Master's Paper of the Department of Statistics, the University of Chicago

Ancestry-eGenes: Differentially Expressed Genes by Local Ancestry and Genotype

Tae Hyun Kim

Advisor: Dan Nicolae

Approved _____

Date _____

June 14, 2017

Abstract

Ancestry is known to play a large role in phenotypic variation from physical appearances to susceptibility to diseases. In this work, we use the RNA-seq data in muscle-skeletal tissue to explore the relationship between gene expression level and ancestry. While there have been studies in the past that found a set of eQTLs and eGenes, this study aims to find the ancestry-eGenes whose local ancestry shows a strong interaction effect on their expression level with the genotype of a certain SNP. We used a permutation test to control for Type I error while accounting for the linkage disequilibrium structure. As a result, we found 201 ancestry-eGenes and analyzing their biological functions showed a significant enrichment of the genes in the MHC region, possibly implying a difference in the immunological evolutionary process between Africans and Europeans.

Acknowledgements

First, I would like to thank my advisor Professor Dan Nicolae for his guidance, encouragement, and patience throughout this project. He never hesitated to share his experience and knowledge whenever I needed them, and every comment he made helped me grow as a researcher. I also thank Professor Haky Im and her lab for providing the data and computational resources. I feel thankful to all the fellow students in the department. Their hard work and genuine enthusiasm in statistics inspire me to strive to become a better scientist. I am also grateful to my friends Jiunn Song and Hye Joon Min who offered not only their expertise in biology, but also their empathy for going through the similar stages in life. Lastly, I would like to thank Junyoung Park and my family for their unconditional emotional support.

Introduction

The advancement in the microarray sequencing technology and the subsequent genome-wide association studies (GWAS) have brought rich knowledge of human genetic variants and their association with many phenotypes. Some of the variants' specific roles in determining traits were clarified through further research in biology, but there are numerous functional common variants that are still left unexplained, especially in the non-coding region [14] [15] [16]. The molecular mechanism of how these variants affect human traits raised the question of the relationship between gene expression level and SNPs. The Genotype Tissue Expression (GTEx) project provides resources of gene expression level, genotypes, and phenotypes from the donors so that researchers can explore the particular question, and the GTEx pilot study has cataloged a comprehensive list of eQTLs, the SNPs that affect the gene expression level of a certain gene, and eGenes, the genes that have at least one eQTL [2].

Meanwhile, population genetics has long been an interest to many biologists and geneticists. Admixed population, whose recent ancestry includes populations from more than one continent, can be particularly useful in the field because it brings together the genomes from multiple populations that have diverged by selective pressures and genetic drift. The statistical tools that analyze these admixed genomes and identify the ancestry of chromosomal segments have been constantly improved in terms of both speed and accuracy, and inferred local ancestry has provided insights to disease mapping and to differential gene expression levels among different populations [10] [12] [17].

This paper builds on the two streams of research to discover the joint effects of the variants in a cis-region and the local ancestry on gene expression level. Our goal is to find 'ancestry-eGenes'. We define ancestry-eQTLs as the SNPs that influence the gene expression level differently based on the gene's local ancestry, and ancestry-eGenes as the genes that have at least one ancestry-eQTL in their cis-region. One illustration of this phenomenon would be that Africans and Europeans developed different transcription factors to a certain gene as a result of evolutionary divergence, and African transcription factor binds particularly well to a certain allele, while European transcription factor doesn't. Under this scenario, the genotypes and local ancestry will have an interaction effect because having that allele will increase the expression of the African gene but decrease (or not affect) the expression of the European gene.

We studied the African American and European American samples from GTEx due to the limited size of available admixed population data. Also, we only focused on the gene expression level of the muscle skeletal tissue, but the method is easily expandable to other tissues as well. As a result, we found 201 ancestry-eGenes and the analysis of their biological functions revealed that the genes in the MHC region show a significant enrichment. We also investigated how many of the ancestry-eGenes were caused by the different linkage disequilibrium structure between African and

European chromosomes. The comprehensive list of ancestry-eGenes and their biological functions will contribute to explaining on a molecular level the difference in the evolutionary process between the African and European populations.

Data [2]

Subjects

We studied African Americans and European Americans who have their gene expression level measured for the muscle skeletal tissue in GTEx. The sample includes 356 subjects, 51 of whom self-reported as African Americans (AA) and 305 as European Americans (EA). Through local ancestry inference and principal component analysis, we decided to classify four individuals from the self-reported European Americans as African Americans (See **Global and Local Ancestry** and Figure 1). We later provide evidence that jointly modeling European Americans and Africans is statistically valid and improves power.

Genotype Data and Imputation

The 356 subjects in GTEx were genotyped through Illumina OMNI 5M SNP array, and the rest of the SNPs were imputed using 1000 Genomes Project Phase I, version 3. Post-imputation genotype filters such as call rate threshold 95% and info score threshold 0.4 were applied. From the imputed genotype data, only the SNPs with minor allele frequency $> 5\%$ were included in the analysis. The number of SNPs per chromosome are summarized in Table 1.

Expression Level

We used the pre-processed gene expression level data from the GTEx portal. The data includes 22,248 genes in the autosomal chromosomes, and it is truncated for having at least 0.1 RPKM, normalized, log-transformed, and corrected for technical artifacts. The number of genes per chromosome in the available data is summarized in Table 2.

Number of SNPs			
chr1	502,003	chr12	335,827
chr2	576,281	chr13	264,149
chr3	488,450	chr14	225,404
chr4	514,658	chr15	201,010
chr5	447,505	chr16	210,634
chr6	471,029	chr17	188,733
chr7	408,784	chr18	202,046
chr8	382,608	chr19	164,372
chr9	298,030	chr20	153,560
chr10	350,319	chr21	99,372
chr11	343,657	chr22	95,734
		Total	6,954,165

Table 1

Number of genes			
chr1	2,387	chr12	1,190
chr2	1,544	chr13	393
chr3	1,315	chr14	753
chr4	841	chr15	810
chr5	1,088	chr16	1,048
chr6	1,174	chr17	1,386
chr7	1,141	chr18	328
chr8	818	chr19	1,512
chr9	946	chr20	550
chr10	910	chr21	267
chr11	1,268	chr22	579
		Total	22,248

Table 2

Global and Local Ancestry

Local ancestry was inferred through the software LAMP. The global ancestry, a value between 0 and 1 that quantifies the proportion of African chromosome in an individual, was computed by averaging the local ancestry, and it was cross-checked with the result of principal component analysis (Figure 1 (d)).

For reference allele frequency of pure populations, we used 186 Yoruban (YRI) subjects and 183 Northern Europeans from Utah (CEU) subjects from 1000 Genome Project (1000GP). Therefore, only the SNPs that were genotyped both in 1000GP and GTEx were included for the local ancestry inference. The phased data of 1000GP had missing values, and we excluded the SNPs that were missing more than 50% of the haplotype values. The number of SNPs used for local ancestry inference is summarized in S.Table 1.

Principal Component Analysis

Principal component analysis can effectively cluster the subjects into subpopulations [8] [11]. The covariance matrix of the genotypes of AAs and EAs from GTEx and YRIs and CEUs from 1000GP was eigen-decomposed to create the plot in Figure 1 (a) and (b), once with CHB (Han Chinese in Beijing, China) and once without. In Figure 1 (a), the first principal component distinguishes YRI from CEU population. As predicted, African Americans are placed between CEU and YRI. There are four outliers among self-reported GTEx European Americans. For the rest of the paper, we treat these 4 outliers as African Americans, making our pool of subjects a mix of 55 AAs and 301 EAs.

The second principal component reflects the sources of the data. The GTEx individuals were placed at the bottom while individuals with 1000GP were placed on the top. This means that principal component analysis is capable of detecting technical artifacts such as imputation procedure, missing values, and lab environment, and we can conclude that the genotype data from GTEx has no heterogeneity introduced by technical confounders.

Figure 1 (b) shows that when Asian population is added, PCA classifies different populations without the division from the genotyping procedure.

LAMP

For local ancestry inference, we used LAMP that reaches as high as 98% accuracy level for distinguishing YRI and CEU ancestry [7]. The software received as input the minor allele frequency of pure population CEU and YRI, the genotype data of the subjects from GTEx, and the chromosomal positions of the SNPs. For other parameters, we used 7 for the number of generations of admixture, 0.2 and 0.8 for initial proportion of CEU and YRI population, and 10^{-8} for recombina-

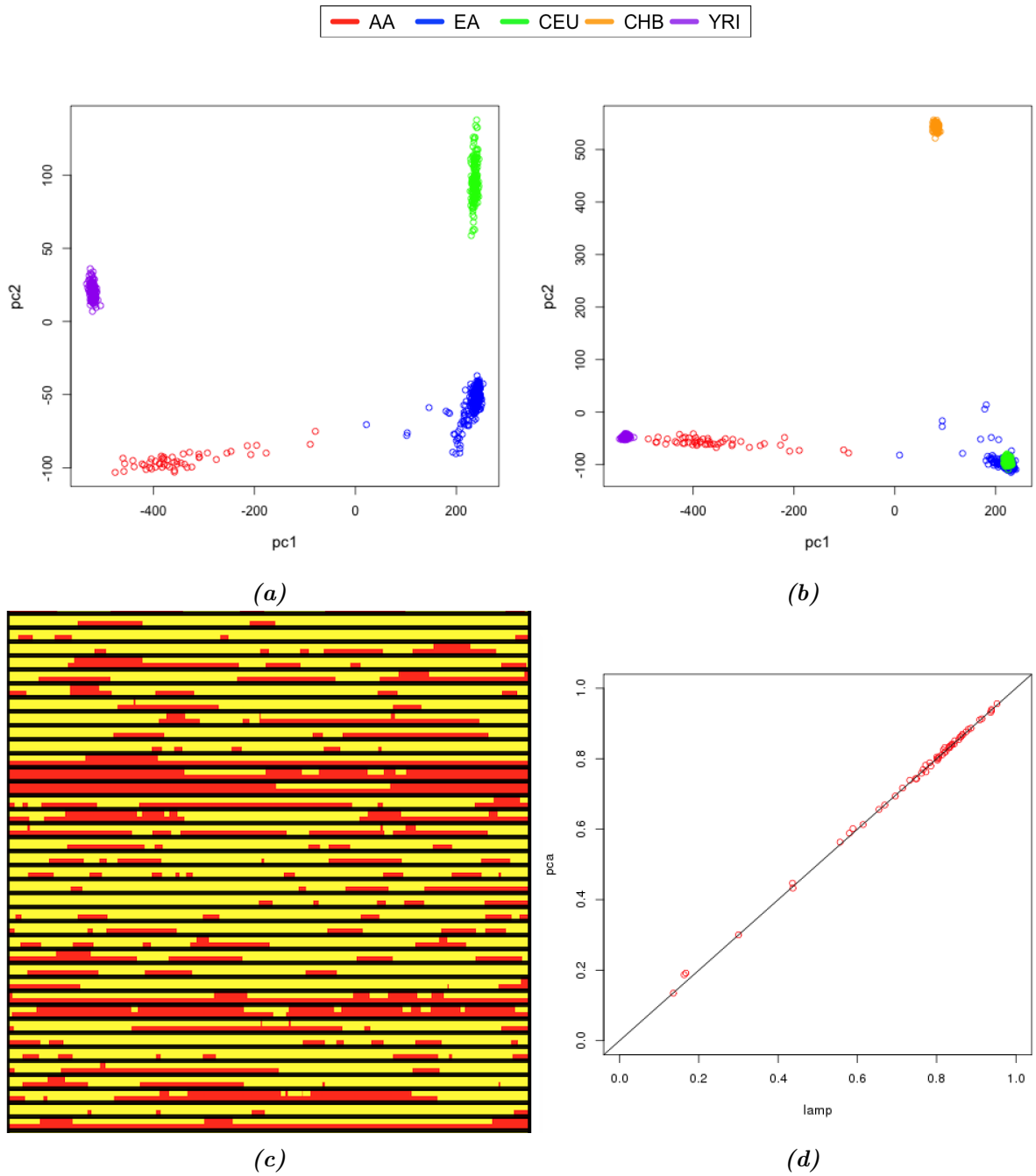


Figure 1: Local Ancestry Inference. *(a) PCA result of genotypes from common SNPs in 1000GP and GTEx. The first principal component divides the sample by ancestry, and the second principal component by the source of the data - 1000GP or GTEx. The four blue outliers with the lowest first principal components were treated as African Americans and were included for local ancestry inference. (b) PCA result with CHB population. The first two principal components only contain ancestry information. (c) Sample output from LAMP. Each row is chromosome 1 of one subject. Yellow blocks mean African and red blocks mean European ancestry. Dominantly red rows are from the outliers of self-reported European Americans. (d) Comparison of global ancestry inferred from LAMP and from PCA. LAMP's result is partly verified through PCA in terms of global ancestry.*

tion rate, and the results were robust to these initial parameters. LAMP returns 0, 1, or 2 African chromosomes for each locus. An example output of LAMP is visualized in Figure 1 (c). Each row indicates an individual's chromosome 1 (not all the individuals are shown) where yellow represents the African ancestry and red represents the European ancestry. The aforementioned outliers, who reported their ethnicity as 'white' but had a non-negligible proportion of African ancestry in their genome, are already included in this figure, and the three dominantly red rows with few yellow blocks are from those outliers.

Figure 1 (d) compares the global ancestry inferred by PCA and by LAMP. The global ancestry of individual i inferred from PCA was computed by

$$\text{global ancestry}_i = \frac{\text{pc1}_i - \overline{\text{pc1}_{CEU}}}{\overline{\text{pc1}_{YRI}} - \overline{\text{pc1}_{CEU}}}$$

where pc1_i indicates the value of the first principal component of subject i in Figure 1 (b), $\overline{\text{pc1}_{CEU}}$ means the average of the first principal component values of CEU population, and $\overline{\text{pc1}_{YRI}}$ means teh average of the first principal component values of YRI population. This plot shows that the returned local ancestry in LAMP is in part verified through PCA by matching global ancestry.

Framework

Most studies about local ancestry only take admixed populations as their sample. However, in this paper, we include European Americans in the analysis as well, and assign 0 for both their global and local ancestry. This aims to increase the power of our tests.

We are interested in the interaction between genotypes and local ancestries, and since most of the GTEx African American subjects have global ancestry around 0.8, their genes often have 1 or 2 for local ancestry, leaving many variants where no admixed individual has local ancestry 0. Therefore, increasing sample size, if possible, can improve power, and it would especially help to include Europeans who have local ancestry 0 everywhere.

On the other hand, we need certain assumptions such as constant variance in order to include both European Americans and African Americans in the same linear model. In this section, we explore the data and provide evidence of the homoscedasticity of the joint model, and we also include an adjustment to the model so that assigning 0 to the global ancestry of European Americans has a consistent interpretation with modeling only with African Americans.

Full Model

Before we present our justification for the joint modeling, we introduce our full model to make the notations clear. We will use the following model for our primary goal of investigating the association between the expression of gene k and its cis-SNP s where θ_{ks} is non-zero.

$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + G_s\lambda_{ks} + LG_{ks}\theta_{ks} + X\nu_{ks}$$

y_k is a $n \times 1$ vector of the gene expression level of $n = 356$ individuals for gene k, normalized to mean 0 and variance 1. μ_{ks} is a mean term, and $\mathbb{1}_{EA}$ is an indicator function assigning 1 to European Americans and 0 to African Americans. This allows different means in African Americans' and European Americans' gene expression. A is a $n \times 1$ vector for global ancestry, and $L_k \in \{0, 1, 2\}^{n \times 1}$ is a vector for the subjects' local ancestry of the gene k, defined as the local ancestry of the SNP that is closest to the middle of the gene. G_s is a $n \times 1$ vector for the genotype of SNP s. Here we limit s to be in a cis-region of the gene k, where 'cis' area is between (*gene start site - 1Mb*) and (*gene end site + 1Mb*). LG_{ks} is also a $n \times 1$ vector for the interaction term, which is an element-wise multiplication of G_s and L_k. α_{ks}, β_{ks}, $\gamma_{ks}, \lambda_{ks}, \theta_{ks}$ are all scalar coefficients. X is the covariate matrix of dimension 356×4 of age, sex and the first two principal components of the expression level matrix, and ν_{ks} is a coefficient vector of length 4 for the covariates X.

Constant Variance

In order to check that EAs and AAs can be included in the same linear model, we tested if the two groups' residuals have the same sample variance under the model $E(y_k) = \mu_k + A\beta_k + L_k\gamma_k$. We used the fact that the expression levels are already normalized: 22241 genes out of 22248 have Shapiro statistics > 0.99 and all genes have Shapiro statistics over 0.96. Denoting $r_{AA,k}$ as the residual vector in African Americans and $r_{EA,k}$ as the residual vector in European Americans from the model $E(y_k) = \mu_k + A\beta_k + L_k\gamma_k$, we tested the ratio of the sample variance of the $r_{AA,k}$ and $r_{EA,k}$ for all k from 1 to 22,248. Under the null hypothesis that the two populations are homoscedastic, the ratio of the residuals' sample variance should follow $F_{n_{AA}-1,n_{EA}-1}$ where n_{AA} is the sample size of African Americans and European Americans, i.e. 55 and 301 respectively. The 22,248 p-values from the F test for each gene are summarized in S.Figure 1. The p-value histogram is very close to uniform, and there was no particularly strong signal under FDR=0.05 indicating non-constant variance from any genes.

Same Mean Test

Next, we checked if an extra term of indicator $\mathbb{1}_{EA}$ is necessary. More specifically, we compared the following two models M1 and M2 for gene k.

$$\text{M1} : E(y_k) = \mu_k + A\beta_k + L_k\gamma_k$$

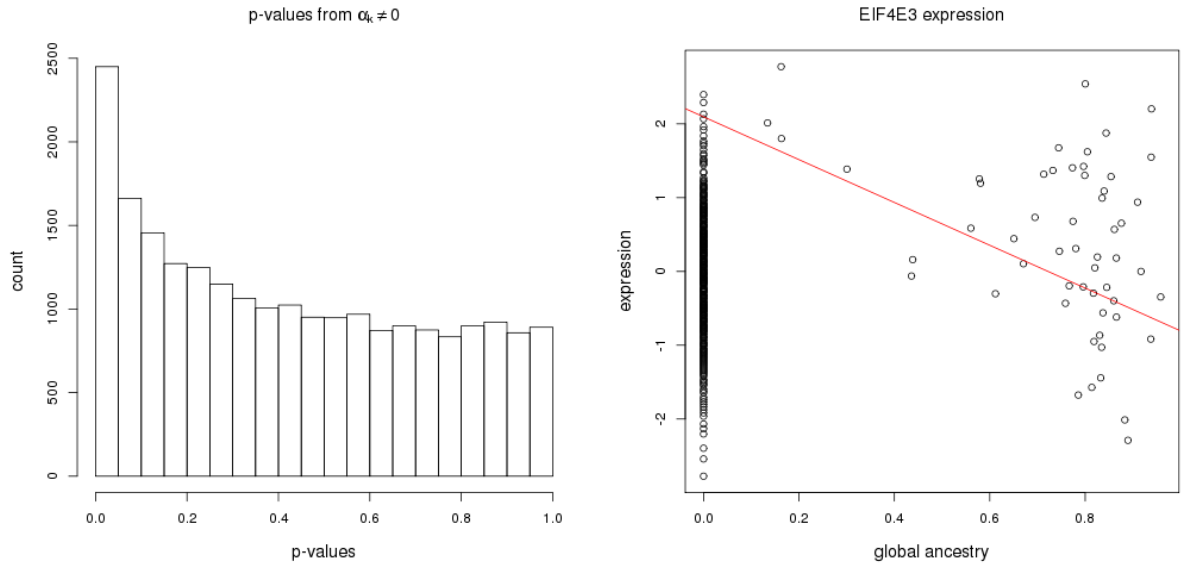
$$\text{M2} : E(y_k) = \mu_k + \mathbb{1}_{EA}\alpha_k + A\beta_k + L_k\gamma_k$$

This aims to see if the intercept of the linear model built only on African Americans is consistent with the mean expression level of European Americans whose global ancestry is set to 0. The resulting histogram of p-values testing $H_0 : \alpha_k = 0$ is shown in Figure 2.

The histogram shows clear signals of two different linear models and suggests that excluding the indicator term can lead to a wrong model. The most extreme scenario (lowest p-value) is observed in gene EIF4E3, as illustrated in Figure 2 (b). The red line is the fitted line from modeling only with African Americans. The intercept of the red line is clearly different from the mean gene expression level of European Americans. Jointly modeling the expression level and ancestry without the indicator term can lead to an imprecise estimate of the effect of ancestry on admixed population.

Intuitively, using the model M1 with only African Americans and using the model M2 with both European Americans and African Americans are equivalent under the constant variance assumption that we confirmed earlier. This shows that jointly modeling African Americans and European Americans maintains the intended interpretation; when genotype effects are ignored, European Americans can be considered homogeneous to admixed population with ancestry 0 with a separate

expression mean term.



(a) ***(b)***

Figure 2: Testing $\mathbb{1}_{\mathbf{EA}}$. *(a) Histogram of p-values from the t-test for the significance of the extra mean term* ($H_0 : \alpha_k = 0$). *(b) The gene with the strongest signal from* α_k. *The red line is the fitted line from modeling only with African Americans. Its intercept clearly doesn't match the mean of European Americans' gene expression.*

Effects of ancestry

Definition of local ancestry

Local ancestry in general is defined for each locus to be 2 if both chromosomes are African, 1 if heterozygous, and 0 if both European. In this work, we define local ancestry of a gene as the local ancestry of the SNP that is closest to the center of the gene. This is designed to estimate the average of local ancestry of the gene from the start site to the end site [10].

In most cases, the genes had the same ancestry information overall, without any recombination event within the gene. However, there are also genes in which we have no genotype information between the start site and end site. In this case, our best approximation of the local ancestry of the gene would be, indeed, the local ancestry of the SNP that is the most closely located to the gene, or equivalently, to the center of the gene.

On the other hand, there are genes in which 1 or more subjects show ancestry switch. The number of subjects with the ancestry switches are cataloged in Table 3; 20,358 genes had no recombination event for all the subjects, and 1,329 genes had one individual who had a recombination event within the gene. Around 92% of the genes show no recombination events in all the subjects, and only less than 3% of the genes have more than one individual with ancestry switch.

# of subjects	0	1	2	3	4	5	6	7	8	9	13
# of genes	20358	1329	379	120	34	12	9	3	2	1	1

Table 3: Genes with ancestry switch.

Therefore, we conclude that our definition of the local ancestry of a gene is a justifiable approximation of the average local ancestry of the gene.

Marginal effects of global and local ancestry

Before investigating the genotypes, we first studied the following model

$$E(y_k) = \mu_k + \mathbb{1}_{EA}\alpha_k + A\beta_k + L_k\gamma_k + X\nu_k$$

to examine how global and local ancestry marginally affect the gene expression level. We tested the hypotheses $\beta_k = 0$ and $\gamma_k = 0$, and the results are summarized in S.Figure 2 and S.Table 2. Most genes with strong signals in either β_k or γ_k show low p-values for both β_k and γ_k. This is in part expected because L_k and A are highly correlated due to the inclusion of European Americans who have 0 for both A and L_k for all genes k, and such collinearity has a negative impact on power. Nevertheless, we included both variables in our final model because first, global ancestry controls for general population structure, a procedure known to prevent inflated type I error, and second,

local ancestry is included in the interaction term, the variable of our main interest.

Covariates X

The covariates matrix X includes gender, age, and the first two principal components of the expression level matrix. The gender effect and age effect are introduced in S.Figure 3. The p-value histogram testing for their coefficients showed many genes are affected by age and gender.

While testing for the marginal global and local ancestry effects, we found that including the first two principal components of the expression matrix in the model particularly affects the p-values β_k, the effect size of global ancestry. The effects of the two PCs are illustrated in Figure 3 (a) and (b) through QQ-plot of p-values before and after controlling the PCs. The Figure 3 (a) was created under the model where X only includes gender and age, while Figure 3 (b) was created under the model where X includes the first two PCs of expression level matrix as well as gender and age. The inflated p-values in (a) were corrected in (b).

Through further investigation, we concluded that the p-values for global ancestry decrease, not because the two PCs explain away the ancestry effect, but because they correct for the unexplained confounders. Figure 3 (c) suggests that the PCs are not much correlated with ancestry. We cannot tell the exact interpretation of the PCs, but we can observe in Figure 3 (d) that they roughly follow the pattern of the cohorts. We suspect that the time between the donor's death and the measurement of gene expression level is different for the three cohorts, and that this difference can cause the variance in expression.

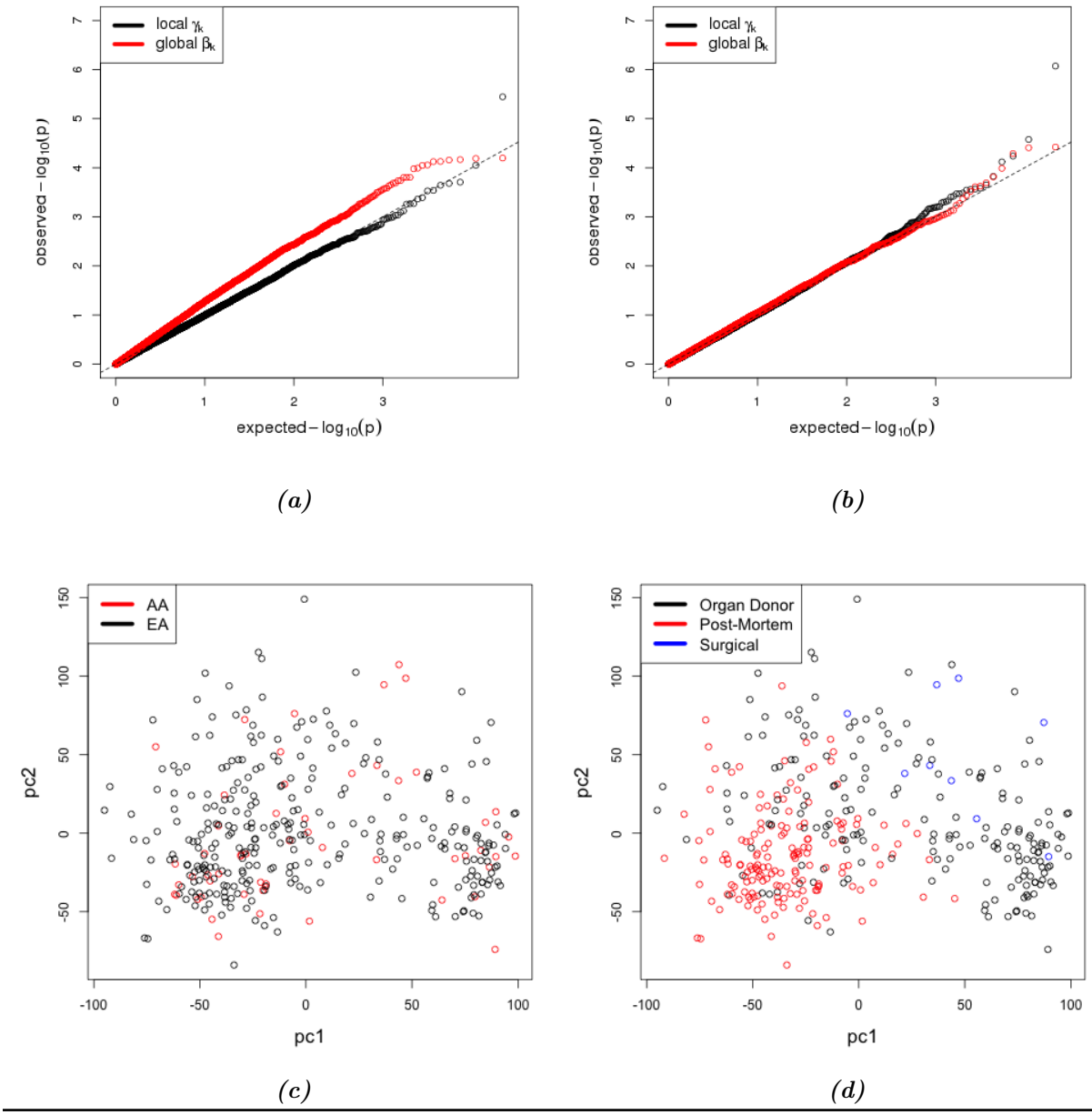


Figure 3: Role of the first two principal components of expression level matrix. *(a) QQ-plot of the p-values from ancestry effects (β_k and γ_k) before controlling the first two principal components of the expression level matrix. (b) The same QQ-plot after controlling the first two PCs of expression level matrix. (c) The first two principal components of expression level matrix comparing African Americans (red) and European Americans (black). (d) The same plot comparing individuals from three cohorts: organ donors, post-mortem, and surgical.*

Ancestry-eGenes

Definition

In this section, we investigate the interaction effects between local ancestry and certain SNP's genotype. The interpretation of a significant non-zero effect of the interaction term would be that minor allele has different effects on the expression of a gene with different local ancestry. An example is illustrated in Figure 4. One individual can have purely European ancestry ($L = 0$) on a gene HLA-C, and another individual can have purely African ancestry ($L = 2$) for the same gene. Then, an allele A on the SNP rs2523578 decreases the gene expression level of the first individual and increases that of the second individual. We define ancestry-eQTLs as these SNPs that differentially affect genes by their local ancestries, and ancestry-eGenes as the genes that have at least one ancestry-eQTL in its cis-region.

We are assuming additive effects for both local ancestry and genotypes, so we will treat them as quantitative variables. The interaction variable LG_{ks} is defined as the element-wise multiplication of the genotype vector G_s and local ancestry L_k vector.

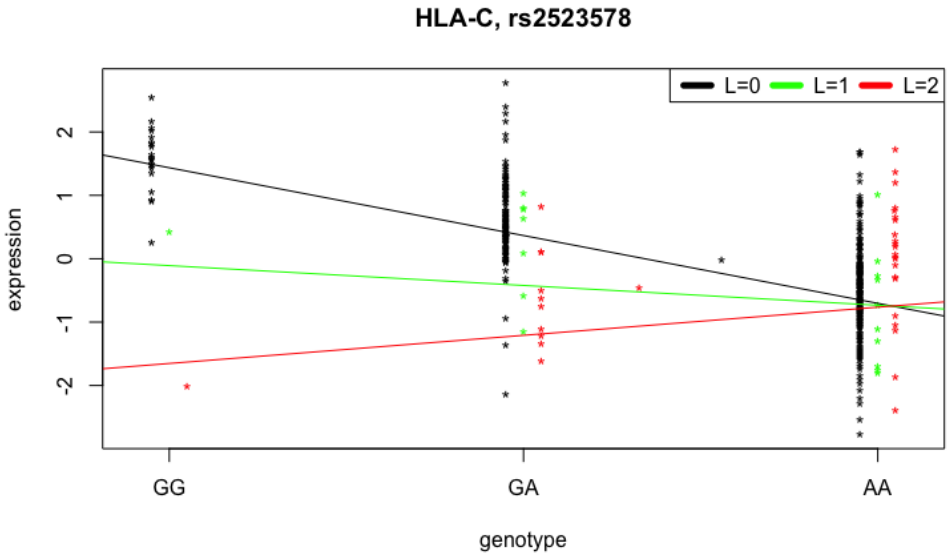


Figure 4: Example of an ancestry-eGene and an ancestry-eQTL. *An illustration of the interaction effect on the expression of HLA-C between genotype of SNP rs2523578 and the gene's local ancestry. The subjects with local ancestry 0 at gene HLA-C are black points (mostly European Americans) with a negative effect of minor allele A on the expression. Those who are locally African (L=2) are marked with red points and the minor allele A has a positive effect on the expression. Subjects marked with green points are heterozygous in local ancestry.*

Computational Challenges

Individually testing the model

$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + G_s\lambda_{ks} + LG_{ks}\theta_{ks} + \nu_{ks}X$$

for gene k and SNP s for each gene-SNP pair requires around $\sim 10^{11}$ linear regressions given we perform 1000 permutations for each gene-SNP pair, but there have been studies that suggested various methods to speed up the process. Matrix eQTL software, for example, redesigns the linear regression into matrix operations [13]. This can be applied to the above model with a small modification.

We denote $n = 356$ as the sample size and S_k as the number of cis-SNPs in gene k. Below, I consider the method for only one gene. Since we can parallelize the procedure in a gene-level process, it is computationally tractable.

Our test statistic for each gene-SNP pair is the t-value computed from the Pearson correlation coefficient between the interaction term LG_{ks} and the expression y_k. Note that if two vectors have mean 0 and sums of squares 1, the inner product is equivalent to the Pearson correlation coefficient. Adjusting this marginal correlation with degrees of freedom can lead to the t-value that tests the association between the interaction term and the expression term. We call this t-value as T_{ks}. Our goal is to get T_{ks} for all gene-SNP pairs (k, s) while controlling for other covariates.

First, we centered all the variables to mean 0 so that we have no need to consider the intercept μ_{ks} term. Then marginal correlation between two variables of interest with other covariates controlled can be acquired by comparing the residuals

$$r_{ks}^{(1)} = y_k - \hat{y}_{ks}$$

$$r_{ks}^{(2)} = LG_{ks} - \hat{LG}_{ks}$$

where $\hat{y}_{ks}$ is the fitted values from linear model

$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + G_s\lambda_{ks} + X\nu_{ks}$$

and $\hat{LG}_{ks}$ is the fitted values from linear model

$$E(LG_{ks}) = \mu_{ks} + \alpha_{ks}\mathbb{1}_{EA} + \beta_{ks}A + \gamma_{ks}L_k + \lambda_{ks}G_s + \nu_{ks}X$$

Then, $r_{ks}^{(1)}$ and $r_{ks}^{(2)}$ were standardized, so that the inner product $< r_{ks}^{(1)}, r_{ks}^{(2)} >$ is the Pearson correlation coefficient. This computation can be done in a large matrix with all s; matrix $r_k^{(1)}$ and

$r_k^{(2)}$ with dimension $n \times S_k$ can be multiplied element-wise, and then the column sums will be the inner products for each SNP.

However, getting the matrix $r_k^{(1)}$ requires running linear regressions $S_k \approx 10^4$ times. Alternatively, we can manually compute the residuals by solving normal equations, but that includes inverting the covariance matrix which can be burdensome. So, we use Gram-Schmidt orthogonalization of all the covariates, including the genotypes. First, we created a covariate matrix with only the variables that do not change for each SNP:

$$C_k = \begin{bmatrix} \mathbb{1}_{EA} & L_k & A & X \end{bmatrix} \in \mathbb{R}^{n \times 6}$$

and make $\tilde{C}_k$ by orthogonalizing each column. In C_k, $\mathbb{1}_{EA}$, L_k and A are vectors but X is a matrix with relevant covariates including gender, age, and the PCs of the expression level matrix. Then we orthogonalized each column of the cis-SNP matrix of G. More specifically, we first orthogonalized C_k by QR decomposition. C has the dimension of only 356×7 ($\mathbb{1}_{EA}$, A, L, gender, age, two PCs of expression level matrix), and this needs to be done only once, so the efficiency of orthogonalization is unimportant in our process. Calling the seven orthogonalized columns $\tilde{c}_1$ .. $\tilde{c}_7$, we used Gram Schmidt algorithm for each column G_{ks} make an orthogonalized genotype matrix $\tilde{G}_k$.

$$\tilde{G}_{ks} = \frac{1}{r_{ss}} \left(G_{ks} - \sum_{j=1}^{7} \tilde{c}_j^T G_{ks} \tilde{c}_j \right)$$

where r_{ss} is the normalizing constant. This is equivalent to

$$\tilde{G}_k = \left(G_k - \tilde{C}_k (\tilde{C}_k^T G_k) \right) \cdot \text{diag}(1/r_{ss})_{s=1,..,S_k}$$

where $\tilde{G}_k, G_k \in \mathbb{R}^{n \times S_k}$ and $\tilde{C}_k \in \mathbb{R}^{n \times 7}$.

Then, we computed the residuals $r_{ks}^{(1)}$ from the orthogonalized covariates:

$$r_{ks}^{(1)} = y_k - \tilde{C}_k (\tilde{C}_k^T y) - \tilde{G}_{ks} (\tilde{G}_{ks}^T y)$$

which can be done through matrix operations as well for all SNPs s.

Computing $r_{ks}^{(2)}$ is a very similar process, and we can re-use the orthogonalized matrices of genotypes and covariates. This procedure allowed us to efficiently compute T_{ks} for all gene-SNP pairs.

Permutation Test

It is difficult to interpret the significance of T_{ks} for each gene-SNP pair due to multiple comparisons and the correlation introduced by the LD structure. In order to circumvent this issue, we define a

gene-specific test statistic as the maximum of the absolute value of T_{ks} and define this test statistic for gene k as $t_k = max_s(|T_{ks}|)$.

To control false discovery rate, we performed a permutation test by re-arranging all variables that are not affected by LD: expression level and the seven covariates. We conducted the same test for each permutation, and recorded the test statistic $t_{\pi_1,k}$, .. $t_{\pi_{1000},k}$ where π_1, .., π_{1000} stands for the 1000 different permutations.

Normally we would calculate the quantile of t_k against $t_{\pi,k}$ to get a p-value, but note that in order to correct for the multiple comparisons, we need the p-values to be accurately specified, especially near 0. Under FDR=0.05, we have to observe the difference between p-values below and above $0.05/22248 \approx 2 \cdot 10^{-6}$. That means we would have to conduct at least a million permutations for each gene, which is practically infeasible. Therefore, we instead fitted a gamma distribution to the $t_{\pi,k}$ values from only 1,000 random permutations, i.e. to the empirical distribution of the test statistics under the null. Figure 5 shows that, for gene HLA-C, the $t_{\pi,k}$ matches gamma distribution very closely, where the parameters for gamma distribution were fitted through the method of moments. Figure 5 also shows that the p-values derived from the fitted gamma are very close to the p-values from the traditional permutation test.

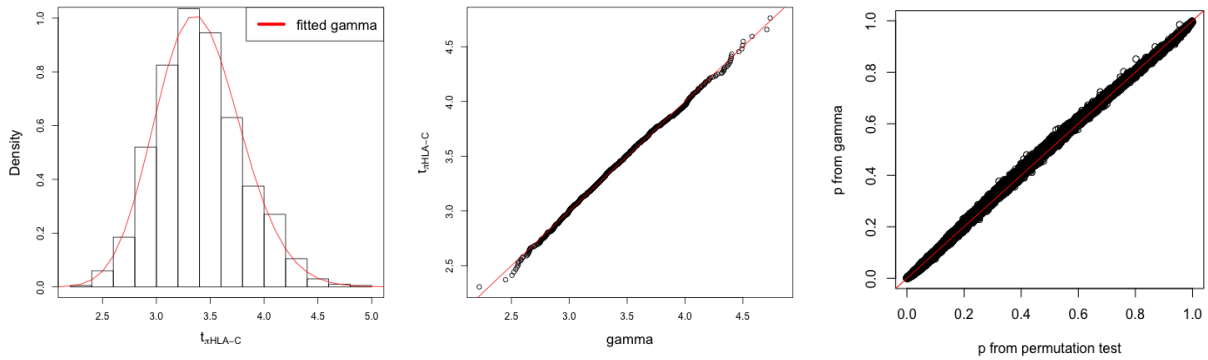


Figure 5: Validity of the Permutation Test. *(a) Histogram of gene-specific test statistics $t_{\pi_j,HLA-C}$ for gene HLA-C under random permutations π_j for $j = 1, ...1000$. The red line is the density of the fitted gamma distribution. (b) QQ-plot of the fitted gamma distribution and the empirical distribution of $t_{\pi_j,HLA-C}$ under random permutations π_j. (c) The p-values from the traditional permutation test plotted against the p-values from the fitted gamma distribution for all gene k in chromosome 6.*

Results and Analysis

Distribution of p-values

The above permutation procedure provides one p-value for each gene from the gamma distribution, and it can be interpreted as the evidence measuring how likely a gene is an ancestry-eGene. The histogram of those p-values are shown in Figure 6. There were 30 genes who had no genotyped cis-SNPs or the interaction term induced a perfect collinearity, so they were removed from the analysis, leaving us 22,218 genes and 22,218 corresponding p-values. The tail part near $p = 1$ shows an unusual behavior, and shows that the null distribution of p-value might not be uniformly distributed. This usually suggests a possible unexplained confounder in the model, but we controlled the population structure and technical artifacts, the two common sources of hidden confounders. Also given the Figure 5 (b) where our fitted p-values through gamma distribution closely matches the p-values from the permutation test, it is likely that the signals are real. Taking these p-values as a true result, we used Benjamini Hochberg procedure with FDR level < 0.05 to identify 201 genes as ancestry-eGenes, while with FDR $<= 0.01$ we identified 110.

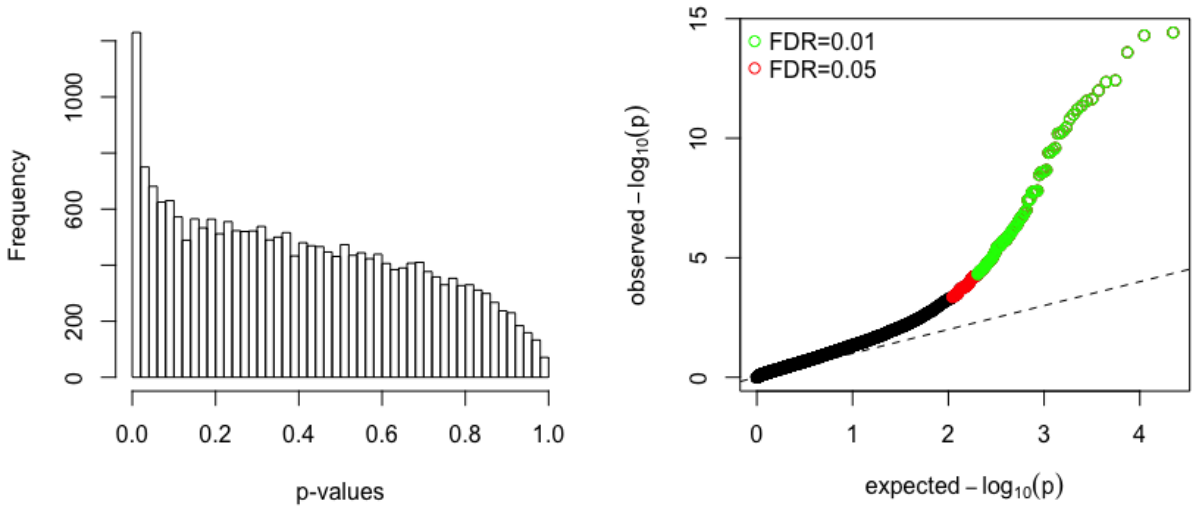


Figure 6: Ancestry-eGenes result. *The left plot shows the histogram of the p-values for each gene from the permutation test with fitted gamma distribution. Each gene's p-value shows the evidence of the gene being an ancestry-eGene. The plot on the right shows the QQ-plot of the same p-values compared to the uniform distribution. The green points are found with FDR level 0.01 and red points are found with FDR level 0.05.*

Functional Clustering

We used DAVID, a software that given a gene list returns a result of functional enrichment by extracting biological roles of the genes, to interpret the result of the list of ancestry-eGenes. One reason that we used DAVID in particular is that it lets the users to specify the background genes for a context-specific interpretation of a gene list [5]. DAVID recommends the users to have more than a hundred genes, so we used the 201 signals that were found with FDR level 0.05. For background reference genes, due to the size restriction of the software, we randomly selected 3,000 genes from the genes that were expressed in muscle skeletal tissue (RPKM > 0.1). DAVID's functional clustering result showed that the two most enriched clusters' keywords were 'MHC' and 'glutathione'. Since it is easy to identify the MHC genes by their locations (Chromosome 6 28,477,797-33,448,354), we counted all signals in MHC regions and conducted the Fisher Exact Test to confirm the enrichment. Out of 201 rejected genes, 14 were in the MHC region, and from the following Table 4

	MHC	non-MHC
Signals (FDR <0.05)	14	187
Non-Signals	165	21852

Table 4: Enrichment of the genes in the MHC region.

the test showed the odds ratio 12.23 and the p-value of 8.56e-13, showing enough evidence for the high enrichment of MHC genes within the discovered ancestry-eGenes. The 14 genes in the MHC region were HLA-C, HLA-DRB5, HLA-DRB1, HLA-DQA2, HLA-DQB2, HLA-K, HLA-U, HLA-DQB1, STK19B, MICA, XXbac-BPG181B23.7, HLA-DPA1, C4A, HLA-DPB1, in the order of the signal strength.

Past works have shown that some of the autoimmune disorders display heterogeneous behaviors for different ethnicities [3] [6]. Given that the MHC region and human leukocyte antigens (HLA) genes are responsible for making receptors for pathogens and play a large role in immunity, this result can be an interesting addition to explaining such phenomenon.

The next enriched group was the keyword glutathione, an important endogenous antioxidant. Especially in the muscle skeletal tissue, it has been suggested to reduce the risk of cellular injury, to improve performance, and to delay muscle fatigue [9], but the mechanism by which ancestry has an interaction effect with minor allele should be investigated further.

Interpretations

One of the possible biological explanations for the interaction effect is the linkage disequilibrium,. Europeans are known to have stronger LD between markers than Africans [1] [4], and such difference can lead to an interaction effect between a tag SNP and expression level. Consider an eQTL s_1 that affects the expression y regardless of the population. Let s_2 be in LD with s_1 for Europeans,

but not for Africans. Then, the variant s_2 will affect y on European chromosome through s_1 but not on African chromosome. In other words, s_2 will have different effects on y on each population, which is interpreted as an interaction between ancestry and s_2 on y.

An example from the real data is illustrated below in Figure 7 for gene SDHDP6. The black points are from s_1 rs654818. Among the SNPs that showed no significant interaction effect, rs654818 had the strongest genotype effect. For all local ancestries, it shows a decreasing effect of the minor allele. The red points are from s_2 rs679429 that showed the strongest interaction effect of local ancestry and genotype. When $L = 0$ (European gene), the genotypes from the two SNPs are almost the same, showing high LD, and therefore they both show decreasing effect of minor allele. When $L = 1$, they show less LD, and when $L = 2$ (African gene), the two genotypes show a completely different pattern. Therefore, we can suspect that the strong LD among Europeans lets the strongest eQTL to pass on its effects onto another variant, leading to an interaction effect.

In order to see how many of our signals were explained by LD, we included genotypes from both s_1 and s_2 in the same model to see if the interaction effect is explained away by the added eQTL.

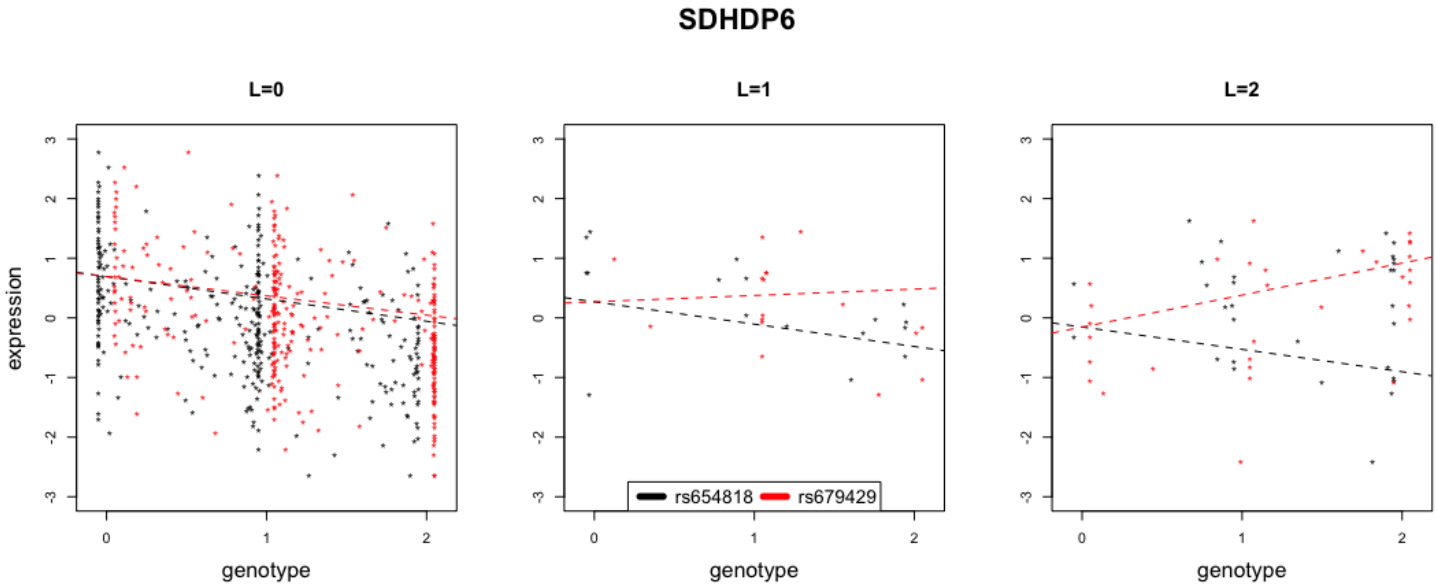


Figure 7: LD causing the interaction effect on gene SDHDP6. *The black SNP rs654818 is an eQTL, and the red SNP rs679429 is an ancestry-eQTL. For individuals with local ancestry 0 (left plot), the two SNPs show high LD through similar genotypes, but for those with local ancestry 1 (middle plot) or 2 (right plot), the two SNPs' genotypes show different patterns. The different LD structure in European and African chromosomes can cause the interaction effect of the rs679429.*

From our full model

$$E(y_k) = \mu_{ks} + \mathbb{1}_{EA}\alpha_{ks} + A\beta_{ks} + L_k\gamma_{ks} + G_s\lambda_{ks} + LG_{ks}\theta_{ks} + X\nu_{ks}$$

we picked s_1 as the variant s with the strongest evidence for $\lambda_{ks} \neq 0$ among s with not enough evidence for $\theta_{ks} \neq 0$ where the significance threshold for the p-value of θ_{ks} was the permutation test threshold with FDR controlled at 0.05 as illustrated in Figure 6: $p = 2 \cdot 10^{-6}$. We picked s_2 as the SNP s with the most significance from testing $\theta_{ks} \neq 0$.

Then we used the model below

$$E(y_k) = \mu_{ks_1s_2} + \mathbb{1}_{EA}\alpha_{ks_1s_2} + A\beta_{ks_1s_2} + L_k\gamma_{ks_1s_2} + G_{s_1}\lambda_{1,ks_1s_2} + G_{s_2}\lambda_{2,ks_1s_2} + LG_{ks_2}\theta_{ks_1s_2} + X\nu_{ks_1s_2}$$

and checked if $\theta_{ks_1s_2}$ is significantly different from 0. The same p-value threshold was used as above ($p = 2 \cdot 10^{-6}$). We found that 79 out of 201 genes lost their significance in interaction effect under the above model, leaving 122 genes with an alternative biological explanation other than LD. Six out of 14 MHC genes lost their signals, but the enrichment was still strong. We attempted to quantify the LD for pure YRI and CEU populations for these two SNPs, but the haplotype information at these loci was not available.

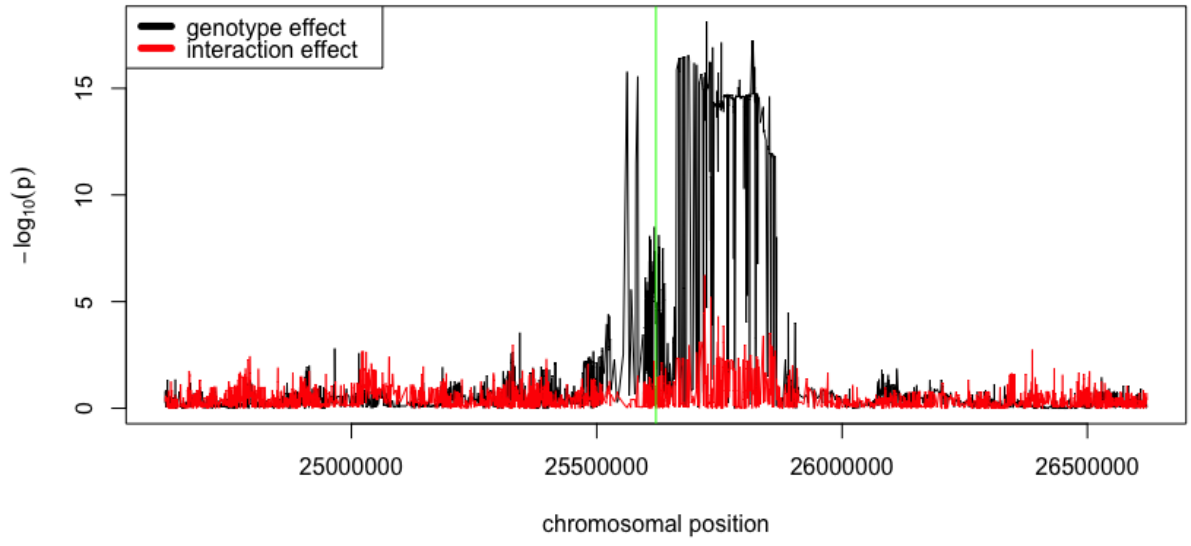


Figure 8: Patterns of eQTLs and ancestry-eQTLs in cis-region of SDHDP6. *This is a typical plot of effect sizes in any gene. We observe strong signals of genotypes around the transcription start site (green), and the strong interaction effects in the similar region with smaller effect sizes.*

We also observed that a strong genotype signal was often followed by a strong interaction signal. The distribution of effect sizes along the cis-region is illustrated in Figure 8. The interaction effects are much smaller than the genotype effects, but they are both concentrated near the transcription start site. It seems as if all eQTLs have at least some interaction effects with the ancestry, and this is a plausible hypothesis especially if the p-value histogram we observed in Figure 6 is an accurate reflection of the effect size, because the histogram suggested that most of the genes have non-zero signals.

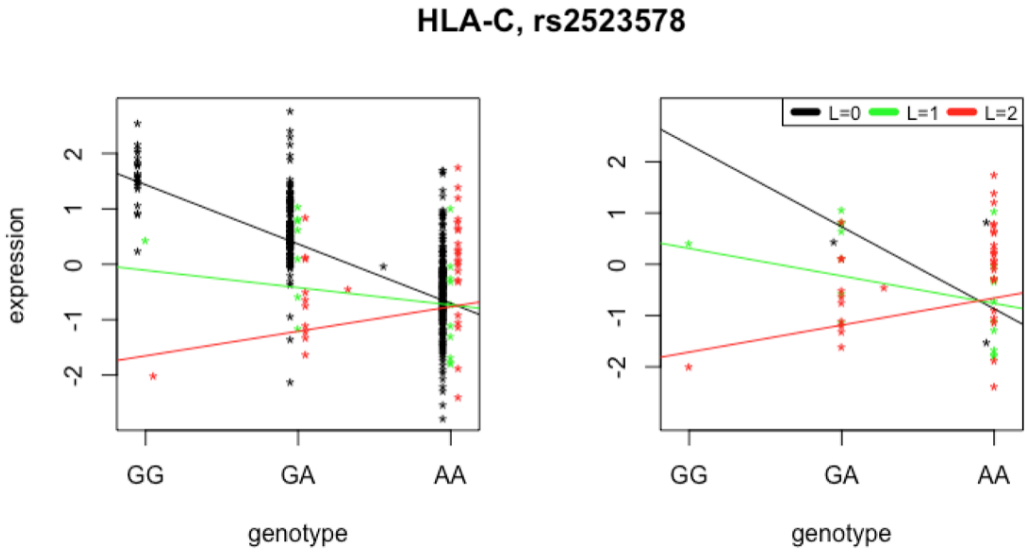
We repeated the above procedure with a different p-value threshold 0.01 for determining the non-significance of interaction effect when we picked the eQTL variant s_1. When we included these SNPs in the model, only 39 of the ancestry-eGenes, including four MHC genes, lost their significance, still preserving the enrichment. This suggests that many of the eQTLs we initially controlled had a fairly strong signal for the interaction. In other words, 40 of the genes' s_1 had the interaction p-value between 0.01 and $2 \cdot 10^{-6}$. The newly selected eQTLs with weaker interaction effect rarely affected the signals of our discovered ancestry-eQTL/ancestry-eGene pair.

It would require further investigation to rigorously state the relationship between the effects of genotype and ancestry-genotype interaction. Since the orthogonalization and marginal correlation trick will no longer be effective, we expect heavy computation to answer such question.

Discussion

We have modeled the expression data and ancestry not only with admixed individuals but also with pure European Americans by assigning 0 to their ancestry, and Figure 9 illustrates the impact in power. Under joint modeling, the interaction effect is significant with p-value $1.97 \cdot 10^{-12}$, but when only the admixed population was used, the p-value was 0.003 which is not low enough after correcting for the multiple comparisons. Therefore, we successfully increased power by including European Americans in our model.

However, we still believe that more samples of admixed population would be helpful for our tests, and although we laid some grounds on the validity of joint modeling with European Americans, only investigating the genome of admixed populations could lead to different results due to unexplained non-genetic variables like lifestyle difference.



	Estimate(SE)	p-value
Intercept	1.44 (0.41)	6e-04
A	0.71 (0.60)	0.24
L	-1.54 (0.26)	5e-09
G	-1.07 (0.06)	9e-50
LG	0.76 (0.10)	2e-12

	Estimate(SE)	p-value
Intercept	2.33 (1.07)	0.03
A	0.30 (0.74)	0.68
L	-2.02 (0.56)	7e-04
G	-1.60 (0.55)	0.01
LG	1.06 (0.34)	3e-03

Figure 9: Increase in power after adding European Americans. *The left plot and the table show the result of fitting linear model on both African Americans and European Americans, while the right plot and table are the result of fitting on only 55 African Americans. The interaction effect and the genotype effect both lose significance after multiple testing adjustments.*

Also, there could be several biological interpretations for genotype-ancestry interaction other than LD. For example, any difference between populations in regulatory elements such as transcription factors or binding sites could have caused the interaction. Another possible explanation is a SNP-SNP interaction. If one variant is a strong ancestry-informative-marker, an interaction with this variant would mimic the effect of the interaction with ancestry. Further research is required to uncover the molecular mechanism of the interaction between ancestry and genotype.

References

[1] Kristin G Ardlie, Leonid Kruglyak, and Mark Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309, 2002.

[2] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[3] Anne Davidson and Betty Diamond. Autoimmune diseases. *New England Journal of Medicine*, 345(5):340–350, 2001.

[4] Katrina AB Goddard, Penelope J Hopkins, Jeff M Hall, and John S Witte. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *The American Journal of Human Genetics*, 66(1):216–234, 2000.

[5] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.

[6] Mikako Mori, Ryo Yamada, Kyoko Kobayashi, Reimi Kawaida, and Kazuhiko Yamamoto. Ethnic differences in allele frequency of autoimmune-disease-associated snps. *Journal of human genetics*, 50(5):264–266, 2005.

[7] Bogdan Paşaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221, 2009.

[8] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genet*, 2(12):e190, 2006.

[9] Scott K Powers, Li Li Ji, and CHRISTIAAN Leeuwenburgh. Exercise training-induced alterations in skeletal muscle antioxidant capacity: a brief review. *Medicine and science in sports and exercise*, 31(7):987–997, 1999.

[10] Alkes L Price, Nick Patterson, Dustin C Hancks, Simon Myers, David Reich, Vivian G Cheung, and Richard S Spielman. Effects of cis and trans genetic ancestry on gene expression in african americans. *PLoS Genet*, 4(12):e1000294, 2008.

[11] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[12] Michael F Seldin, Bogdan Pasaniuc, and Alkes L Price. New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 12(8):523–528, 2011.

[13] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[14] Barbara E Stranger, Eli A Stahl, and Towfique Raj. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, 2011.

[15] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

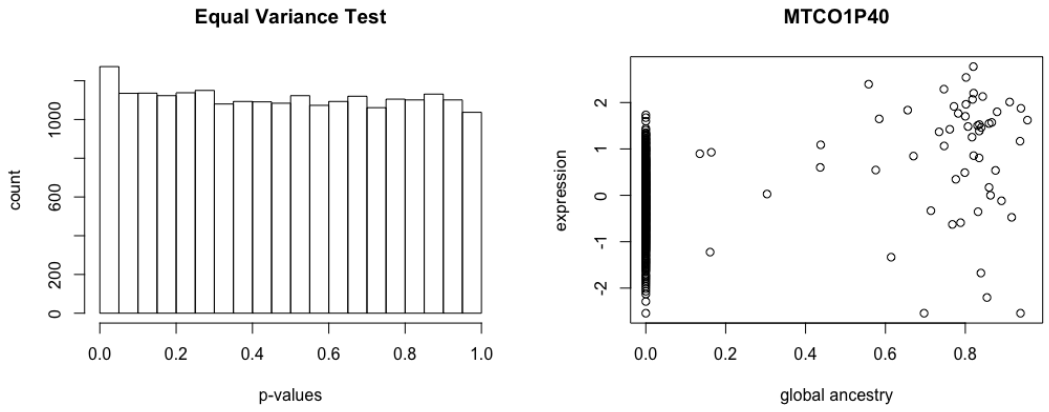
[16] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.

[17] Cheryl A Winkler, George W Nelson, and Michael W Smith. Admixture mapping comes of age. *Annual review of genomics and human genetics*, 11:65–89, 2010.

Appendix

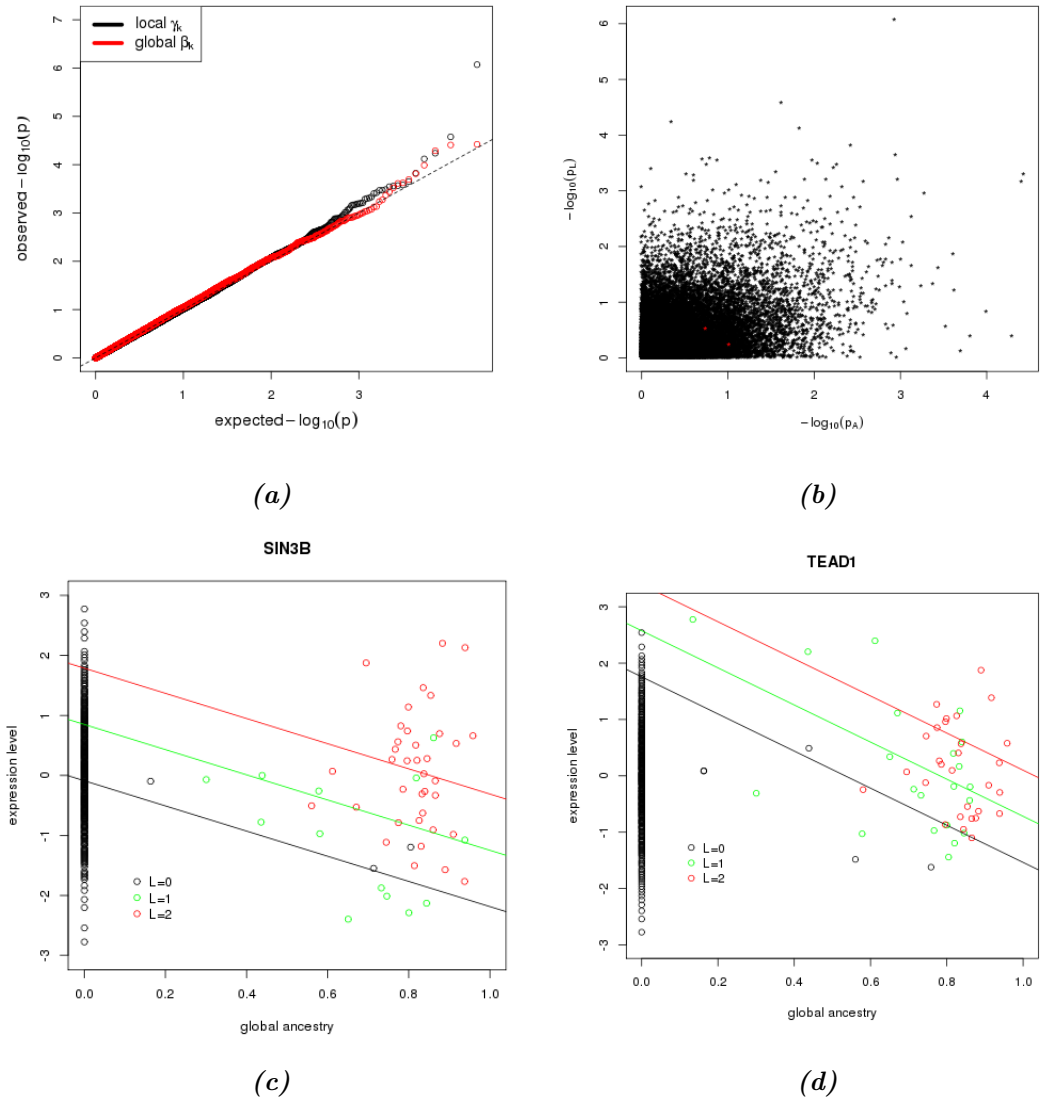
chr1	105,688	chr12	65,037
chr2	112,419	chr13	49,229
chr3	94,733	chr14	45,627
chr4	88,735	chr15	43,355
chr5	83,685	chr16	47,369
chr6	84,896	chr17	40,579
chr7	75,833	chr18	41,711
chr8	74,133	chr19	30,735
chr9	62,243	chr20	35,148
chr10	69,913	chr21	20,189
chr11	67,355	chr22	22,648
		Total	1,361,260

S.Table 1: *Number of SNPs per chromosome whose local ancestry was inferred through LAMP. Only the common SNPs in GTEx and 1000GP were used.*



(a) ***(b)***

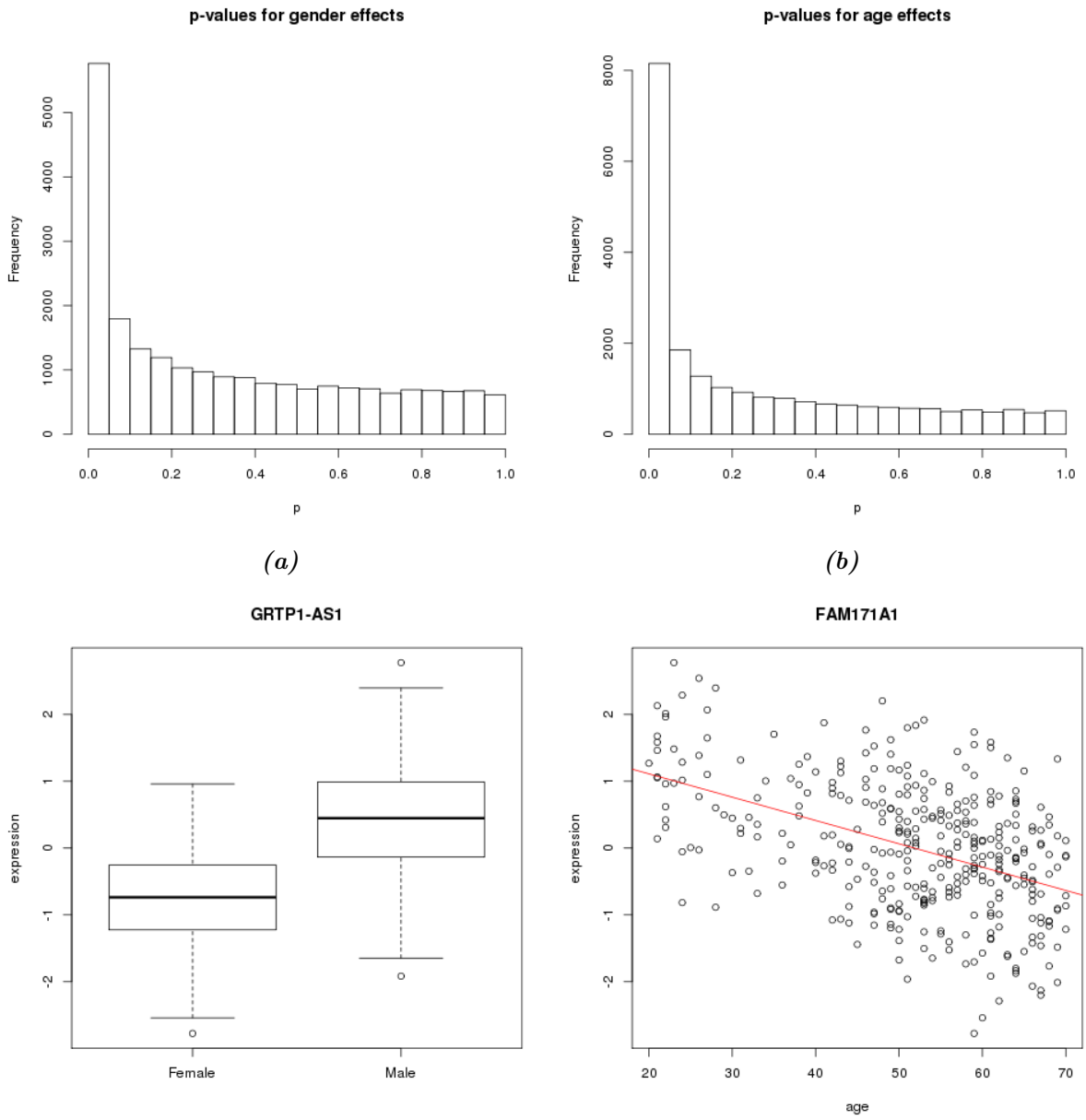
S.Figure 1: Result of constant variance test. *(a) Histogram of p-values from the sample variance ratio F-test of 22,248 genes. (b) The worst case gene with p-value of* $7.04e-06$ *from the F-test.*



S.Figure 2: Local and global ancestry effects. *(a) A QQ-plot of the p-values testing the strength of signals from local ancestry and global ancestry. (b) Relationships between the p-values testing local ancestry and the p-values testing global ancestry. (c) Gene SIN3B that showed the strongest local ancestry effect. (d) The gene TEAD1 showed the strongest global ancestry effect. The linear models fitted for (c) and (d) are shown in Table 2*

	SIN3B			TEAD1		
	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-0.09	0.42	0.83	1.76	0.53	1.11e-03
$\mathbb{1}_{EA}$	0.24	0.38	0.52	-1.31	0.49	7.52e-03
Global	-2.10	0.64	1.17e-03	-3.29	0.79	3.79e-05
Local	0.94	0.19	8.49e-07	0.82	0.23	4.91e-04

S.Table 2: Linear model for genes with strong ancestry effects. *The fitted models for the plotted genes. The covariates like gender and age were fitted but not shown, and not taken into account in the plots (c) and (d).*



(a) *(b)* *(c)* *(d)*

S.Figure 3: Age and gender effects. *(a),(b): p-value histograms for the gender and age effect. (c),(d): genes where gender and age effects respectively are the strongest after controlling for ancestry and 2 PCs of expression. In (d), the red line shows the fitted linear model of the age effect.*