapped and genotyped in all populations, that the eQTL effects are identical across populations, and that a linear model of eQTLs is assumed. For cases where eQTL architecture is not fully shared across populations, we see that prediction from each population into the other improves as the proportion of shared eQTLs increases (Figure 4). The cross-population correlation between predicted gene expression versus measurement is highest from YRI to AA (0.238 to 0.338), intermediate from CEU to AA (0.218 to 0.310), and lowest between CEU and YRI (0.0020 to 0.326). Prediction quality from AA to CEU and YRI interpolates that of YRI to AA and CEU to AA, with correlations ranging from 0.223 to 0.338. Prediction quality from AA to CEU or YRI shows a slight upward trend as more eQTLs are shared, an artifact of eQTL inheritance in our simulations; as described previously, AA eQTL models are largest (20 eQTLs) when CEU and YRI share no eQTLs and smallest (10 eQTLs) when CEU and YRI share all eQTLs. Consequently, when predicting between two populations, the choice of which population is used to train predictive models can produce differences in prediction quality. Prediction quality between AA to CEU and AA to YRI is not significantly different (p-value ~ 1, Dunn test). All other train/test scenarios are significantly different from each other (Supplementary Table 7). The results for k = 10, 20, and 40 eQTLs show a consistent trend of prediction quality driven primarily by different in eQTL architecture, with additional minor influence from ancestral similarity between populations (k = 10, Figure 4, similar plots in Supplementary Figure 6 and Supplementary Figure 7). Although less realistic for most

genes[5,6,18], we also analyzed models with a single causal eQTL. Trends for single-eQTL models are more difficult to analyze due the limitations of binary inference as to whether the causal SNP is identified or not. Nevertheless, when the causal eQTL is identified and shared across populations, prediction quality is high in call cases. If the causal eQTL differs across populations, then cross-population prediction between AA and YRI or CEU is noticeably better than prediction between CEU and YRI (Supplementary Figure 8), in line with results for other values of *k* that suggest that eQTL sharing is the primary driver of gene expression prediction quality.

Power to detect associations declines with decreasing shared ancestry

Simulation of gene expression demonstrates that gene expression prediction quality is modulated by both shared eQTL architecture and shared genetic ancestry. These results suggest possible effects of cross-population generalizability on the power to detect associations between a phenotype and gene expression measures in a TWAS. For each of our three populations (AA, CEU, and YRI), we used the simulated gene expression measures to simulate a continuous phenotype whose variation depends on expression of a single causal gene. For simplicity, the phenotypes shared the same causal gene, the same effect size, and the same environmental noise model. We tested various effect sizes from 1×10^{-5} to 1 and drew the environmental noise from a zero-mean normal distribution with variance 0.01. The effect sizes produced a continuous spectrum of genetic heritability values h^2 spanning the full range of heritability for gene expression. We then regressed the phenotype onto predicted gene expression measures, resulting in nine association tests, one for each train-test scenario. For simplicity, we focused on the prediction scenario with $k = 10$ causal eQTLs per gene. To see how shared eQTL architecture

346 affects power, we used predicted expression measures with 0%, 50%, and 100% shared eQTLs
347 per gene.
348
349 Figure 5 shows power curves for the association tests for the nine prediction scenarios for all
350 three tested eQTL architectures. Unsurprisingly, power improves as populations share more
351 causal eQTLs. For example, with 100% shared eQTLs and phenotypic heritability 0.205, cross-
352 population power ranges between 0.69 – 0.86. In contrast, average power under a 0% shared
353 eQTL model ranges more broadly, from 0.02 (CEU to YRI, YRI to CEU) to 0.38 (AA to YRI, AA to
354 CEU) to 0.82 (CEU to AA) and 0.88 (YRI to AA), indicating some ability to predict gene expression
355 at genetically controlled genes, even without shared eQTLs. Power also improves with shorter
356 genetic distance between populations. Figure 6, which is a cross-section of Figure 5, shows power
357 for each train-test scenario across various shared eQTL architectures for $\beta = 0.05$, corresponding
358 to a phenotype heritability of $h^2 = 0.205$, indicating moderate genetic control. TWAS in this case
359 using gene expression imputed from matched populations has higher power across all eQTL
360 architectures, from 0.33-0.85, compared to cross-population TWAS, where power varies
361 substantially. For an architecture with no shared eQTLs, power between CEU and YRI is 0, while
362 power is higher for CEU to AA (0.25) and YRI to AA (0.30). TWAS power for expression imputed
363 from AA to CEU (0.05) or YRI (0.06) is much lower due to the aforementioned structure of eQTL
364 inheritance. As the proportion of shared eQTLs jumps from 0% to 50% and 100%, power increases
365 across all cross-population scenarios, reaching up to 0.31 (YRI to AA, 100% shared eQTLs). When
366 eQTLs are fully shared, power from YRI to AA (0.31) is higher than from CEU to AA (0.25),
367 indicating an effect of genetic distance on prediction quality. Indeed, when controlling for eQTL

368 architecture, increasing genetic similarity between reference and target populations yields more
369 significant median association test *t*-statistics (Supplementary Figure 9).

370 **Admixture proportion interpolates power in two-way admixture**

371 The results in Figure 6 show how genetic distance affects power in TWAS association tests for
372 one particular admixture proportion, but offer limited insight about how power changes across
373 the admixture spectrum. To understand how admixture proportion affects TWAS power in a
374 general admixed population with two ancestral populations, we simulated multiple admixed
375 populations from CEU and YRI with admixture proportions varying at 10% increments. When the
376 admixed population has 0% YRI admixture, it is fully drawn from haplotypes from CEU, whereas
377 a population with 100% YRI admixture is drawn exclusively from haplotypes from YRI. It is
378 important to note that in neither case does the admixed population exactly match the reference
379 CEU or YRI since the genotypes for the admixed population are formed from an independent
380 shuffling of the CEU or YRI haplotypes. For each admixed population, we estimated prediction
381 models of gene expression as done in our previous analyses. For computational efficiency, we
382 investigated the scenario of 50% shared eQTLs across reference populations and the number of
383 eQTLs per gene to 10. Populations still shared the same causal gene, effect size, and
384 environmental noise model.
385
386 Figure 7 shows power across admixture proportions for all cross-population scenarios. The
387 phenotypes were simulated at effect sizes β = 0.005, 0.01, and 0.025, and environmental
388 variance σ^2 = 0.01, corresponding to heritability h^2 = 0.06, 0.20, and 0.58, respectively. To avoid

389 confusion with previous references to AA, which had a fixed admixture proportion, here we
390 denote the admixed population for all proportions as AD. As expected, statistical power
391 increases with the genetic heritability of the phenotype for all prediction scenarios. However,
392 the different admixture proportions yield directional changes in power when gene expression is
393 predicted to or from AD. For example, when $h^2 = 0.20$ and gene expression is predicted from AD
394 to CEU, power at 0% YRI admixture is 0.56 (95% CI: 0.462 – 0.658) and declines linearly with
395 increasing YRI admixture; at 100% YRI, statistical power for AD to CEU is 0.46 (95% CI: 0.362 –
396 0.558). For AD to YRI, power at 0% YRI admixture starts at 0.42 (95% CI: 0.323 – 0.512) and
397 increases linearly to 0.53 (95% CI: 0.431 – 0.628) at 100% YRI. We observe similar changes in
398 power for CEU to AD (decreasing power as YRI proportion increases) and YRI to AD (increasing
399 power as YRI proportion increases). The four directional trends also hold for $h^2 = 0.06$ and $h^2 =$
400 0.58, though power for cross-population scenarios involving AD is much lower in the former
401 case and almost universally high in the latter case. In essence, the varying admixture
402 proportions in this two-way admixed population yield a continuous linear trend of statistical
403 power between the two ancestral populations: when AD is genetically closer to CEU, power for
404 gene expression predicted these populations is highest, and declines as AD becomes genetically
405 closer to YRI. Similarly, when predicting from AD to YRI or vice versa, power is lowest when the
406 two populations are genetically distinct, intermediate as the two populations become more
407 genetically similar, and maximized when they are most alike.

408 ## Discussion

409 Our goal with this study was to understand the extent to which gene expression prediction
410 models estimated in one population can accurately predict the genetic component of gene

411 expression in a different population. Cross-population generalizability of gene expression
412 prediction models is an important but understudied issue for TWAS analyses. Among TWAS
413 resources, we focused on PrediXcan as a test case with openly distributed prediction models
414 available for multiple populations.[6,38] Using 39 subjects from the SAGE study[39–42] we
415 compared predicted expression values from PrediXcan models to measured gene expression on
416 the same subjects and found that predictions matched poorly to measurements. Our
417 investigation with the GEUVADIS dataset[43] offered us a more homogenous environment in
418 which to train and test gene expression prediction models. Prediction quality in GEUVADIS using
419 both continental and constituent subpopulations provided stronger evidence of cross-population
420 generalizability issues with transcriptome prediction, but could not control for eQTL predictors
421 that vary between populations. To that end, our simulation of an admixed population from 1000
422 Genomes CEU and YRI haplotypes[4,44] allowed us to finely control eQTL positions and effects
423 as well as the causal genes in a TWAS. The simulation results show that both gene expression
424 prediction accuracy and statistical power decrease as population eQTL models begin to diverge
425 and genetic distance increases between populations for varying admixture proportions.
426
427 Our results highlight two points: firstly, since prediction within populations is better than
428 prediction between populations, our results reaffirm prior investigations[38] that population
429 matching matters for optimally predicting gene expression. This is consistent with our results of
430 impaired transcriptome prediction performance in SAGE with currently available resources.
431 Secondly, despite decreased prediction accuracy when predicting between different populations,
432 the populations that are more closely genetically related demonstrate somewhat better cross-

433 population prediction and power to detect associations in TWAS. Our simulations of prediction
434 between ancestral populations and an admixed one under varying admixture proportions neatly
435 summarize this relationship: the admixture proportion from each ancestral population
436 interpolates the power available from each ancestral population, and power is maximized when
437 the admixed population is most closely related to one or the other ancestral population.
438 However, while the differences in power under varying admixture are statistically meaningful,
439 they are smaller than differences attributable to different eQTL architectures or to different
440 levels of genetic heritability of a phenotype.
441
442 Prediction results from GTEx, DGN, and MESA into SAGE suggest that current predictive models,
443 even for genes with greater heritability, perform worse than expected despite matching tissue
444 types. Our investigation into cross-population prediction accuracy with GEUVADIS data replicates
445 this lack of cross-population generalizability as observed with current predictive models from
446 PredictDB, demonstrating that heterogeneity in RNA-Seq protocols does not fully explain our
447 observations. Since transcriptome prediction models use multivariate genotype predictors
448 trained on a specific outcome, the impaired cross-population application can be viewed as an
449 analogous observation to that seen previously in polygenic scores.[35]
450
451 Our simulations control for many technical issues that are otherwise difficult to overcome with
452 real data, such as oracular knowledge of positions and effect sizes of causal eQTLs. Nevertheless,
453 in our simulations we see issues with cross-population prediction that we first observed when
454 applying existing PrediXcan models to SAGE genotype data. Certainly, SAGE differs in important

ways from GTEx, DGN, and MESA: SAGE is a pediatric asthma case-control cohort study in African-American children, so we cannot rule out technical heterogeneity introduced by differences in age, study design, and ethnicity. Furthermore, our SAGE sample includes RNA-Seq data for $n = 39$ subjects, a dataset leveraged previously to validate genetic associations, but is nevertheless somewhat small by contemporary standards.[39] However, technical heterogeneity between SAGE and existing PrediXcan models cannot solely explain the poor prediction performance. Our simulation results strongly suggest that problematic cross-population prediction performance between PrediXcan models and SAGE is deeper than differences in expression data.

Our investigations into the architecture of gene expression indicate that the power to detect associations is primarily determined by the degree of shared eQTLs across populations. In our simulations, this can be approximated as a (quasi-)linear interpolation of the prediction in the ancestral or reference populations into the admixed populations. However, the same is not true of overall levels of power in the admixed population: under 100% shared eQTL scenarios, cross-population generalizability is high, so the choice of training population matters less. In practical terms, this result bodes well for prediction of genes with eQTLs that do not vary by population. It is curious that in high-heritability genes, even models that share no eQTLs still retain power to detect scenarios: for genetically distant populations (CEU and YRI), power ranges from 0.10-0.14. Without shared eQTLs, this implies that local linkage disequilibrium between population-specific eQTLs, combined with high heritability, enables some degree of cross-population prediction. When cross-population statistical power is driven by LD and h^2 instead of expression signals, then

476 subsequent interpretation of association hits, such as direction and strength of effect, becomes
477 difficult to link to actual biological relationships between phenotype and gene expression.
478
479 It is important to note that our observations do not reflect shortcomings of either the initial
480 PrediXcan or TWAS frameworks. Nor do our findings affect the positive discoveries made using
481 these frameworks over the past several years. These methods fully rely on the data used as input
482 for training, and the most commonly used datasets for model training are overwhelmingly of
483 European descent. Here we note that the current models fail to capture the complexity of the
484 cross-population genomic architecture of gene expression for populations of non-European
485 descent. Failing to account for this could lead researchers to draw incorrect conclusions from
486 their genetic data, particularly as these models would lead to false negatives.
487
488 To this end, our simulations strongly suggest that predicting gene expression in a target
489 population is improved by using predictive models constructed in a genetically similar training
490 population. Maximizing prediction quality crucially depends on both genetic architecture and
491 eQTL architecture. If populations share the exact same eQTL architecture, then they are
492 essentially interchangeable for the purposes of gene expression prediction so long as eQTLs are
493 genotyped and accurately estimated, which remains a technological and statistical challenge. As
494 the proportion of shared eQTL architecture decreases between two populations, both cross-
495 population prediction quality and TWAS power decrease as well. In both SAGE and GEUVADIS,
496 we observe cross-population patterns consistent with an imperfect overlap of eQTLs across
497 populations. Ensuring representative eQTL architecture for all populations in genotype-

498 expression repositories will require a solid understanding of true cross-population and
499 population-specific eQTLs. However, expanding the amount of global genetic architecture
500 represented in genotype-expression repositories, which can be accomplished by sampling more
501 populations, provides the most desirable course for improving gene expression prediction
502 models. Additionally, this presents an opportunity for future research in methods that could
503 improve cross-population generalizability, particularly when one population is over-represented
504 in reference data. Tools from transfer learning could facilitate porting TWAS eQTL models from
505 reference populations to target populations using little or no RNA-Seq data.
506
507 In light of the surging interest in gene expression prediction and TWAS, we see a pressing need
508 for freely distributed predictive models of gene expression estimated from coupled
509 transcriptome-genome data sampled in a variety of populations and tissues. The recently
510 published predictive models with multi-ethnic MESA data constitute a crucial first step in this
511 direction for researchers working with admixed populations.[38] However, the clinical and
512 biomedical research communities must push for more diverse genotype-expression resources to
513 ensure that the fruits of genomic studies benefit all populations.

514 **Online Resources**

515 PredictDB: http://predictdb.org/

516 GTEx: http://gtexportal.org/

517 DGN: http://dags.stanford.edu/dgn/

518 GEUVADIS: https://www.ebi.ac.uk/Tools/geuvadis-das/

519 Source code: https://github.com/asthmacollaboratory/sage-geuvadis-predixcan

520 Results and simulation data: https://ucsf.box.com/v/sage-geuvadis-predixcan

521

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.[57] Genotypes were then imputed on the Michigan Imputation Server[58] with EAGLE v2.3[59] and the 1000 Genomes panel phase 3 v5[44] 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.[6,9,38,60] 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 prediction quality by comparing PrediXcan predictions to normalized gene expression from SAGE using linear regression and correlation tests.

Estimation of 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

566 models cannot easily differentiate predictions of 0 (no gene expression) and NA (missing
567 expression). We addressed this with two additional filters. Firstly, we removed genes that did not
568 have any eQTLs in their predictive models. Secondly, genes where fewer than half of the
569 individuals had nonmissing predictions were removed from further analysis. Coefficients of
570 determination (R^2) were computed with the `lm` function in R. Spearman correlations were
571 computed with the `cor.test` function in R.

572 **Simulation of gene expression**

573 We downloaded a sample of 20,085 HapMap 3 SNPs[54] from each of CEU and YRI on
574 chromosome 22 as provided by HAPGEN2.[55] The data include 234 phased haplotypes for CEU
575 and 230 phased haplotypes for YRI. We forward-simulated from these haplotypes to obtain two
576 populations of n = 1000 individuals each. We then sampled haplotypes in proportions of 80% YRI
577 and 20% CEU to obtain a mixture of CEU and YRI where the ancestry patterns roughly mimic
578 those of African Americans. For computational simplicity, and in keeping with the high ancestry
579 LD present in African Americans[61,62], for each gene we assumed local ancestry was constant
580 for each haplotype. For each of the three simulated populations, we applied the same train-test-
581 validate scheme used for cross-population analysis in GEUVADIS. Genetic data for model
582 simulation were downloaded from Ensembl 89 and included the largest 100 genes from
583 chromosome 22. We defined each gene as the start and end positions corresponding to the
584 canonical transcript, plus 1 megabase in each direction. Two genes, PPP6R2 and MOV10L1,
585 spanned no polymorphic markers in our simulated data, resulting in 98 gene models used for
586 analysis. To simulate predictive eQTL models, we tested multiple parameter configurations for
587 each gene: we varied the number of causal eQTL (k = 1, 10, 20, and 40) and the proportion of

shared eQTL positions (p = 0.0, 0.1, 0.2, …, 0.9, 1) between ancestral populations. The admixed population always inherited all eQTLs from the ancestral populations. 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.

Simulation of TWAS

Using the simulated gene expression measures with k = 10 eQTLs per gene, we simulated a continuous phenotype with a known genetic architecture that depended on 1 causal gene. We tested prediction scenarios with 0, 5, and 10 eQTLs shared across populations. For each eQTL architecture, the three populations AA, CEU, and YRI shared the same causal gene G, the same causal effect size β, and the same environmental noise ε. G was chosen randomly. Effect sizes were fixed, and we tested various effect magnitudes $\beta = 1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, \ldots, 1 \times 10^{-1}, 5 \times 10^{-1}, 1$. The environmental noise ε was drawn from an $N(0,0.1^2)$ distribution. Consequently, phenotypes therefore only varied with the expression measures from G. For a given population c, the phenotype y_c was then simulated as

$$y_c = G\beta + \varepsilon.$$

For each combination of shared eQTL architecture, G, and β, this procedure yielded one y_c per individual in a population. We then performed a TWAS with y_c onto the *predicted* gene expression values, yielding three TWAS per y_c, one for each reference prediction population. We then queried the resulting association p-value at G and tabulated whether it was declared significant (yes) or not (no) against a Bonferroni-corrected threshold of 0.05 / 98, accounting for all 98 genes in the TWAS. We ran this procedure for 100 random instantiations of (G, ε) and computed association test power with a logistic interpolation of the yes/no results.

610 **Analysis tools**

611 Analyses used GNU `parallel`[63]. The R packages used for analysis include `argparser`,
612 `assertthat`, `data.table`, `doParallel`, `dunn.test`, `knitr`, `optparse`, `peer`, the
613 Bioconductor packages `annotate`, `biomaRt`, and `preprocessCore`, and the `tidyverse`
614 bundle.[64–75] All plots were generated with `ggplot2`.[76]

615 ## ACKNOWLEDGEMENTS

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

628

629 The authors wish to acknowledge the following SAGE co-investigators for subject recruitment,
630 sample processing and quality control: Luisa N. Borrell, DDS, PhD, Emerita Brigino-Buenaventura,
631 MD, Adam Davis, MA, MPH, Michael A. LeNoir, MD, Kelley Meade, MD, Fred Lurmann, MS and

632 Harold J. Farber, MD, MSPH. The authors also wish to thank the staff and participants who
633 contributed to the SAGE study.

634

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

641

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

645

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

648

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

650 **Conflict of Interest Statement**

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

REFERENCES

1. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015;12. doi:10.1371/journal.pmed.1001779
2. NHLBI Trans-Omics for Precision Medicine [Internet]. [cited 13 Nov 2018]. Available: https://www.nhlbiwgs.org/
3. NHGRI Genome Sequencing Program (GSP). In: National Human Genome Research Institute (NHGRI) [Internet]. [cited 13 Nov 2018]. Available: https://www.genome.gov/10001691/nhgri-genome-sequencing-program-gsp/
4. The 1000 Genomes Consortium. An integrated map of genetic variation from 1,092 human genomes | Nature [Internet]. [cited 13 Nov 2018]. Available: https://www.nature.com/articles/nature11632
5. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48: 245–252. doi:10.1038/ng.3506
6. 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. Nat Genet. 2015;47: 1091–1098. doi:10.1038/ng.3367
7. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585. doi:10.1038/ng.2653
8. 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 Res. 2014;24: 14–24. doi:10.1101/gr.155192.113
9. Barbeira AN, Dickinson SP, Torres JM, Bonazzola R, Zheng J, Torstenson ES, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9. doi:10.1038/s41467-018-03621-1
10. 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. Nat Genet. 2016;48: 481–487. doi:10.1038/ng.3538
11. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat Neurosci. 2018;21: 811. doi:10.1038/s41593-018-0154-9
12. Ferreira MAR, Jansen R, Willemsen G, Penninx B, Bain LM, Vicente CT, et al. Gene-based analysis of regulatory variants identifies four putative novel asthma risk genes related to nucleotide synthesis and signaling. J Allergy Clin Immunol. 2017;139: 1148–1157. doi:10.1016/j.jaci.2016.07.017
13. Lamontagne M, Bérubé J-C, Obeidat M, Cho MH, Hobbs BD, Sakornsakolpat P, et al. Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. Hum Mol Genet. 2018;27: 1819–1829. doi:10.1093/hmg/ddy091
14. Thériault S, Gaudreault N, Lamontagne M, Rosa M, Boulanger M-C, Messika-Zeitoun D, et al. A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. Nat Commun. 2018;9: 988. doi:10.1038/s41467-018-03260-6
15. Porcu E, Rüeger S, Consortium eQTLGen, Santoni FA, Reymond A, Kutalik Z. Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. bioRxiv. 2018; 377267. doi:10.1101/377267

16. Gusev A, Lawrenson K, Segato F, Fonseca M, Kar S, Lee J, et al. Multi-Tissue Transcriptome-Wide Association Studies Identify 21 Novel Candidate Susceptibility Genes for High Grade Serous Epithelial Ovarian Cancer. bioRxiv. 2018; 330613. doi:10.1101/330613
17. Huckins LM, Dobbyn A, Ruderfer D, Hoffman G, Wang W, Pardinas AF, et al. Gene expression imputation across multiple brain regions reveals schizophrenia risk throughout development. bioRxiv. 2017; 222596. doi:10.1101/222596
18. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550: 204–213. doi:10.1038/nature24277
19. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011;475: 163–165. doi:10.1038/475163a
20. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538: 161–164. doi:10.1038/538161a
21. Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? J Community Genet. 2017;8: 255–266. doi:10.1007/s12687-017-0316-6
22. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nat Rev Genet. 2018;19: 175–185. doi:10.1038/nrg.2017.89
23. Asimit JL, Hatzikotoulas K, McCarthy M, Morris AP, Zeggini E. Trans-ethnic study design approaches for fine-mapping. Eur J Hum Genet. 2016;24: 1330–1336. doi:10.1038/ejhg.2016.1
24. Wang X, Cheng C-Y, Liao J, Sim X, Liu J, Chia K-S, et al. Evaluation of transethnic fine mapping with population-specific and cosmopolitan imputation reference panels in diverse Asian populations. Eur J Hum Genet. 2016;24: 592–599. doi:10.1038/ejhg.2015.150
25. Li YR, Keating BJ. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. 2014;6: 91. doi:10.1186/s13073-014-0091-5
26. Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, Colangelo L, et al. Genetic ancestry in lung-function predictions. N Engl J Med. 2010;363: 321–330. doi:10.1056/NEJMoa0907897
27. Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nat Genet. 2011;43: 237–241. doi:10.1038/ng.763
28. Acuña-Alonzo V, Flores-Dorantes T, Kruit JK, Villarreal-Molina T, Arellano-Campos O, Hünemeier T, et al. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. Hum Mol Genet. 2010;19: 2877–2885. doi:10.1093/hmg/ddq173
29. Adeyemo A, Rotimi C. Genetic variants associated with complex human diseases show wide variation across multiple populations. Public Health Genomics. 2010;13: 72–79. doi:10.1159/000218711
30. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016;375: 655–665. doi:10.1056/NEJMsa1507092
31. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. 2016;17: 157. doi:10.1186/s13059-016-1016-y
32. Oh SS, White MJ, Gignoux CR, Burchard EG. Making Precision Medicine Socially

745 Precise. Take a Deep Breath. Am J Respir Crit Care Med. 2016;193: 348–350.
746 doi:10.1164/rccm.201510-2045ED
747 33. Oh SS, Galanter J, Thakur N, Pino-Yanes M, Barcelo NE, White MJ, et al. Diversity in
748 Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. PLoS Med. 2015;12.
749 doi:10.1371/journal.pmed.1001918
750 34. Belbin GM, Nieves-Colón MA, Kenny EE, Moreno-Estrada A, Gignoux CR. Genetic
751 diversity in populations across Latin America: implications for population and medical genetic
752 studies. Curr Opin Genet Dev. 2018;53: 98–104. doi:10.1016/j.gde.2018.07.006
753 35. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human
754 Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum
755 Genet. 2017;100: 635–649. doi:10.1016/j.ajhg.2017.03.004
756 36. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, et al. Multi-
757 Ethnic Study of Atherosclerosis: objectives and design. Am J Epidemiol. 2002;156: 871–881.
758 37. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, et al.
759 Methylomics of gene expression in human monocytes. Hum Mol Genet. 2013;22: 5065–5074.
760 doi:10.1093/hmg/ddt356
761 38. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, et al. Genetic
762 architecture of gene expression traits across diverse populations. PLOS Genet. 2018;14:
763 e1007586. doi:10.1371/journal.pgen.1007586
764 39. Mak ACY, White MJ, Eckalbar WL, Szpiech ZA, Oh SS, Pino-Yanes M, et al. Whole-
765 Genome Sequencing of Pharmacogenetic Drug Response in Racially Diverse Children with
766 Asthma. Am J Respir Crit Care Med. 2018;197: 1552–1564. doi:10.1164/rccm.201712-2529OC
767 40. Thakur N, Oh SS, Nguyen EA, Martin M, Roth LA, Galanter J, et al. Socioeconomic
768 status and childhood asthma in urban minority youths. The GALA II and SAGE II studies. Am J
769 Respir Crit Care Med. 2013;188: 1202–1209. doi:10.1164/rccm.201306-1016OC
770 41. Borrell LN, Nguyen EA, Roth LA, Oh SS, Tcheurekdjian H, Sen S, et al. Childhood
771 Obesity and Asthma Control in the GALA II and SAGE II Studies. Am J Respir Crit Care Med.
772 2013;187: 697–702. doi:10.1164/rccm.201211-2116OC
773 42. Nishimura KK, Galanter JM, Roth LA, Oh SS, Thakur N, Nguyen EA, et al. Early-life air
774 pollution and asthma risk in minority children. The GALA II and SAGE II studies. Am J Respir
775 Crit Care Med. 2013;188: 309–318. doi:10.1164/rccm.201302-0264OC
776 43. Lappalainen T, Sammeth M, Friedländer MR, ‘t Hoen PA, Monlong J, Rivas MA, et al.
777 Transcriptome and genome sequencing uncovers functional variation in humans. Nature.
778 2013;501: 506–511. doi:10.1038/nature12531
779 44. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An
780 integrated map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81.
781 doi:10.1038/nature15394
782 45. Mikhaylova AV, Thornton TA. Accuracy of Gene Expression Prediction From Genotype
783 Data With PrediXcan Varies Across and Within Continental Populations. Front Genet. 2019;10.
784 doi:10.3389/fgene.2019.00261
785 46. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population
786 genomics of human gene expression. Nat Genet. 2007;39: 1217–1224. doi:10.1038/ng2142
787 47. Viñuela A, Brown AA, Buil A, Tsai P-C, Davies MN, Bell JT, et al. Age-dependent
788 changes in mean and variance of gene expression across tissues in a twin cohort. Hum Mol
789 Genet. 2018;27: 732–741. doi:10.1093/hmg/ddx424
790 48. McCall MN, Illei PB, Halushka MK. Complex Sources of Variation in Tissue Expression

Data: Analysis of the GTEx Lung Transcriptome. Am J Hum Genet. 2016;99: 624–635. doi:10.1016/j.ajhg.2016.07.007

49. Zhu Y, Wang L, Yin Y, Yang E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. Sci Rep. 2017;7: 5435. doi:10.1038/s41598-017-05882-0

50. Ferreira PG, Muñoz-Aguirre M, Reverter F, Godinho CPS, Sousa A, Amadoz A, et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. Nat Commun. 2018;9: 490. doi:10.1038/s41467-017-02772-x

51. Martin AR, Karczewski KJ, Kerminen S, Kurki MI, Sarin A-P, Artomov M, et al. Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. Am J Hum Genet. 2018;102: 760–775. doi:10.1016/j.ajhg.2018.03.003

52. Yuan Y, Tian L, Lu D, Xu S. Analysis of Genome-Wide RNA-Sequencing Data Suggests Age of the CEPH/Utah (CEU) Lymphoblastoid Cell Lines Systematically Biases Gene Expression Profiles. Sci Rep. 2015;5: 7960. doi:10.1038/srep07960

53. Çalışkan M, Pritchard JK, Ober C, Gilad Y. The Effect of Freeze-Thaw Cycles on Gene Expression Levels in Lymphoblastoid Cell Lines. PLOS ONE. 2014;9: e107166. doi:10.1371/journal.pone.0107166

54. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467: 52–58. doi:10.1038/nature09298

55. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011;27: 2304–2305. doi:10.1093/bioinformatics/btr341

56. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, et al. The Great Migration and African-American Genomic Diversity. PLOS Genet. 2016;12: e1006059. doi:10.1371/journal.pgen.1006059

57. Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics. 2011;98: 422–430. doi:10.1016/j.ygeno.2011.08.007

58. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287. doi:10.1038/ng.3656

59. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48: 1443–1448. doi:10.1038/ng.3679

60. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Consortium Gte, et al. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLOS Genet. 2016;12: e1006423. doi:10.1371/journal.pgen.1006423

61. Gravel S. Population genetics models of local ancestry. Genetics. 2012;191: 607–619. doi:10.1534/genetics.112.139808

62. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009;5: e1000519. doi:10.1371/journal.pgen.1000519

63. Tange O. GNU Parallel 2018 [Internet]. Ole Tange; 2018. doi:10.5281/zenodo.1146014

64. Shih DJH. argparser: Command-Line Argument Parser [Internet]. 2016. Available: https://CRAN.R-project.org/package=argparser

65. Wickham H. assertthat: Easy Pre and Post Assertions [Internet]. 2019. Available: https://CRAN.R-project.org/package=assertthat

66. Dowle M, Srinivasan A, Gorecki J, Chirico M, Stetsenko P, Short T, et al. data.table: Extension of "data.frame" [Internet]. 2019. Available: https://CRAN.R-project.org/package=data.table

67. Calaway R, Corporation M, Weston S, Tenenbaum D. doParallel: Foreach Parallel Adaptor for the "parallel" Package [Internet]. 2018. Available: https://CRAN.R-project.org/package=doParallel

68. Dinno A. dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums [Internet]. 2017. Available: https://CRAN.R-project.org/package=dunn.test

69. Xie Y, Vogt A, Andrew A, Zvoleff A, http://www.andre-simon.de) AS (the C files under inst/themes/ were derived from the H package, Atkins A, et al. knitr: A General-Purpose Package for Dynamic Report Generation in R [Internet]. 2019. Available: https://CRAN.R-project.org/package=knitr

70. Davis TL. optparse: Command Line Option Parser [Internet]. 2019. Available: https://CRAN.R-project.org/package=optparse

71. Gentleman R. annotate: Annotation for microarrays [Internet]. 2018. Available: http://bioconductor.org/packages/annotate/

72. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma Oxf Engl. 2005;21: 3439–3440. doi:10.1093/bioinformatics/bti525

73. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4: 1184–1191. doi:10.1038/nprot.2009.97

74. Bolstad B. preprocessCore [Internet]. 2017. Available: https://github.com/bmbolstad/preprocessCore

75. Wickham, Hadley, Grolemund, Garrett. R for Data Science [Internet]. O'Reilly Media, Inc.; 2017. Available: https://r4ds.had.co.nz/

76. Wickham, Hadley. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available: http://ggplot2.org

77. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467: 1061–1073. doi:10.1038/nature09534

Figures and Tables

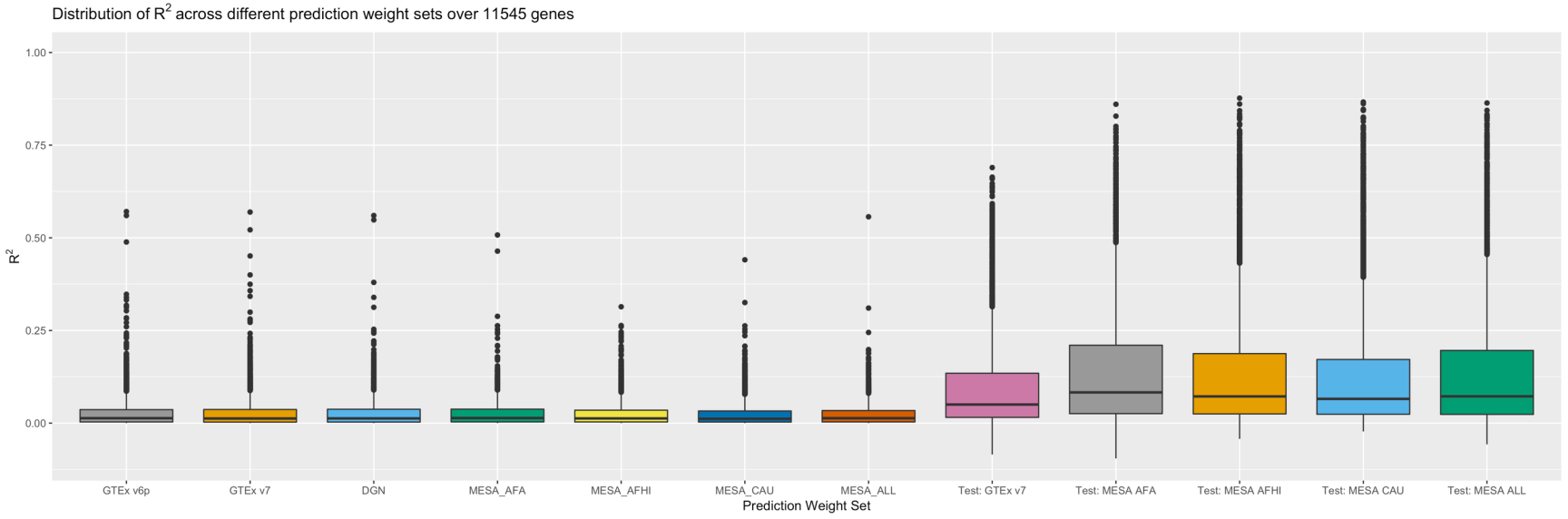


Figure 1: A comparison of R^2 between prediction and measurement in SAGE, with PredictDB test metrics as benchmarks, for 11,545 genes total. The prediction weights used here are, from left to right: GTEx v6p, GTEx v7, DGN, MESA African Americans, MESA African Americans and Hispanics, MESA Caucasians, and all MESA subjects. Test R^2 from model training in GTEx 7 and MESA ("test_R2_avg" in PredictDB) appear on the right and provide a performance baseline. The number of genes per weight set varies; see Supplementary Table 1.

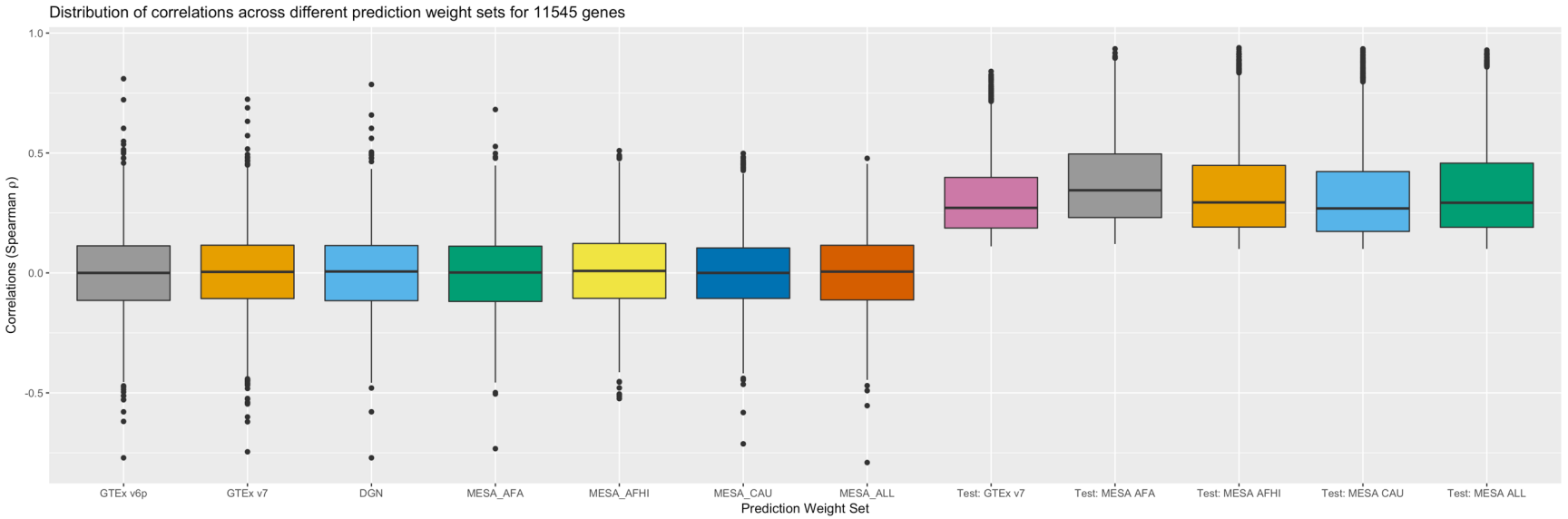
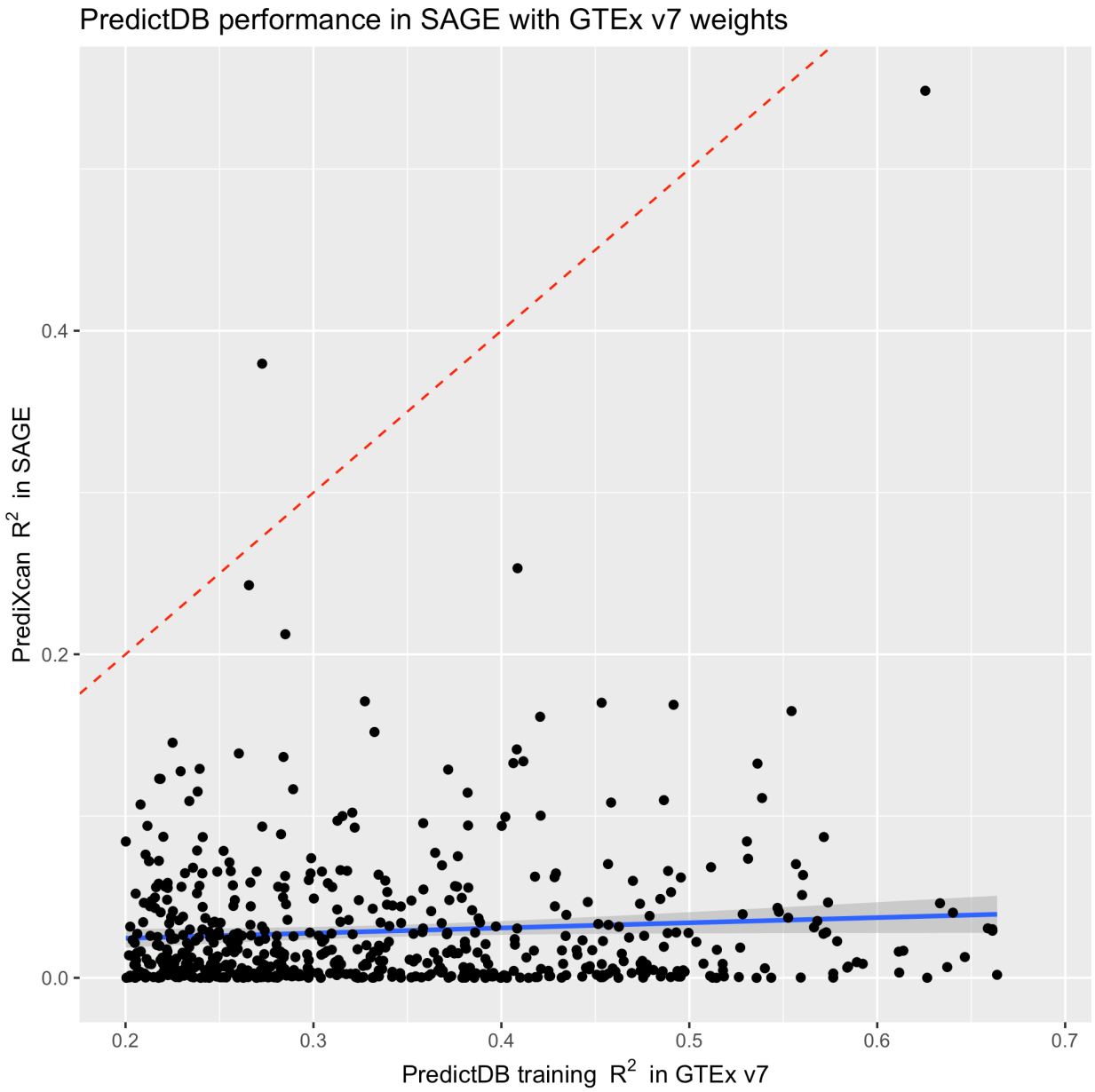


Figure 2: Spearman correlations of measured gene expression versus predicted expression from PrediXcan. The order of the weight sets matches

Figure 1. Test correlations for GTEx v7 and MESA correspond to "rho_avg" from PredictDB.

883



884

885 *Figure 3: A comparison of R^2 from SAGE and GTEx v7 training diagnostics. The SAGE R^2 are*
886 *computed from regressing PrediXcan predictions onto gene expression measurements. The GTEx*
887 *v7 R^2 are taken from PredictDB ("test_R2_avg"). The red dotted line marks where R^2 between*
888 *the two groups 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: Prediction 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: Prediction 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.

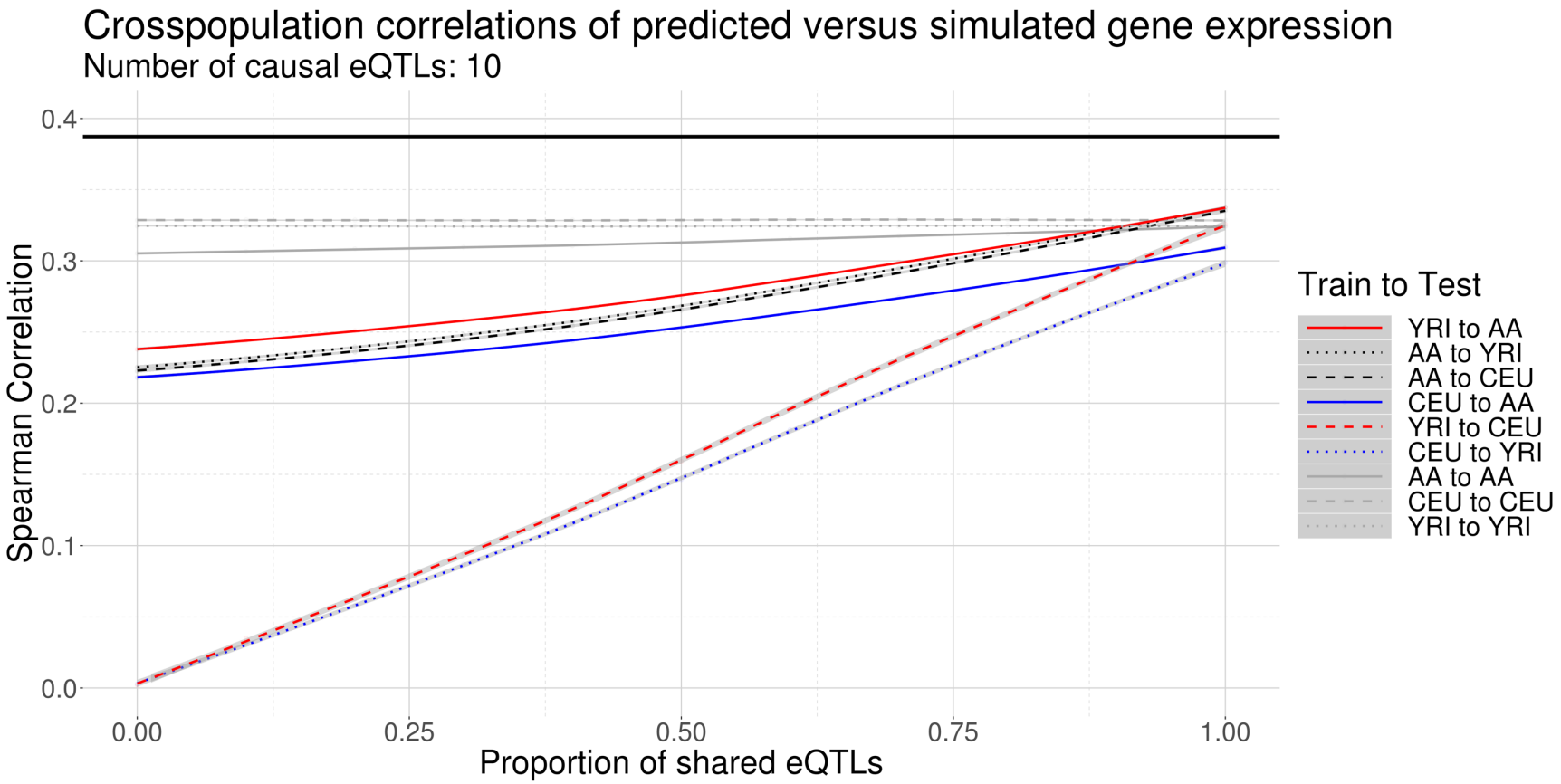


Figure 4: Correlations between predictions and simulated gene expression measurements from simulated populations across various proportions of shared eQTL architecture with 10 causal cis-eQTLs. Here YRI is simulated from the 1000 Genomes Yoruba, CEU is simulated from the Utahns, and AA is constructed from YRI and CEU. The black line represents the upper bound of correlation 0.387 dictated by our choice $h^2 = 0.15$ for the genetic heritability of expression. Each trend line represents an interpolation of correlation versus shared eQTL proportion. Gray areas denote 95% confidence regions of LOESS-smoothed mean correlations conditional on the proportion of shared eQTLs.

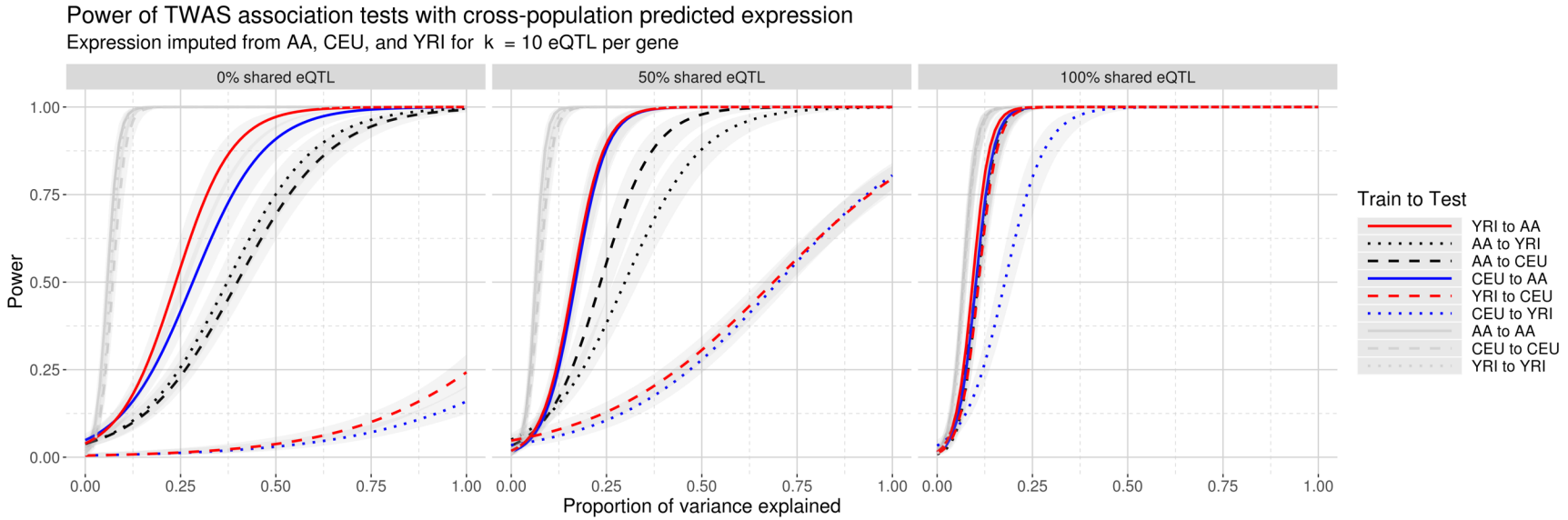


Figure 5: Curves depicting power to detect association under various TWAS scenarios. The x-axis represents the proportion of phenotypic variance explained by gene expression. As in **Figure 4**, *AA reflects simulated African-Americans constructed from YRI and CEU. The curves represent logistic interpolations of whether or not the causal gene was declared significant in an association test of a phenotype from the testing population with gene expression predicted from a training population into the testing population. Gray areas denote 95% confidence regions of mean power conditional on the effect size. A dotted red line at $h^2 = 0.95$ marks the power values shown in*

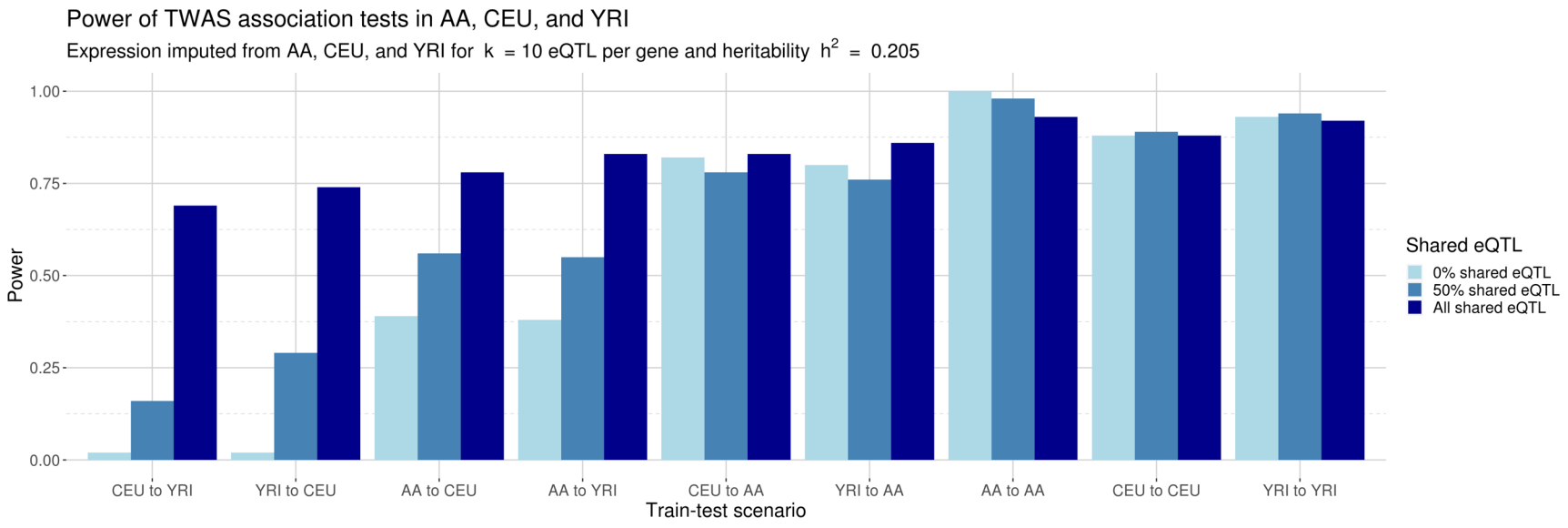


Figure 6: Power for phenotype-expression association tests with cross-population imputed gene expression for heritability $h^2 = 0.205$. The cross-population scenarios are ordered left to right from least shared ancestry (CEU to YRI, 0.0 shared ancestry) to most shared ancestry (YRI to AA, 0.8 shared ancestry). Power increases on two axes: (1) as the proportion of shared eQTL architecture increases, and, to a lesser extent, (2) as genetic distance decreases between reference and target populations. Power is consistently high when training and testing populations match.

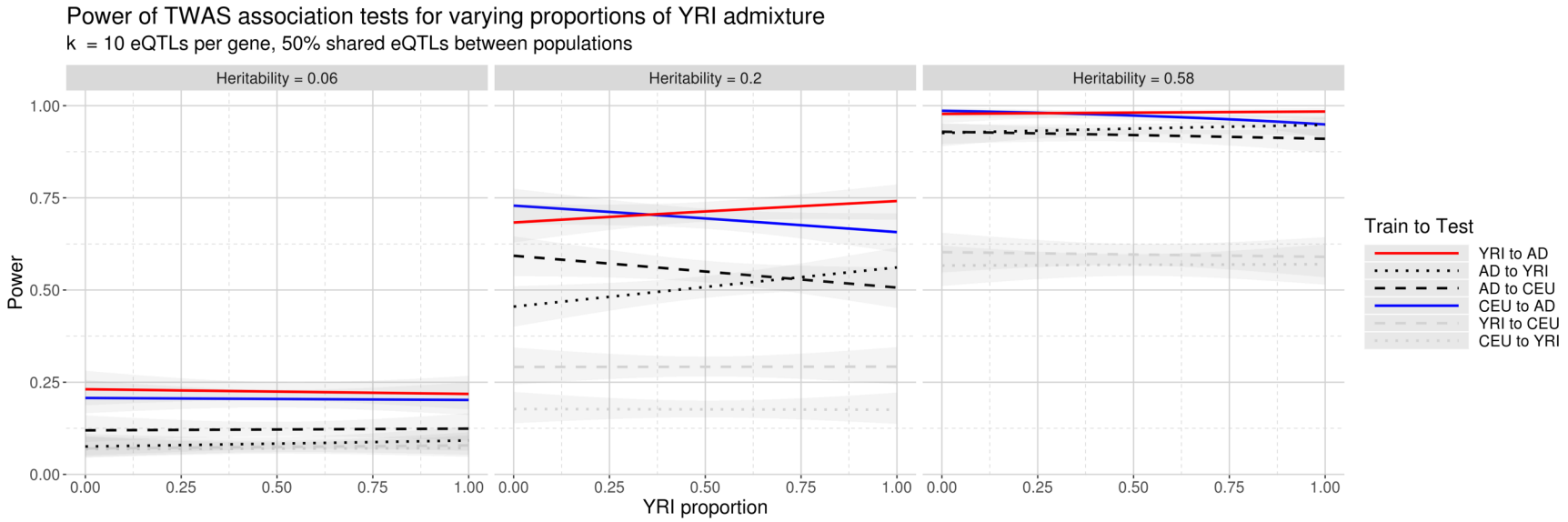


Figure 7: Power for various cross-population train-test scenarios with varying YRI admixture for three phenotypic heritability levels h^2 = 0.06, 0.20, and 0.58, corresponding to effect sizes 0.005, 0.01, and 0.025, respectively. Power increases as heritability increases, but also as populations become more genetically similar.

Supplementary Figures and Tables

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.

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

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

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

947

Training Pop	Testing Pop	Shared eQTL Proportion	Correlation (Mean)	Correlation (StdErr)
AA	AA	0	0.305	0.039
AA	AA	0.1	0.307	0.039
AA	AA	0.2	0.308	0.038
AA	AA	0.3	0.310	0.038
AA	AA	0.4	0.311	0.038
AA	AA	0.5	0.313	0.038
AA	AA	0.6	0.315	0.036
AA	AA	0.7	0.318	0.036
AA	AA	0.8	0.319	0.036
AA	AA	0.9	0.321	0.036
AA	AA	1	0.324	0.035
CEU	CEU	0	0.329	0.035
CEU	CEU	0.1	0.329	0.035
CEU	CEU	0.2	0.329	0.035
CEU	CEU	0.3	0.328	0.035
CEU	CEU	0.4	0.329	0.035
CEU	CEU	0.5	0.328	0.035
CEU	CEU	0.6	0.329	0.035
CEU	CEU	0.7	0.329	0.035
CEU	CEU	0.8	0.329	0.035
CEU	CEU	0.9	0.328	0.035
CEU	CEU	1	0.329	0.035
YRI	YRI	0	0.324	0.035
YRI	YRI	0.1	0.325	0.035
YRI	YRI	0.2	0.325	0.035
YRI	YRI	0.3	0.324	0.035
YRI	YRI	0.4	0.324	0.035
YRI	YRI	0.5	0.324	0.035
YRI	YRI	0.6	0.325	0.035
YRI	YRI	0.7	0.324	0.035
YRI	YRI	0.8	0.325	0.035
YRI	YRI	0.9	0.324	0.035
YRI	YRI	1	0.324	0.035

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

948

Correlation Mean (Std Err)		Train-test direction		
		AA	CEU	YRI
Training Pop	AA	0.324 (0.0352)	0.309 (0.0399)	0.337 (0.0306)
	CEU	0.335 (0.0335)	0.328 (0.0348)	0.325 (0.0389)
	YRI	0.337 (0.0302)	0.298 (0.0459)	0.324 (0.0347)

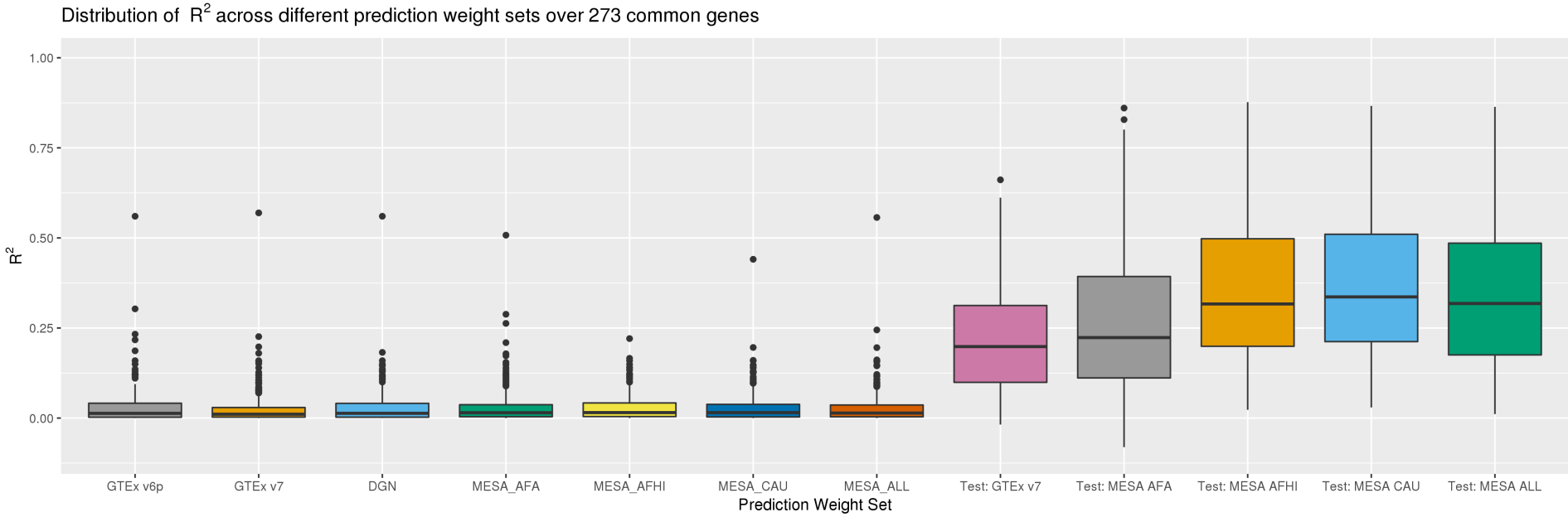
Supplementary Table 5: Prediction performance under fully shared eQTL architecture for k = 10 eQTLs yields reliable cross-population gene expression prediction. Results for other sizes of eQTL models are similar.

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}

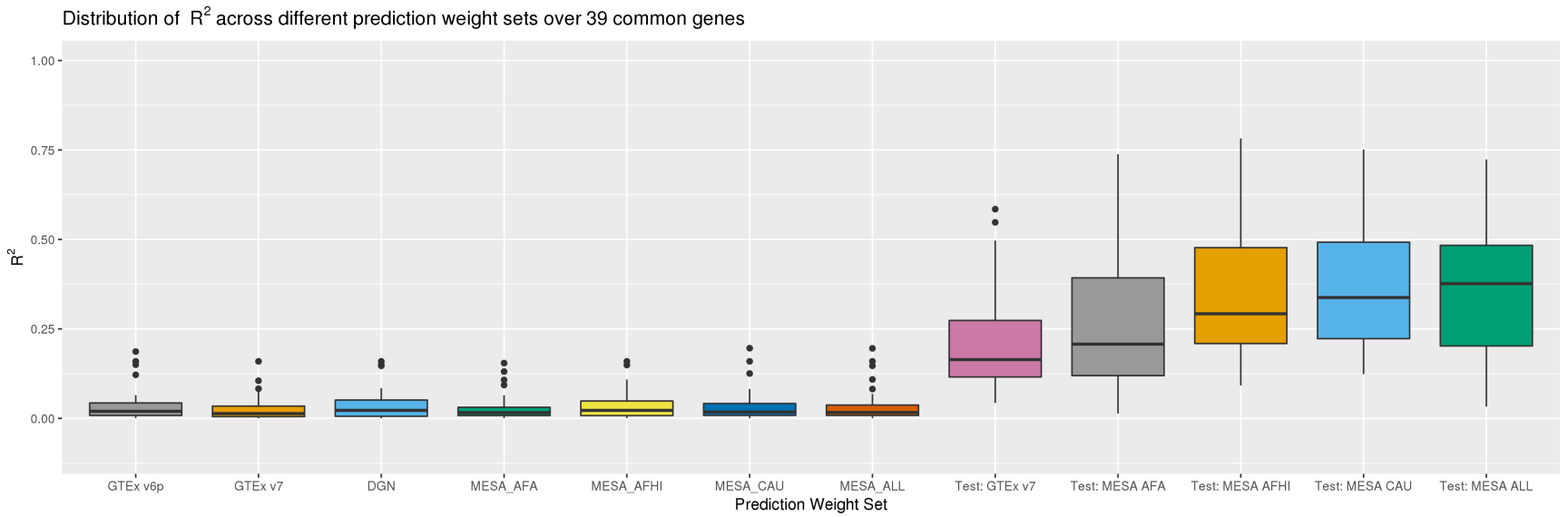
Supplementary Table 6: A Dunn test shows statistically significant differences when predicting between AFR and EUR populations versus predicting between EUR populations.

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**	-1.401 (p < 1.00E+00)				
	CEU to AA	6.712 (p < 1.44E-10)	8.113 (p < 3.68E-15)			
	CEU to YRI	28.517 (p < 5.32E-178)	29.919 (p < 8.34E-196)	21.805 (p < 1.55E-104)		
	YRI to AA	-5.391 (p < 5.23E-07)	-3.990 (p < 4.95E-04)	-12.104 (p < 7.51E-33)	-33.909 (p < 3.62E-251)	
	YRI to CEU	24.146 (p < 0.00E+00)	25.547 (p < 0.00E+00)	17.433 (p < 0.00E+00)	-4.371 (p < 0.00E+00)	29.538 (p < 0.00E+00)

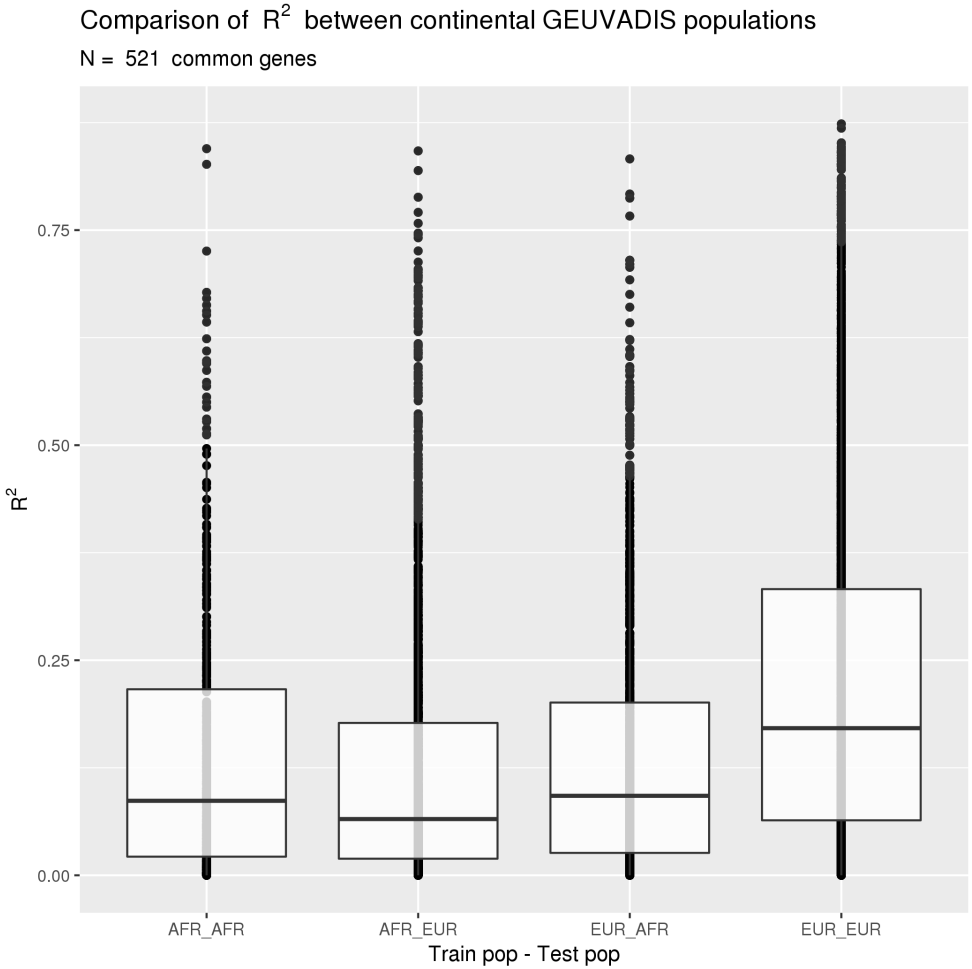
Supplementary Table 7: Differences in cross-population prediction performance are statistically significant, with a few notable exceptions. Prediction from AA to CEU or YRI is essentially the same, but all other scenarios are different, indicating that the direction of prediction does matter.



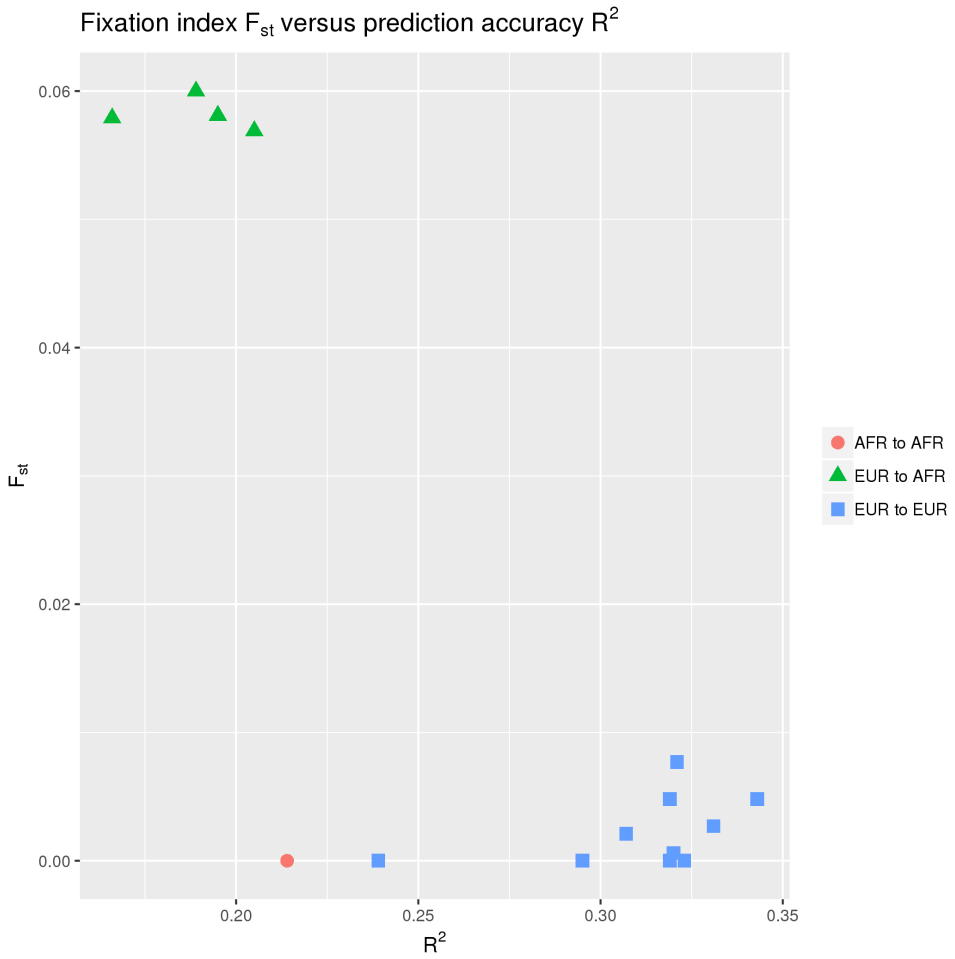
Supplementary Figure 1: R^2 of measured gene expression versus predictions from PrediXcan. The prediction weights used here are, from left to right: GTEx v6p, GTEx v7, DGN, MESA African Americans, MESA African Americans and Hispanics, MESA Caucasians, and all MESA subjects. Test R^2 from model training in GTEx 7 and MESA ("test_R2_avg" in PredictDB) appear on the right and provide a performance baseline.



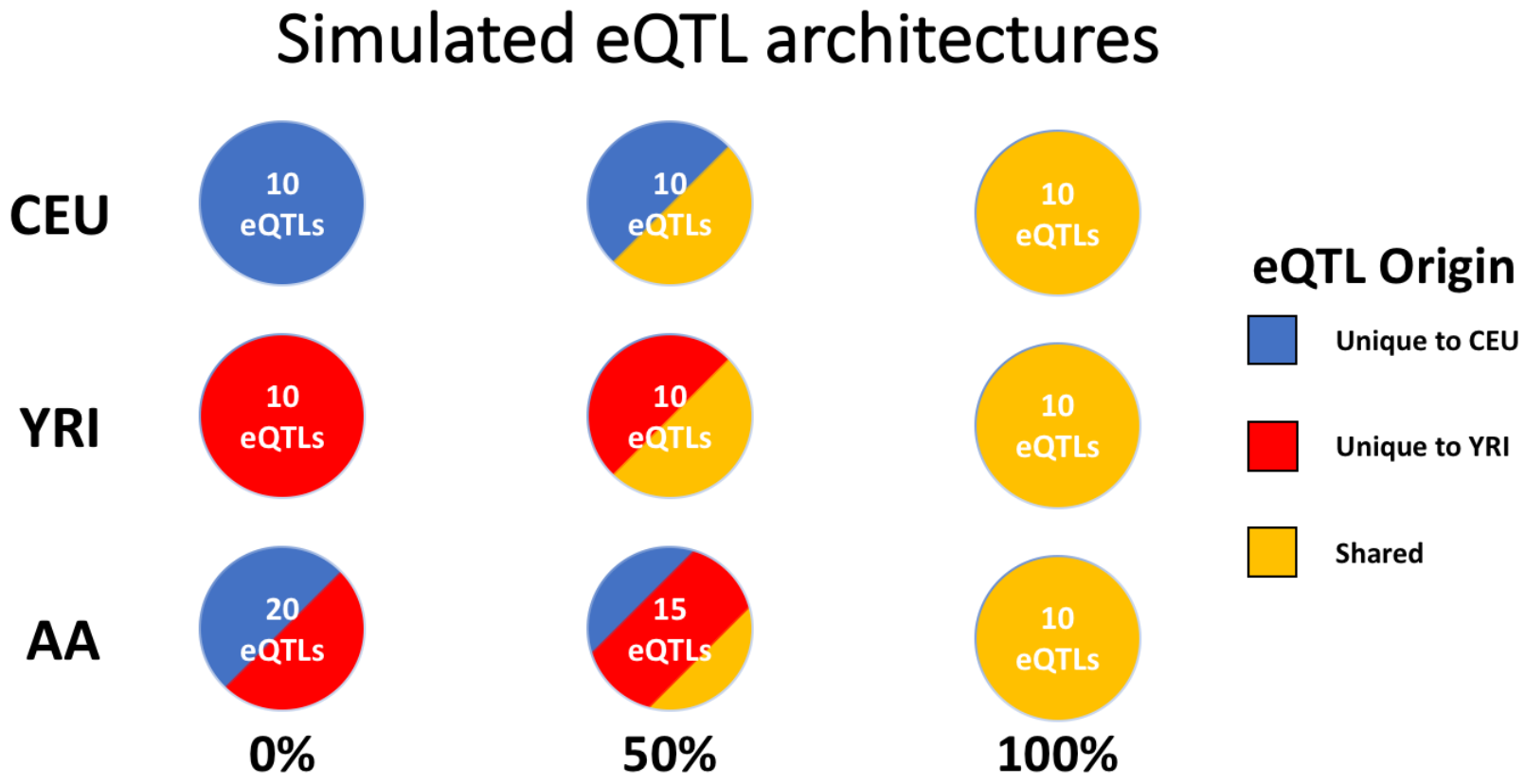
Supplementary Figure 2: R^2 between prediction and measurement in SAGE only using the 39 genes with positive correlation between prediction and measurement in all weight sets and benchmarks.



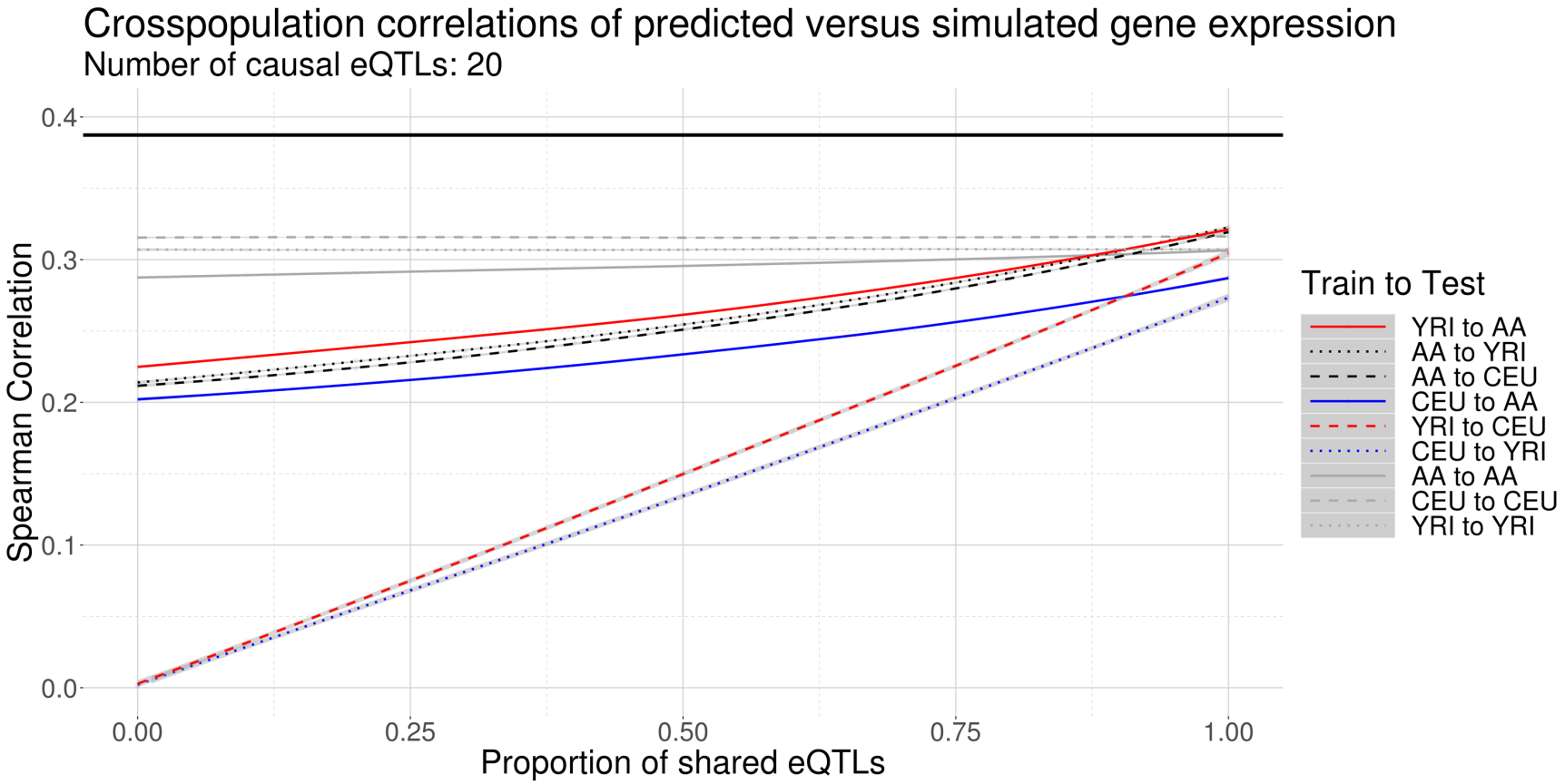
Supplementary Figure 3: Prediction R^2 between AFR (YRI) and EUR (CEU, TSI, GBR, and FIN). Predicting into and from AFR produces consistently lower R^2 than predicting within EUR, suggesting a potential decrease in prediction accuracy when predicting across continental population groups.



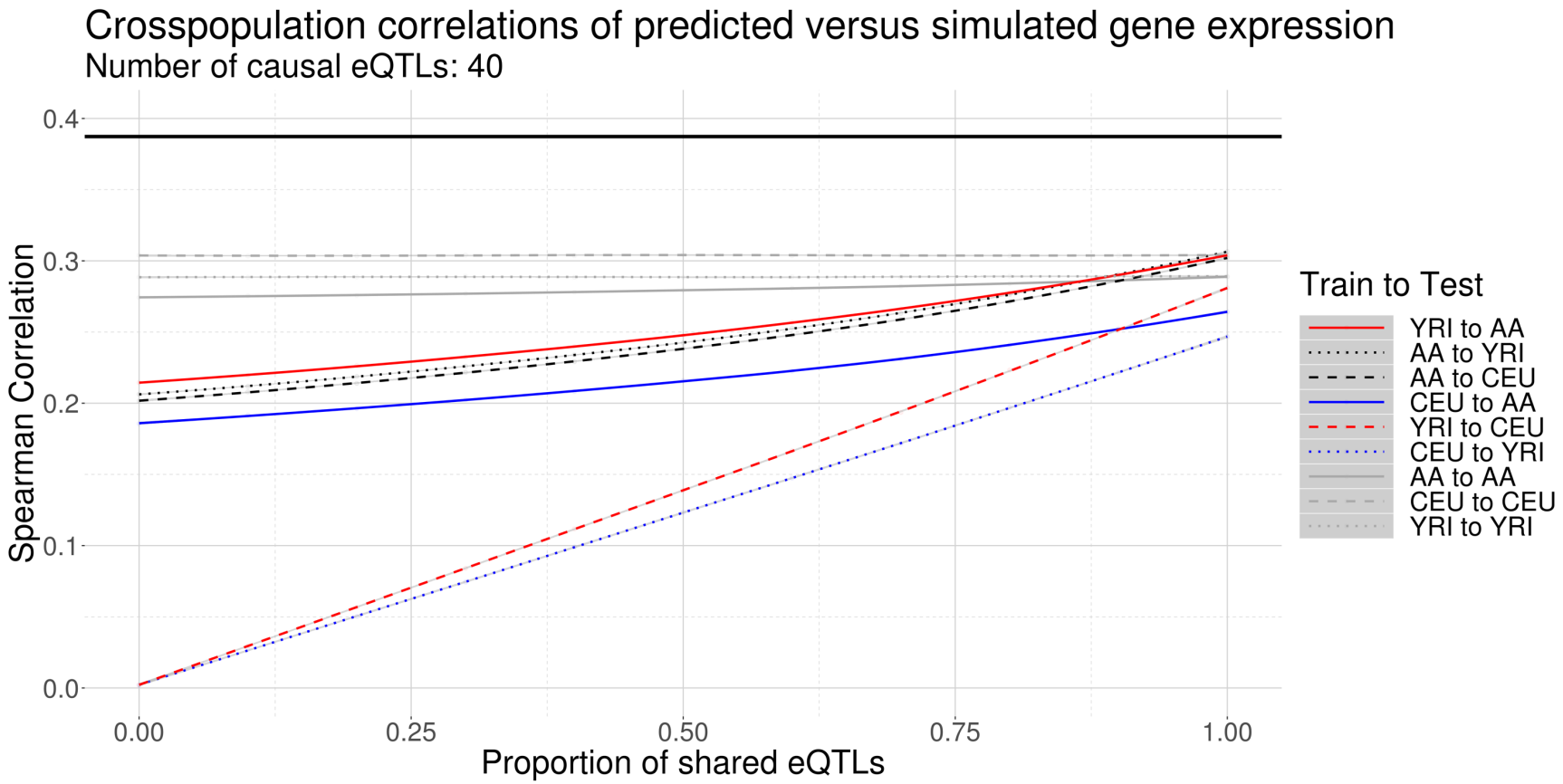
Supplementary Figure 4: Genetic distance versus prediction accuracy over 142 genes with positive correlation across all train-test scenarios. Here the GEUVADIS populations are arranged into three groups. AFR to AFR includes prediction from YRI into itself; EUR to AFR includes prediction into YRI from CEU, GBR, TSI, and FIN; and EUR to EUR includes prediction within and between all European populations in GEUVADIS. Clustering by genetic distance separates prediction between European populations from prediction between European populations and AFR. F_{ST} are taken from the 1000 Genomes Project (Table S11).[77]



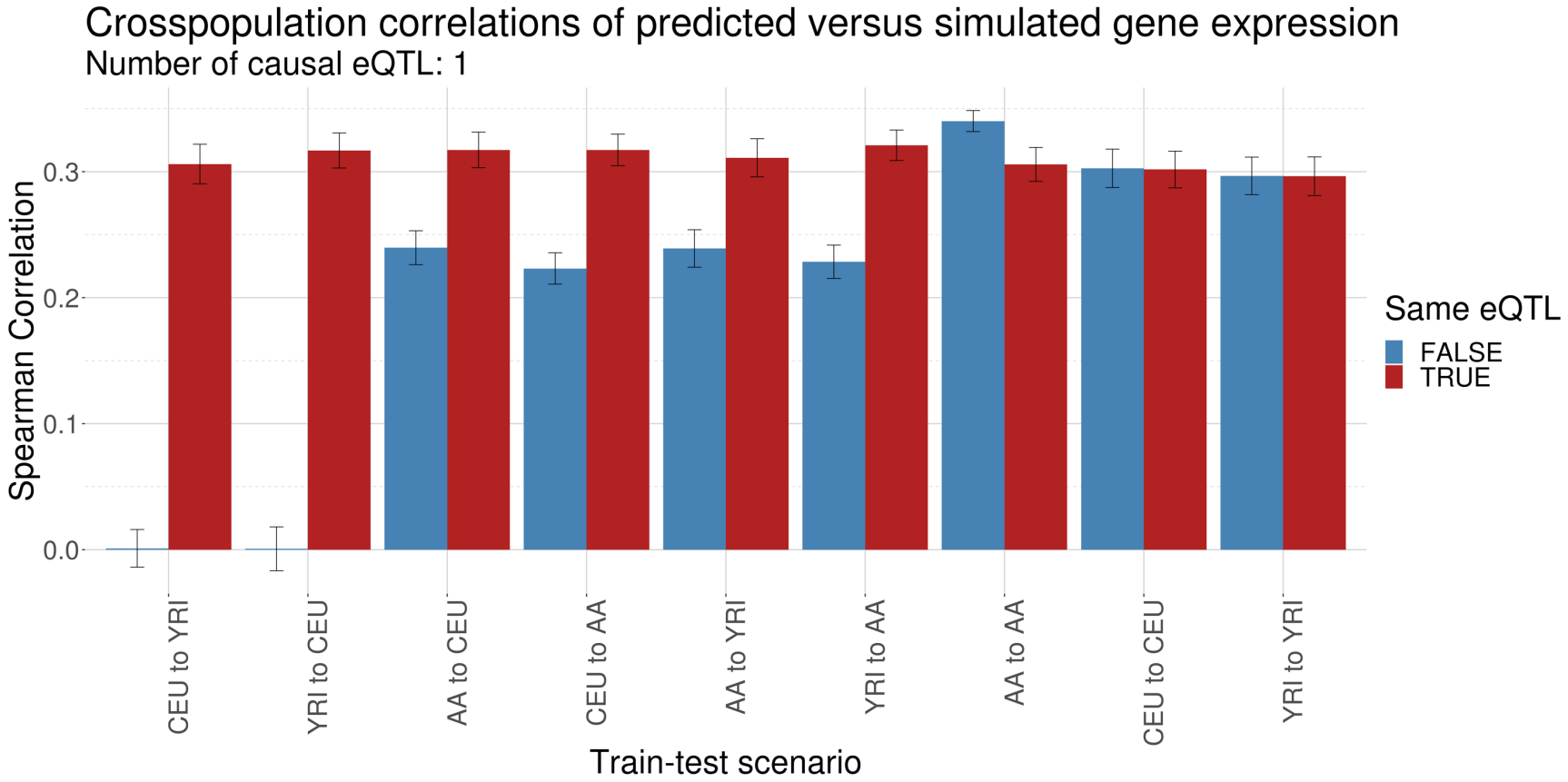
Supplementary Figure 5: A schematic of three shared eQTL architectures for the case of k = 10 eQTLs per gene. Blue encodes eQTLs specific to CEU; red encodes eQTLs specific to YRI; and gold encodes eQTLs shared between CEU and YRI. Models for CEU and YRI always had k eQTLs. AA always inherited all eQTLs from the ancestral populations. Consequently, the number of eQTLs in AA varied depending on how many eQTLs CEU and YRI shared.



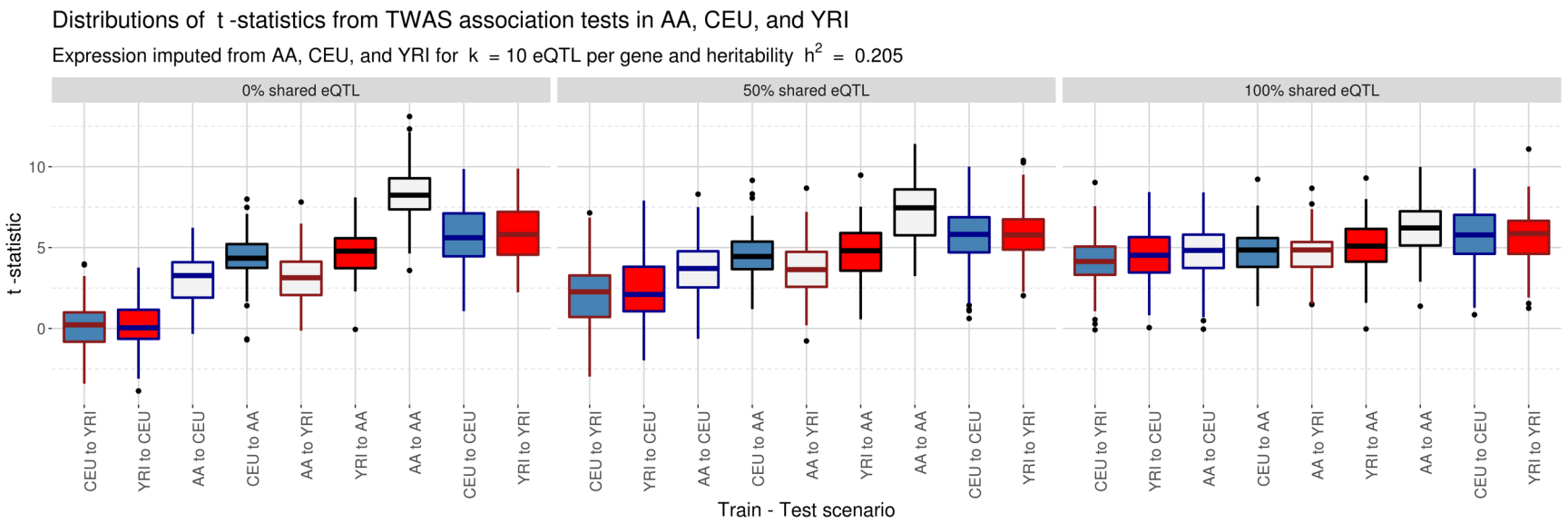
Supplementary Figure 6: Correlations between predictions and simulated gene expression measurements from simulated populations across various proportions of shared eQTL architecture with 20 causal cis-eQTLs.



Supplementary Figure 7: Correlations between predictions and simulated gene expression measurements from simulated populations across various proportions of shared eQTL architecture with 40 causal cis-eQTLs.



Supplementary Figure 8: Mean correlations between predictions and simulated gene expression measurements from simulated populations for a single causal cis-eQTL. For this simplified eQTL architecture, the ancestral populations (CEU and YRI) either share the causal eQTL (TRUE) or not (FALSE). In the TRUE case, AA has 1 eQTL shared with CEU and YRI; in the FALSE case, it has 2 unique eQTLs, one from each of CEU and YRI. Error bars denote 95% confidence intervals.



Supplementary Figure 9: Distributions of t-statistics across various shared eQTL proportions for all nine train-test scenarios with 1000 Genomes populations for a fixed TWAS effect size and fixed number of causal eQTLs. The labels are ordered from left to right from least shared ancestry (CEU to YRI, shared ancestry proportion 0) to most shared ancestry (YRI to AA, shared ancestry proportion 0.8), with train-test scenarios from a population into itself on the right of each panel. Increasing proportions of shared eQTLs yield stronger association statistics from cross-population predictions. Fully shared eQTL architectures yield consistently high power across populations. Median t-statistics increase as populations share more haplotypes, while association tests with gene expression predicted in the same population show consistently high power.