bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Accuracy of gene expression prediction from genotype data with PrediXcan varies across diverse populations

Anna Mikhaylova [1,*] and Timothy Thornton [1,*]

[1]Department of Biostatistics, University of Washington, Seattle , WA, USA.
*Correspondence to avmikh@uw.edu and tathornt@uw.edu.

**Abstract**

Predicting gene expression with genetic data has garnered significant attention in recent years. PrediXcan is one of the most widely used gene-based association methods for testing imputed gene expression values with a phenotype due to the invaluable insight the method has shown into the relationship between complex traits and the component of gene expression that can be attributed to genetic variation. The prediction models for PrediXcan, however, were obtained using supervised machine learning methods and training data from the Depression and Gene Network (DGN) and the Genotype-Tissue Expression (GTEx) data, where the majority of subjects are of European descent. Many genetic studies, however, include samples from multi-ethnic populations, and in this paper we assess the accuracy of gene expression predictions with PrediXcan in diverse populations. Using transcriptomic data from the GEUVADIS (Genetic European Variation in Health and Disease) RNA sequencing project and whole genome sequencing data from the 1000 Genomes project, we evaluate and compare the predictive performance of PrediXcan in an African population (Yoruban) and four European populations. Prediction results are obtained using a range of models from PrediXcan weight databases, and Pearson's correlation coefficient is used to measure prediction accuracy. We demonstrate that the predictive performance of PrediXcan varies across populations (F-test p-value $< 0.001$), where prediction accuracy is the worst in the Yoruban sample compared to European samples. Moreover, the performance of PrediXcan varies not only among distant populations, but also among closely related populations as well. We also find that the qualitative performance of PrediXcan for the populations considered is consistent across all weight databases used.

## 1 Introduction

In the past decade, genome-wide association studies (GWAS) have identified thousands of genetic variants significantly associated with a wide range of human phenotypes. The vast majority of these studies, however, were conducted in samples from European ancestry populations [1–5]. Differences in allele frequencies, genetic architecture, and linkage disequilibrium (LD) patterns across ancestries suggest that GWAS discoveries can fail to generalize across populations, and recent publications have provided compelling evidence that GWAS findings often do not transfer from European

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

populations to other ethnic groups. For example, Carlson et al. analyzed multi-ethnic data from the PAGE Consortium and concluded that some GWAS-identified variants from European ancestry population had different magnitude and direction of allelic effects in non-European populations and the differential effects were more persistent in African Americans [6]. Moreover, genetic risk prediction models derived from European GWAS were unreliable when applied to other ethnic groups [6]. Martin et al. examined the impact of population history on polygenic risk scores and demonstrated that they were biased and confounded by population structure. [7]. Since genetic risk prediction accuracy depends on genetic similarity between the target and discovery cohorts, Martin et al. advised against interpreting the scores across populations and recommended computing them in genetically similar cohorts.

Associations between genetic variation and molecular traits, such as gene expression, have advanced our understanding of the mechanisms underlying trait-variant associations [8]. Prior studies have shown that a large proportion of GWAS variants identified for complex traits are expression quantitative trait loci (eQTLs): i.e., they play a role in regulating gene expression [9]. Thus, eQTLs can aid in prioritizing likely causal variants among the ones identified by GWAS, especially if they are found in non-coding regions, and uncover the mechanisms by which genotypes influence phenotypes [8]. So having three types of data – genotype, phenotype and gene expression – on the same set of subjects can be advantageous for investigating the relationships between phenotypes and genetic background of a subject and underlying processes. However, collecting all of these data types is often not feasible due to cost and tissue availability. Additionally, eQTL studies have the same pitfalls as GWASs – the majority of the detected eQTLs are not causal, but may be in LD with causal variants. Similar to variants identified through GWAS, eQTL findings might fail to replicate in diverse populations due to LD patterns that differ across populations.

Recently methods, such as PrediXcan, have been proposed for integrating eQTL studies and GWASs [10]. Such methods have multiple advantages over traditional GWAS methods, especially where expression data from the tissue of interest are not available and in cases when gene expression is in the causal pathway between genotypic variants and phenotype. PrediXcan can lead to an increase in power to detect associations for multiple reasons. First, it removes environmental noise and focuses on the genetically regulated component of gene expression. Second, PrediXcan bases gene expression prediction on a limited number of variants that are 1Mb upstream and downstream from the gene and then tests for association between the predicted expression and a phenotype. So, by including fewer variants that are potentially causal for every gene, the method has better power to detect eQTLs. Lastly, by conducting tests on aggregated variants instead of testing every variant, PrediXcan dramatically reduces multiple testing burden.

However, PrediXcan models were built using data from the Depression Genes and Networks (DGN) and the Genotype-Tissue Expression (GTEx) Project – both of which consist primarily of European-ancestry subjects. This poses the question of how accurate PrediXcan expression predictions are for non-European ancestry populations. Previous research has reported differences in gene expression levels across diverse populations from the HapMap3 project noting that 77%

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of eQTLs are population specific and only 23% are shared between two or more populations [11]. More distantly related populations have more differentially expressed genes, although this can often be explained by the expression of different gene transcripts across populations [12].

In this work, we investigated whether the predictive performance of PrediXcan differs across four European populations and one African populations using the Genetic European Variation in Health and Disease (GEUVADIS) [12] and 1000 Genomes Projects data [12, 13]. We predicted gene expression levels using seven PrediXcan weight databases derived from whole blood and lymphoblastoid cell lines (LCL) expression data. To test prediction accuracy across populations, we compared observed and predicted gene expression levels by calculating Pearson's correlation coefficients and then using linear mixed models to assess significant differences. In addition, we also evaluated the utility of whole-blood-based models when making predictions for LCL expression data. The results suggests that accuracy of PrediXcan for gene expression prediction differs across populations, even among closely related European ancestry populations. Furthermore, PrediXcan prediction accuracy is the worst in Africans across all weight databases we considered.

# 2 Materials and Methods

## 2.1 Datasets

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

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

## 2.2 Filtering out poorly predicted genes

Linear regression models were used to identify genes whose predicted values are not associated with the observed values at significance level of 0.05 in order to filter out the genes with poor prediction accuracy across all subjects. For each gene, we fit a linear regression model with observed gene expression as the outcome, and predicted gene expression as the predictor of interest. We performed

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the Wald test to assess the significance of the coefficient for each gene and excluded the genes whose corresponding p-values were above the significance level of 0.05.

We then calculated Pearson's correlation coefficient, $r$, between observed and predicted expression values for every gene, in each population separately. A few genes had constant predicted gene expression levels across all subjects. Since we could not calculate the correlation if one of the variables was constant, we excluded those genes. Thus, for every gene we had five Pearson's correlation coefficients, one per population. Note that we used $r$ instead of the square of Pearson correlation, $r^2$, in order to take the directionality of correlation into account. Using $r^2$ as a measure of predictive accuracy can be misleading because a large proportion of genes predicted and observed expression values that are negatively correlated.

## 2.3 Prediction accuracy differences across populations and across tissues

To assess how the training of prediction models with different populations affects prediction accuracy, we used a linear mixed effect model approach. After filtering out poorly predicted genes, we fit the following model:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{FIN,i} + \beta_2 \mathbb{I}_{GBR,i} + \beta_3 \mathbb{I}_{TSI,i} + \beta_4 \mathbb{I}_{YRI,i} + \epsilon_{ij}, \tag{1}$$

where $r_{ij}$ is the correlation coefficient for gene $i$ in population $j$; and $\mathbb{I}_{FIN,i}$, $\mathbb{I}_{GBR,i}$, $\mathbb{I}_{TSI,i}$, and $\mathbb{I}_{YRI,i}$ are indicator variables that are equal to 1 if the gene correlation was calculated on the population indicated in the subscript, and otherwise are equal to 0. Thus, we modeled population as a categorical predictor, with the CEU population as a reference. To account for variation between genes, we included a random intercept $\gamma_i$ for each gene and we assumed that $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$. We also included an error term $\epsilon_{ij}$, such that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. To simultaneously test for differences in correlation coefficients across populations, we used repeated measures ANOVA. To assess the association between the change in correlation coefficient and population, we tested the coefficients for each population using the likelihood-ratio test.

We also ran an additional analysis where we excluded the CEU population due to potentially lower quality of the CEU cell lines, as reported in the literature [14, 15]. We fit a model identical to (1), excluding the CEU and using the FIN population as a reference:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{GBR,i} + \beta_2 \mathbb{I}_{TSI,i} + \beta_3 \mathbb{I}_{YRI,i} + \epsilon_{ij}, \tag{2}$$

where the notation is the same as above. Again, we performed a repeated measures ANOVA to test for differences in gene correlations across the populations and the likelihood-ratio test to separately test the change in gene correlations for each population compared to the reference population.

To evaluate how PrediXcan performance with whole-blood (WB) databases differed from LCL databases, we restricted the set of genes to only those that were present in both the WB and LCL databases. We compared each pair of GTEx WB and GTEx LCL databases using a paired t-test. All the statistical analyses described above were performed in R version 3.3.3.

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 3 Results

## 3.1 Overview of PrediXcan weight databases

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

To avoid repetition, we focus our attention on DGN, GTEx v7 WB and GTEx v7 LCL databases in the main text, and report our findings for the other four databases in the Supplementary material.

## 3.2 PrediXcan prediction accuracy differs across diverse populations

Using DGN, GTEx WB and GTEx LCL models and sequence data, we predicted gene expression for 10387, 5432 and 2777 genes, respectively (see Table 2). The number of genes with available predictions varied by population: the four European populations had similar counts and YRI had a slightly lower count. Because there was no variation in predicted expression values in at least one of the populations, we excluded 33 genes from DGN, 13 from GTEx WB, and 10 from GTEx LCL. From the remaining genes, we filtered out the ones with poor prediction accuracy based on associations between observed and predicted values, as described in the Materials and Methods section. Two-thirds of genes were excluded by this criteria from the genes predicted with DGN database, and slightly less than a half were excluded from gene sets predicted with the GTEx databases.

Next, we computed gene correlation coefficients, separately in each of the five populations. Violin plots display the correlation coefficients by population across genes before and after filtering (see Figures 1A and 1B, respectively). We note that prediction accuracy is slightly lower for the African populations than for any of the European populations, regardless of the weight database used, and this trend is even more obvious after the filtering process.

Afterwards, we binned the genes into six categories based on the gene correlation coefficients (see Table 3). The majority of genes have very poor prediction accuracy – of the genes predicted with whole-blood databases, a third have negative correlations and a half have correlations between 0 and 0.2. Of the genes predicted with LCL, a fifth have negative correlations and over a third have correlations between 0 and 0.2. The distribution of gene correlation coefficients is fairly similar across the four European populations, although predictive accuracy seems worse in CEU compared to FIN, GBR, and TSI. The predictive accuracy is the worst in the African sample. Across all populations, only a small number of genes were predicted with high accuracy (with $r > 0.6$). Furthermore, all

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

European populations have a greater number of well-predicted genes than the African population, regardless of the weight database used.

Next, we assessed the association between the prediction accuracy (as gene correlation coefficients) and population category via repeated measures ANOVA and linear mixed models. We present the parameter estimates and their 95% confidence intervals calculated using model-based standard errors for the model 1 in Table 4. Based on the repeated measures ANOVA, we find that prediction accuracy differs across populations, regardless of the weight database used (p-values for all databases were $< 0.001$). From the linear mixed model 1, we find that the prediction accuracy is significantly higher in FIN, GBR and TSI and significantly lower in YRI, compared to CEU (all p-values $< 0.001$). This suggests that predictive performance varies not only among distant populations, but also among closely related populations.

Finally, we repeated the analysis described above, this time excluding the CEU population. We present the parameter estimates and the corresponding 95% confidence intervals in Table 5. From the repeated measures ANOVA, we find that prediction accuracy differs across the four populations (p-values for all databases were $< 0.001$). Moreover, based on the coefficients and the corresponding p-values from the linear mixed model 2, we estimate the prediction accuracy to be significantly higher in GBR and significantly lower in TSI and YRI, compared to the FIN population (see corresponding p-values in Table 5). This difference in prediction accuracy is the greatest between YRI and FIN when GTEx v7 LCL weight database was used. Like in the analysis above, we notice that predictive performance differs across populations, including European populations.

### 3.3 PrediXcan prediction accuracy differs between tissues

As can be seen in the violin plots in Figure 1, both databases based on whole blood perform similarly, and LCL-based database displays improved prediction accuracy. In order to compare pairwise gene correlations, we restricted our analyses to the 1,587 genes common in both GTEx v7 WB and GTEx v7 LCL. Scatter plots presented in Figure 2 suggest that the majority of genes have very similar correlation coefficients when using WB and LCL databases across all populations. However, we see more genes in the upper left corner, above the dotted line, indicating that using the LCL database results in more genes have better prediction accuracy. This result is not surprising since the expression data we used were derived from LCL. The results of the paired t-test are consistent with the visual examination of the data: the mean difference between gene correlations based on the GTEx v7 LCL model and based on the GTEx v7 WB model is 0.03 ($p$-value $< 0.0001$), with predictions based on the LCL model having better performance.

## 4 Discussion

In this work, we evaluated PrediXcan performance and compared it across five geographically diverse populations using multiple weight databases. Models from all seven weight databases were trained mostly on subjects of European ancestry; three of the databases were derived from LCL and the

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

remaining four from whole blood. As a measure of prediction accuracy, we computed correlation coefficients for each gene in all populations and used the linear mixed models framework to quantify the differences in prediction performance across populations. We also investigated whether whole blood models could be used for predicting gene expression levels in LCL.

Overall, PrediXcan accurately predicted gene expression for some genes; however, the majority of genes had very poor correlation between measured and predicted expression levels. For almost half the genes, the correlation was negative. As expected, prediction accuracy was higher when the training and testing cohorts were of similar ancestry; i.e., models trained on Europeans performed better in the subjects of European descent and the worst in the African subjects. Surprisingly, prediction accuracy varied even among the European populations, with Finnish, British, and Italian populations having significantly higher accuracy than the CEU. These results held under all the weight databases we considered. Lastly, LCL-trained models outperformed whole-blood-trained models, although the prediction accuracy was similar for many of the genes.

A recent study reported consistent results to our findings and suggested that gene expression models should be trained on genetically similar populations [16]. Lack of genomic data from diverse populations limits the ability to effectively interpret and translate genomic results into clinical applications for individuals from admixed and other non-European populations. Our results in this paper emphasize the need to develop methods that account for ancestry and incorporate ancestral LD structure and allele frequencies differences. We also corroborate the importance of including more ancestrally diverse individuals in medical genomics to ensure that everyone gets the benefits of precision medicine and to avoid further exacerbating healthcare inequality.

We conclude this paper with some important caveats. LCLs are derived from B cells found in whole blood, and they provide a continuous supply of genetic material for GWAS and gene expression studies. However, they do undergo a transformation to become immortal that can change their biology and they do not have the same properties as native tissue [17]. Storage conditions, freeze-thaw cycles, and maturity of cell lines can also affect gene expression patterns [14, 15]. The CEU cell lines were collected much earlier than the other cell lines and LCL age can have a confounding effect and bias downstream analyses [14]. This factor could have contributed to the differences in prediction accuracy among European populations. Lastly, our study had modest sample sizes and only one non-European population. Future work is needed to investigate the performance and prediction accuracy of PrediXcan and other related approaches for gene expression in other multi-ethnic and ancestrally diverse populations.

# Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Author Contributions

AM and TT conceived the idea, designed the analysis, interpreted the results, and wrote the paper. AM ran the analysis.

# Data Availability Statement

GEUVADIS expression data is avalable at Array Express (E- MTAB-264 and E-GEUV-1) at `https://www.ebi.ac.uk/arrayexpress/experiments/` and 1000 Genomes project genotype data is available at `http://www.internationalgenome.org/`.

# Tables

Table 1: Summary of PrediXcan databases used in analyses.

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

Table 2: Number of genes for which Pearson correlation coefficients are available by population and by PrediXcan weight database.

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

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3: Binned gene correlation coefficients for the five populations using DGN, GTEx WB and GTEx LCL weight databases.

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

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

| | DGN | | | GTEx v7 WB | | | GTEx v7 LCL | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value |
| FIN | 0.019 | (0.014, 0.025) | < 0.001 | 0.021 | (0.015, 0.028) | < 0.001 | 0.038 | (0.030, 0.046) | < 0.001 |
| GBR | 0.029 | (0.023, 0.034) | < 0.001 | 0.032 | (0.025, 0.039) | < 0.001 | 0.051 | (0.043, 0.059) | < 0.001 |
| TSI | 0.010 | (0.004, 0.016) | < 0.001 | 0.013 | (0.007, 0.020) | < 0.001 | 0.027 | (0.019, 0.035) | < 0.001 |
| YRI | -0.054 | (-0.059, -0.048) | < 0.001 | -0.070 | (-0.077, -0.063) | < 0.001 | -0.097 | (-0.105 -0.089) | < 0.001 |

Table 5: Results from linear mixed models for population category (excluding CEU, with FIN as a reference) and change in gene correlation coefficient among filtered genes.

| | DGN | | | GTEx v7 WB | | | GTEx v7 LCL | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value |
| GBR | 0.010 | (0.004, 0.015) | < 0.001 | 0.011 | (0.004, 0.018) | 0.003 | 0.013 | (0.005, 0.021) | 0.002 |
| TSI | -0.009 | (-0.015, -0.003) | 0.002 | -0.008 | (-0.015, -0.001) | 0.028 | -0.011 | (-0.019, -0.003) | 0.009 |
| YRI | -0.073 | (-0.079, -0.067) | < 0.001 | -0.091 | (-0.098, -0.084) | < 0.001 | -0.134 | (-0.143, -0.126) | < 0.001 |

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Figure captions

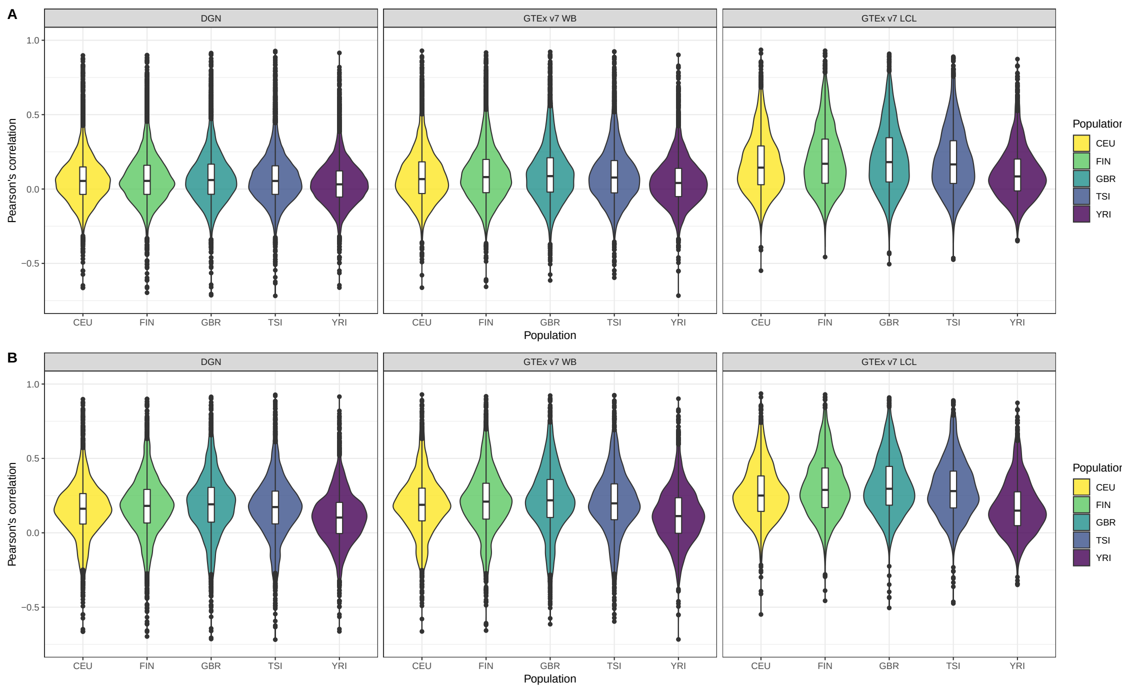


Figure 1: Violin plots of gene expression correlation coefficients by five populations using DGN, GTEx v7 WB and GTEx v7 LCL weight databases; **(A)** before and **(B)** after filtering out poorly predicted genes.

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

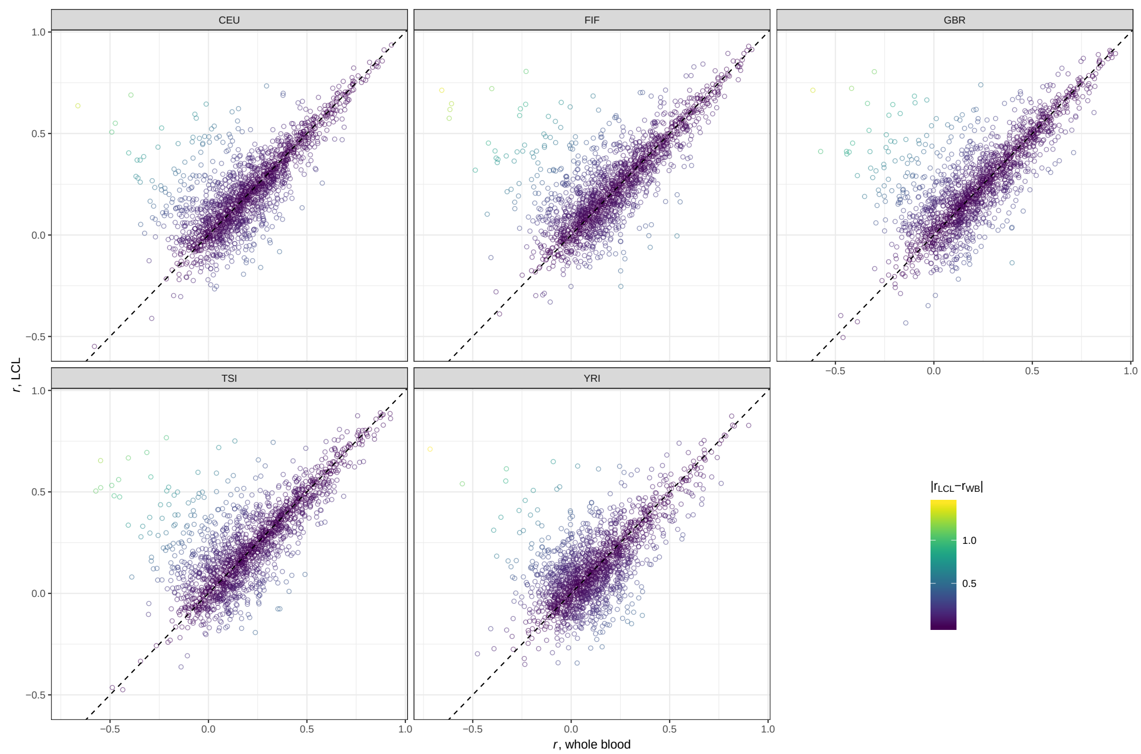


Figure 2: Scatter plots comparing gene correlation coefficients by population using GTEx v7 LCL vs GTEx v7 WB databases.

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# References

[1] A. C. Need and D. B. Goldstein, "Next generation disparities in human genomics: concerns and remedies," *Trends in Genetics*, vol. 25, no. 11, pp. 489–494, 2009.

[2] C. D. Bustamante, E. G. Burchard, and F. M. De la Vega, "Genomics for the world.," *Nature*, vol. 475, no. 14 July, pp. 163–165, 2011.

[3] S. Petrovski and D. B. Goldstein, "Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine," *Genome Biology*, vol. 17, no. 1, p. 157, 2016.

[4] A. B. Popejoy and S. M. Fullerton, "Genomics is failing on diversity," *Nature*, vol. 538, no. 7624, pp. 161–164, 2016.

[5] L. A. Hindorff, V. L. Bonham, L. C. Brody, M. E. Ginoza, C. M. Hutter, T. A. Manolio, and E. D. Green, "Prioritizing diversity in human genomics research," *Nature Reviews Genetics*, vol. 19, no. 3, pp. 175–185, 2018.

[6] C. S. Carlson, T. C. Matise, K. E. North, C. A. Haiman, M. D. Fesinmeyer, S. Buyske, F. R. Schumacher, U. Peters, N. Franceschini, M. D. Ritchie, D. J. Duggan, K. L. Spencer, L. Dumitrescu, C. B. Eaton, F. Thomas, A. Young, C. Carty, G. Heiss, L. Le Marchand, D. C. Crawford, L. A. Hindorff, and C. L. Kooperberg, "Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study," *PLoS Biology*, vol. 11, no. 9, 2013.

[7] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny, "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations," *American Journal of Human Genetics*, vol. 100, no. 4, pp. 635–649, 2017.

[8] F. W. Albert and L. Kruglyak, "The role of regulatory variation in complex traits and disease," *Nature Reviews Genetics*, vol. 16, no. 4, pp. 197–212, 2015.

[9] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. Eileen Dolan, and N. J. Cox, "Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS," *PLoS Genetics*, vol. 6, no. 4, 2010.

[10] E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, and H. K. Im, "A gene-based association method for mapping traits using reference transcriptome data," *Nature Genetics*, vol. 47, no. 9, pp. 1091–1098, 2015.

[11] B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley,

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

R. Durbin, P. Deloukas, and E. T. Dermitzakis, "Patterns of Cis regulatory variation in diverse human populations," *PLoS Genetics*, vol. 8, no. 4, 2012.

[12] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. 'T Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. Van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. Macarthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A. C. Syvänen, G. J. Van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis, "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, vol. 501, no. 7468, pp. 506–511, 2013.

[13] A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M. L. Yaspo, L. Fulton, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, C. L. Campbell, Y. Kong, A. Marcketta, F. Yu, L. Antunes, M. Bainbridge, A. Sabo, Z. Huang, L. J. Coin, L. Fang, Q. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, E. P. Garrison, D. Kural, W. P. Lee, W. F. Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, E. Banks, G. Bhatia, G. Del Angel, G. Genovese, H. Li, S. Kashin, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, T. Rausch, M. H. Fritz, A. M. Stütz, K. Beal, A. Datta, J. Herrero, G. R. Ritchie, D. Zerbino, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, B. Barnes,

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

M. Bauer, R. K. Cheetham, A. Cox, M. Eberle, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Herwig, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. F. Willems, J. T. Simpson, M. D. Shriver, J. A. Rosenfeld, C. D. Bustamante, S. B. Montgomery, F. M. De La Vega, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F. C. Hyland, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. J. Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, S. E. Devine, H. M. Kang, J. M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. K. Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, B. L. Browning, S. R. Browning, F. Hormozdiari, P. H. Sudmant, E. Khurana, C. Tyler-Smith, C. A. Albers, Q. Ayub, Y. Chen, V. Colonna, L. Jostins, K. Walter, Y. Xue, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, C. Hartl, K. Shakir, J. Degenhardt, S. Meiers, B. Raeder, F. P. Casale, O. Stegle, E. W. Lameijer, I. Hall, V. Bafna, J. Michaelson, E. J. Gardner, R. E. Mills, G. Dayama, K. Chen, X. Fan, Z. Chong, T. Chen, M. J. Chaisson, J. Huddleston, M. Malig, B. J. Nelson, N. F. Parrish, B. Blackburne, S. J. Lindsay, Z. Ning, Y. Zhang, H. Lam, C. Sisu, D. Challis, U. S. Evani, J. Lu, U. Nagaswamy, J. Yu, W. Li, L. Habegger, H. Yu, F. Cunningham, I. Dunham, K. Lage, J. B. Jespersen, H. Horn, D. Kim, R. Desalle, A. Narechania, M. A. Sayres, F. L. Mendez, G. D. Poznik, P. A. Underhill, D. Mittelman, R. Banerjee, M. Cerezo, T. W. Fitzgerald, S. Louzada, A. Massaia, F. Yang, D. Kalra, W. Hale, X. Dan, K. C. Barnes, C. Beiswanger, H. Cai, H. Cao, B. Henn, D. Jones, J. S. Kaye, A. Kent, A. Kerasidou, R. Mathias, P. N. Ossorio, M. Parker, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, Z. Tian, S. Tishkoff, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, X. Deng, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. S. Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. T. Hien, S. J. Dunstan, N. ThuyHang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, A. Duncanson, M. Dunn, and J. A. Schloss, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[14] Y. Yuan, L. Tian, D. Lu, and S. Xu, "Analysis of Genome-Wide RNA-Sequencing Data

bioRxiv preprint doi: https://doi.org/10.1101/524728; this version posted January 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Suggests Age of the CEPH/Utah (CEU) Lymphoblastoid Cell Lines Systematically Biases Gene Expression Profiles," *Scientific Reports*, vol. 5, pp. 1–5, 2015.

[15] M. Çalkan, J. K. Pritchard, C. Ober, and Y. Gilad, "The Effect of Freeze-Thaw Cycles on Gene Expression Levels in Lymphoblastoid Cell Lines," *PLoS ONE*, vol. 9, no. 9, p. e107166, 2014.

[16] L. S. Mogil, A. Andaleon, A. Badalamenti, S. P. Dickinson, X. Guo, J. I. Rotter, W. C. Johnson, H. K. Im, Y. Liu, and H. E. Wheeler, "Genetic architecture of gene expression traits across diverse populations Author summary," *PLoS Genetics*, pp. 1–17, 2018.

[17] D. E. Kelly, M. E. Hansen, and S. A. Tishkoff, "Global variation in gene expression and the value of diverse sampling," *Current Opinion in Systems Biology*, vol. 1, pp. 102–108, 2017.