An Ancestry Based Approach for Detecting Interactions

Danny S. Park[1*], Itamar Eskin[2], Eun Yong Kang[3], Eric R. Gamazon[4], Celeste Eng[5], Christopher R. Gignoux[1,6], Joshua M. Galanter[5], Esteban Burchard[1,5], Chun J. Ye[7], Hugues Aschard[8], Eleazar Eskin[3], Eran Halperin[2], Noah Zaitlen[1,5*]

Affiliations:

1. Department of Bioengineering and Therapeutic Sciences. University of California San Francisco. San Francisco, CA.
2. The Blavatnik School of Computer Science. Tel-Aviv University. Tel Aviv, Israel.
3. Department of Computer Science. University of California Los Angeles. Los Angeles, CA.
4. Division of Genetic Medicine, Department of Medicine. Vanderbilt University. Nashville, TN.
5. Department of Medicine. University of California San Francisco. San Francisco, CA.
6. Department of Genetics. Stanford University. Palo Alto, CA.
7. Institute of Human Genetics. University of California San Francisco. San Francisco, CA.
8. Department of Epidemiology. Harvard School of Public Health. Boston, MA.

* Corresponding Author

Email: danny.park@ucsf.edu, noah.zaitlen@ucsf.edu

I. Abstract

Background: Gene-gene and gene-environment interactions are known to contribute significantly to variation of complex phenotypes in model organisms. However, their identification in human associations studies remains challenging for myriad reasons. In the case of gene-gene interactions, the large number of potential interacting pairs presents computational, multiple hypothesis correction, and other statistical power issues. In the case of gene-environment interactions, the lack of consistently measured environmental covariates in most disease studies precludes searching for interactions and creates difficulties for replicating studies.

Results: In this work, we develop a new statistical approach to address these issues that leverages genetic ancestry (θ) in admixed populations. We applied our method to gene expression and methylation data from African American and Latino admixed individuals, identifying nine interactions that were significant at a threshold of $p < 5 \times 10^{-8}$. We replicate two of these interactions and show that a third has previously been identified in a genetic interaction screen for rheumatoid arthritis.

Conclusion: We show that genetic ancestry can be a useful proxy for unknown and unmeasured environmental exposures with which it is correlated

Keywords: Gene-environment Interaction, gene-gene interactions, admixture

II. Background

Genetic association studies in humans have focused primarily on the identification of additive SNP effects through marginal tests of association. There is growing evidence that both gene-gene ($G \times G$) and gene-environment ($G \times E$) interactions contribute significantly to phenotypic variation in humans and model organisms[1-5]. In addition to explaining additional components of missing heritability, interactions lend insights into biological pathways that regulate phenotypes and improve our understanding of their genetic architectures. However, identification of interactions in human studies has been complicated by the multiple testing burden in the case of $G \times G$ interactions, and the lack of consistently measured environmental covariates in the case of $G \times E$ interactions[6,7].

To overcome these challenges, we leverage the unique nature of genomes from recently admixed populations such as African Americans, Latinos, and Pacific Islanders. Admixed genomes are mosaics of different ancestral segments[8] and for each admixed individual it is possible to accurately estimate θ, the proportion of ancestry derived from each ancestral population (e.g. the fraction of European/African ancestry in African Americans)[9]. Studies have demonstrated that an array of environmental and biomedical covariates are correlated with θ[10-13], and we therefore consider its use as a surrogate for unmeasured and unknown environmental exposures. θ is also correlated with the genotypes of SNPs that are highly differentiated between the ancestral populations. Thus θ may also be used as a proxy for detecting epistatic interactions. Therefore, we propose a new SNP by θ test of interaction (AITL) in order to detect evidence of interaction in admixed populations.

We first investigate the properties of our method through simulated genotypes and phenotypes of admixed populations. In our simulations we demonstrate that differential linkage-disequilibrium (LD) between ancestral populations can produce false positive SNP by θ interactions when local ancestry is ignored. To accommodate differential LD, we include local ancestry in our statistical model and demonstrate that this properly controls this confounding factor. We also show that AITL is well powered to detect gene-environment interactions when θ is correlated with the environmental covariates of interest. However, the power for detecting pairwise $G \times G$ interactions at highly differentiated SNPs is lower than direct interaction tests even after accounting for the additional multiple testing burden.

We applied our method to gene expression data from African Americans and DNA methylation data from Latinos. We identified one genome-wide significant interaction($p < 5 \times 10^{-8}$) associated with gene expression in the African Americans and eight significant interactions ($p < 5 \times 10^{-8}$) associated with methylation in the Latinos. We replicated three of the eight interactions associated with DNA methylation in the Latinos and show that the interaction associated with gene expression has also been previously been found to have epistatic effects in the Welcome Trust Case Control Consortium (WTCCC) rheumatoid arthritis case/control dataset[14]. Together, these results provide evidence for the existence of interactions regulating expression and methylation.

III. Results

Simulated Data

To determine the utility of using θ as a proxy for unmeasured and unknown environmental covariates, we applied the AITL to simulated 2-way admixed individuals. We tested θ_1, the proportion of ancestry from ancestral population 1, for interaction with simulated SNPs (see Simulation Framework). Power was computed over 1,000 simulations, assuming 10,000 SNPS being tested, and using a Bonferroni correction p-value cutoff of 5×10^{-6}. We calculated the power using assumed interaction effect sizes (either $\beta_{G \times G}$ or $\beta_{G \times E}$) of 0.1, 0.2, 0.3, and 0.4 (see Simulation Framework). Although the few interactions reported for human traits and diseases show much smaller effect sizes, we simulated large effects because genetic and environmental effect sizes in omics data, such as the expression and methylation data considered here, are known to be of larger magnitude. For example, some cis-eQTL SNPs explain up to 50% of the variance of gene expression[15].

Power When Using θ as a Proxy for Highly Differentiated SNPs

To determine whether using θ as a proxy for a highly differentiated SNPs is more powerful than testing all pairs of potentially interacting SNPs directly, we simulated two interacting SNPS in 1000 admixed individuals (see Simulation Framework). We then tested for an interaction using AITL by replacing the genotypes at the highly differentiated SNP with θ_1. We observed that even with moderate effect sizes, using θ in place of the actual genotypes does not provide any increase in power even after accounting for multiple corrections (see Figure 1a). This is in agreement with recent work showing the limited utility of local ancestry by local ancestry interaction test to identify underlying SNP by SNP interaction when genotype data is available[28]. For the larger effect sizes we simulated, we do see power increasing as the delta between ancestral frequencies increase. The plots show that AITL would be unable to detect anything unless the effect was very strong. Figure 1b reveals that even with the multiple correction penalty, testing all pairwise SNPS directly is always more powerful. We note that when testing the interacting SNPs directly, we used a cutoff p-value of 1×10^{-9} since in theory we were testing all unique pairs of 10,000 SNPs. Based on these results, we would recommend testing for pairs of interacting SNPs directly if pairwise $G \times G$ interactions are a subject of interest in the study. However, when multi-way interactions are considered, AITL may become more powerful (see Discussion).

Power When Using θ as a Proxy Environmental Covariate

When assessing the utility of θ as a proxy for an environmental covariate E, we simulated 3000 individuals. E was simulated such that it was correlated with the individuals' global ancestries in varying degrees (see Simulation Framework). Figure 2 shows the power of the AITL as a function of the Pearson correlation between θ_1 and E. The power of testing E directly is exactly the power of the AITL when the correlation is equal to 1. As expected, as the correlation increases, the power increases as well. When the effect size is 0.1, the power to detect a gene-environment interaction is low whether one uses θ_1 or E. However, both tests are much better powered for effect sizes greater or equal to 0.2, with the AITL's power being dependent on the level of correlation.

Differential LD

To demonstrate that differential LD has the potential to cause inflated test statistics, we ran 10,000 simulations of 1000 admixed individuals. For each individual we simulated 2 SNPs, a causal SNP and a tag SNP. The LD between the tag SNP and causal SNP was different based on the ancestral background the SNPs were on (see Simulation Framework). Over 10,000 simulations, we computed the mean χ_1^2 test-statistic for the AIT and the AITL. We note that the phenotypes for these simulations were generated under a model that assumed no interaction. We observed a mean $\chi_1^2 = 0.996$ with a standard deviation of 1.53 for AITL. AIT, which does not condition on local ancestry, had a mean $\chi_1^2 = 3.59$ with a standard deviation of 3.60. We also looked at λ_{GC}or genomic control, as another indicator of test-statistic inflation[16]. λ_{GC} compares the median observed χ^2 test-statistic versus the true median under the null. In our simulations, we observed $\lambda_{GC} = 5.81$ for AIT and $\lambda_{GC} = 0.980$ for AITL (see Supplementary Figure S1). Last, we computed the proportion of test-statistics that passed a p-value threshold of .05 and .01 in our simulations. The AIT had 3687 statistics passing a p-value of .05 and 1687 at a threshold of .01, whereas AITL had 464 and 96 at the same p-value thresholds. The results for AITL are as expected under a true null. The results from our simulations show that not accounting for local ancestry can result in inflated test-statistics and can potentially lead to false positive findings.

Real Data

Coriell Gene Expression Results

We first applied our method to the Coriell gene expression dataset[17]. The Coriell cohort is composed of 94 African-American individuals and the gene expression values of ~8800 genes in lymphoblastoid cell lines (LCLs). Since African Americans derive their genomes from African and European ancestral backgrounds, we tested for interaction between a given SNP and the proportion of European ancestry, θ_{EUR}. Each SNP by θ_{EUR} term was tested once for association with the expression of the gene closest to the SNP. We observed well-calibrated statistics with a λ_{GC} equal to 1.04 (see Supplementary Figure S2). In the LCLs, we found that interaction of rs7585465 with θ_{EUR} was associated with ERBB4 expression (AITL $p = 2.95 \times 10^{-8}$, Marginal $p = 0.404$) at a genome-wide significant threshold ($p \leq 5 \times 10^{-8}$).

Given that the gene expression values come from LCLs (all cultured according to the same standards), the SNPs are either interacting with epigenetic alterations due to environmental exposures that have persisted since transformation into LCLs or the signals are driven by epistatic interactions. In our simulations, we showed that using θ as a proxy for a single highly differentiated SNP is underpowered compared to testing all pairs of potentially interacting SNPs directly. However, there are many SNPs that are highly differentiated across the genome with which θ will be correlated. It is therefore possible that θ is capturing the interaction between the aggregate of all differentiated trans-SNPs (i.e. global genetic background) and the candidate SNP. This is consistent with a recently reported finding, conducted in human iPS cell lines, that genetic background accounts for much of the transcriptional variation[2,18].

GALA II Methylation Results

We searched for interactions in methylation data derived from a study of asthmatic Latino individuals called the Genes-environments and Admixture in Latino Americans (GALA II)[19]. The methylation data is composed of 141 Mexicans and 184 Puerto Ricans. As the phenotype, we used DNA methylation measurements on ~300,000 markers from peripheral blood. As we had done with gene expression, we tested for interaction between a given SNP and θ_{EUR} using AITL. All SNPs within a 1 MB window centered around the methylation probe were tested. We used the European component of ancestry because it is the component shared most between Mexicans and Puerto Ricans (see Table 1). We observed well-calibrated test statistics with λ_{GC} equal to 1.06 in the Mexicans and 0.96 in the Puerto Ricans (see Supplementary Figure S3). We tested 128,794,325 methylation-SNP pairs which results in a Bonferroni corrected p-value cutoff of 3.88×10^{-10}. However, this cutoff is extremely conservative given the tests are not all independent. We therefore we report all results that are significant at 5×10^{-8} in either set as an initial filter. We found 5 interactions in the Mexicans and 3 in the Puerto Ricans that are significant at this threshold (see Table 2).

Unlike the Coriell individuals, who are 2-way admixed, the GALA II Latinos are 3-way admixed and derive their ancestries from European, African, and Native American ancestral groups. Consequently, to confirm that incomplete modeling or better tagging on one of the non-European ancestries was not driving the results, we retested all significant interactions including a second component of ancestry for AITL. In the case of the Mexicans, we included African and European ancestry, and in the case of the Puerto Ricans, we included European and Native American ancestry. Even after adjusting for the second ancestry the interactions between SNP and θ_{EUR} remained highly significant (see Supplementary Table 1).

We then performed a replication study of the significant Puerto Rican associations in the Mexican cohort and vice versa. To account for the fact that we are replicating eight total results across both populations, we used a Bonferroni corrected p-value threshold equal to $.05/8 = 6.25 \times 10^{-3}$. The interaction of rs4312379 and rs4312379 with ancestry in the Puerto Ricans replicated in the Mexicans. Furthermore, there was a highly significant overall trend of association in the replication study (permutation $p < 1 \times 10^{-4}$). The lack of direct replication for other specific interactions might be driven in part by the fact that Mexicans and Puerto Ricans have distinct genetics and environmental exposures. Overall, our results from the GALA II cohort suggest there are both genetic and environmental interactions that have yet to be discovered in admixed individuals.

IV. Discussion and Conclusions

For many disease architectures, interactions are believed to be a major component of missing heritability[20]. Finding new interactions has proven to be difficult for logistical, statistical, biological, and computational reasons. In this study, we have demonstrated that in admixed populations, testing for gene by θ interactions can be leveraged to overcome some of the difficulties typically encountered when searching for interactions. Although our method does not provide details as to which covariate is interacting with a genetic locus, it can show whether an interaction effect exists in a given dataset. Furthermore, the drawback of not having consistently measured environmental covariates is addressed by our method. Genetic ancestry is nearly perfectly replicable, especially with respect to environmental measurements that can be influenced by a myriad of factors between studies. Testing for the presence of interaction using a nearly perfectly reproducible covariate may enhance our understanding of the genetic basis of disease and other traits. Our method also provides the additional benefit of not being confounded by interactions between unaccounted-for covariates[21].

Our simulations showed that genetic ancestry can be a good proxy for an environmental covariate depending on the correlation between the two. On the other hand, our simulations also revealed that testing SNP by θ where genetic ancestry is a proxy for a single highly differentiated SNP is severely underpowered. Although genetic ancestry in our simulations was not a good proxy for a single SNP, our results from cell lines suggest that genetic ancestry is a good proxy for genetic background, since all highly differentiated SNPs across the genome will be correlated with genetic ancestry. There are also other contexts in which modeling SNP by θ may be useful, such as in heritability estimation. We have previously shown that local ancestry from admixed populations can be leveraged to estimate the total additive heritability of a phenotype[22]. We could also use the SNP by θ interaction terms to estimate heritability in a mixed-model framework because genetic ancestry is correlated with many genetic markers and environmental covariates[23]. To do so, we would introduce an additional variance component computed from SNP by θ across the genome in addition to the component computed from SNPs only. In this scenario, genetic ancestry would represent an aggregate of potential interacting genetic and environmental covariates. It will be interesting to see whether such estimations yield more accurate measures of heritability.

In our analysis of real data, we discovered gene by θ interactions associated with genes that have known interactions. In the Coriell data, we found that ERBB4 gene expression was associated with a SNP by θ interaction. Notably, ERBB4 gene expression has been previously shown to be modulated by SNP-SNP interactions in Schizophrenic individuals of European background[24,25]. Furthermore, the SNP rs7585465 in ERBB4 that we identified has been shown to be part of multiple epistatic interactions from the results of interaction analysis for rheumatoid arthritis in the WTCCC; of note, this SNP was in interaction for this disease with a highly population-differentiated SNP rs163673 (which has allele "A" frequency of 0.11 in the reference African population YRI and 1.0 in the reference European ancestry population CEU)[26]. In the GALA II Mexicans, the interaction of rs925736 with ancestry was associated with the methylation of HDAC4, a known histone deaceytlase (HDAC). In concert with DNA methylases, HDACs function to regulate gene expression by altering chromatin state[27]. In Europeans, HDACs have been shown to be associated with lung function through direct genetic effects and through environmental interactions[28,29]. For the GALA II Puerto Ricans, rs17091085 showed an interaction associated with the methylation state of SERPINA6. Of note, interaction between birth weight and SERPINA6 has been previously associated with Hypothalamic-Pituitary-Adrenal axis function[30]. Further investigations of our interaction findings are thus warranted.

Our analysis revealed the existence of interactions but does not provide a direct way to determine the covariate that is interacting with a SNP. Further work will need to be done to uncover the exact environmental exposures or genetic loci with which SNPs are interacting. The existence of gene by θ interactions in GALA II underscores why modeling interactions should be considered for future association studies and heritability estimation in admixed populations.

V. Materials and Methods

Our approach is best illustrated with an example. First consider testing a SNP s for interaction with an environmental covariate E. θ can serve as a proxy for E if the two are correlated, even if E is unknown or unmeasured (see Figure 3a). Now consider testing s for interaction with a SNP $j \neq s$ that is highly differentiated in terms of ancestral allele frequencies. For example, a SNP that has a high allele frequency in one ancestral population and a low allele frequency in the other ancestral population. θ can be used as a proxy for j because θ and the genotypes of SNP j will be correlated. Consider the case where j has a frequency of 0.9 in population 1 and frequency of 0.1 in population 2. Individuals with large values of θ_1 are more likely to have derived j from population 1 and on average have greater genotype values at j. Similarly, individuals with small values of θ_1 are more likely to have derived *j* from population 2 and on average have smaller genotype values. Thus, θ will be correlated with the genotypes of the individuals for highly differentiated SNPs and can serve as a proxy for detecting interactions (see Figure 3b).

Consider an admixed individual *i* who derives his or her genome from *k* ancestral populations. We denote individual *i*'s global ancestry proportion as $\theta_i = \langle\theta_{i1}, \theta_{i2}, _{\dots}, \theta_{ik}\rangle$, where $\sum_k \ \theta_{ik} = 1$. The local ancestry of individual *i* at a SNP *s* is denoted as $\gamma_{ais} \in \{0, 1, 2\}$ and is equal to the number of alleles from ancestry $a \in \{1 \dots k\}$ inherited at SNP *s.* Current methods allow us to estimate ancestry directly from genotype data both globally and at specific SNPs[9,31,32]. We denote the genotype of an individual *i* at SNP *s* as $g_{is} \in \{0, 1, 2\}$ and the corresponding phenotype as y_i.

In this work, we model phenotypes in an additive linear regression framework, but note that our method can easily be extended to a logistic framework for case-control data. Assuming *n* (unrelated) individuals, define $\vec{y}$ to be the vector of all individuals' phenotypes. The model for the phenotype is then

$$\vec{y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$$

where $\vec{\varepsilon} \sim \mathcal{N}(0, \sigma)$ is a *n*×1 vector of error terms, $\boldsymbol{X}$ is a *n*×*v* matrix of *v* covariates, and $\vec{\beta}$ is a *v*×*1* vector of the covariate effect sizes. We note that in our notation $\vec{v}^2 = \vec{v}^T\vec{v}$ for a vector $\vec{v}$. Assuming independence, the likelihood under this model is:

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2\sigma^2}(\vec{y} - \boldsymbol{X}\beta)^2\right)$$

We can compute the log-likelihood ratio statistic (D) using a maximum likelihood approach:

$$D = \underset{\beta,\sigma}{\text{argmax}}\ (\log L_1 - \log L_0) = -\left(n\log(\hat{\sigma}_{L_1}) + \frac{(\vec{y} - \boldsymbol{X}\hat{\beta}_{L_1})^2}{2\hat{\sigma}^2_{L_1}}\right) + \left(n\log(\hat{\sigma}_{L_0}) + \frac{(\vec{y} - \boldsymbol{X}\hat{\beta}_{L_0})^2}{2\hat{\sigma}^2_{L_0}}\right)$$

The maximum likelihood estimator (MLE) of the effect sizes is $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$, and the MLE of the error variance is $\hat{\sigma}^2 = \frac{1}{n}\left(\vec{y} - \boldsymbol{X}\hat{\beta}\right)^2$. Here, L_1 is the likelihood under the alternative and L_0 is the likelihood under the null. $(\hat{\beta}_{L_1}, \hat{\sigma}^2_{L_1})$ and $(\hat{\beta}_{L_0}, \hat{\sigma}^2_{L_0})$ are the effect sizes and error variance estimates that maximize the respective likelihoods. D is distributed as χ^2 with k degrees of freedom (*df*), where k is the number of parameters constrained under the null.

1-df Ancestry Interaction Test (AIT)

The first test we present is the standard direct test of interaction. We test for a SNP's interaction with θ instead of an environmental covariate or another genotype. Let $\vec{g}_s = \langle g_{1s} \dots g_{ns}\rangle$ be the vector of the individuals' genotypes at SNP *s*, $\vec{\theta}_a = \langle \theta_{1a} \dots \theta_{na}\rangle$ be the vector of their global ancestries for ancestry a, and $\vec{g}_s \times \vec{\theta}_a$ be the vector of interaction terms which result from the component-wise multiplication of the genotype and global ancestry vectors. We test the alternative hypothesis $\left(\hat{\beta}_{G\times\theta} \neq 0\right)$ against the null hypothesis $\left(\hat{\beta}_{G\times\theta} = 0\right)$.

$$H_1: \vec{y} = \vec{g}_s + \vec{g}_s \times \vec{\theta}_a + \vec{\theta}_a$$

$$H_0: \vec{y} = \vec{g}_s + \vec{\theta}_a$$

In this test of interaction, we test a single ancestry versus the other ancestries that may be present in the population of interest. One parameter is constrained under the null which results in a statistic with k=1 *df*. Let $\hat{\beta}_{L_{\{0,1\}}(s)}$, $\hat{\beta}_{L_{\{0,1\}}(G\times\theta)}$, and $\hat{\beta}_{L_{\{0,1\}}(\theta)}$ denote the effect sizes of genotype, interaction, and global ancestry under a given hypothesis respectively. The statistic is given below.

$$D = -\left(n\log\left(\hat{\sigma}_{L_1}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle\hat{\beta}_{L_1(s)}, \hat{\beta}_{L_1(G\times\theta)}, \hat{\beta}_{L_1(\theta)}\rangle\right]^2}{2\hat{\sigma}^2_{L_1}}\right) + \left(n\log\left(\hat{\sigma}_{L_0}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle\hat{\beta}_{L_0(s)}, 0, \hat{\beta}_{L_0(\theta)}\rangle\right]^2}{2\hat{\sigma}^2_{L_0}}\right)$$

where $\boldsymbol{X}$ is an $n \times 3$ matrix composed of $\vec{g}_s$, $\vec{\theta}_a$, and $\vec{g}_s \times \vec{\theta}_a$ as columns.

1-df Ancestry Interaction Test with Local Ancestry (AITL)

Given that the individuals we analyze in this work are assumed to be admixed, there is potential for confounding due to differential LD. An interaction that is not driven by biology could occur due to the possibility that a causal variant may be better tagged by a SNP being tested on one ancestral background versus another (See Figure 3c). We account for the different LD patterns on varying ancestral backgrounds by including local ancestry as an additional covariate in AITL. By including local ancestry, we assume that the SNP being tested is on the same local ancestry block as the causal SNP that it may be tagging. Such an assumption is reasonable because admixture in populations such as Latinos and African Americans are relatively recent events and their genomes have not undergone many recombination events. As a result, local ancestry blocks on average stretch for several hundred kilobases[33,34].

Let $\vec{\gamma}_{as} = \langle \gamma_{a1s} \ldots \gamma_{a1s} \rangle$ be the vector of local ancestry calls for all individuals for ancestry a and let $\vec{g}_s \times \vec{\gamma}_{as}$ be the interaction terms from piecewise multiplication of the two vectors. We use the following alternative and null hypotheses:

$$H_1 \colon \vec{y} = \ \vec{g}_s + \vec{g}_s \times \vec{\theta}_a + \vec{\theta}_a + \vec{\gamma}_{as} + \vec{g}_s \times \vec{\gamma}_{as}$$

$$H_0 \colon \vec{y} = \ \vec{g}_s + \vec{\theta}_a + \vec{\gamma}_{as} + \vec{g}_s \times \vec{\gamma}_{as}$$

Here we are testing for an interaction effect, i.e. $\hat{\beta}_{G\times\theta} \neq 0$, and constrain one parameter under the null resulting in a statistic with k=1 *df*. Let $\hat{\beta}_{L_{\{0,1\}}(G\times\gamma)}$ and $\hat{\beta}_{L_{\{0,1\}}(\gamma)}$ denote the effect sizes of the interaction between genotype and local ancestry and just local ancestry, respectively. The log likelihood ratio statistic is given by

$$D = -\left(n\log\left(\hat{\sigma}_{L_1}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle \hat{\beta}_{L_1(s)}, \hat{\beta}_{L_1(G\times\theta)}, \hat{\beta}_{L_1(\theta)}, \hat{\beta}_{L_1(\gamma)}, \hat{\beta}_{L_1(G\times\gamma)} \rangle\right]^2}{2\hat{\sigma}^2_{L_1}} \right)$$

$$+ \left(n\log\left(\hat{\sigma}_{L_0}\right) + \frac{\left[\vec{y} - \boldsymbol{X}\langle \hat{\beta}_{L_0(s)}, 0, \hat{\beta}_{L_0(\theta)}, \hat{\beta}_{L_0(\gamma)}, \hat{\beta}_{L_0(G\times\gamma)} \rangle\right]^2}{2\hat{\sigma}^2_{L_0}} \right)$$

where $\boldsymbol{X}$ is an $n \times 5$ matrix composed of $\vec{g}_s$, $\vec{\theta}_a$, $\vec{g}_s \times \vec{\theta}_a$, $\vec{\gamma}_{as}$, and $\vec{g}_s \times \vec{\gamma}_{as}$ as columns. All of these test statistics are straightforwardly modified to jointly incorporate several ancestries in the case of multi-way admixed populations.

Simulation Framework

For all our simulations, we simulated 2-way admixed individuals. Global ancestry for ancestral population 1 (θ_1) was drawn from a normal distribution with $\mu = 0.7$ and $\sigma = 0.2$. Individuals with $\theta_1 > 1$ or $\theta_1 < 0$ were assigned a value of 1 or 0, respectively. We simulated phenotypes of individuals to investigate our method in three different scenarios: gene-environment interactions, pairwise gene-gene interactions, and false positive interactions due to local differential tagging.

To simulate phenotypes under the situation of a gene-environment interaction, we simulated a single SNP. For each individual *i*, we assigned the local ancestry or the number of alleles derived from population 1 (γ_{ai}) for each haplotype by performing two binomial trials with the probability of success equal to θ_{i1}. We then drew ancestry specific allele frequencies following the Balding-Nichols model by assuming a $F_{ST} = 0.16$ and drawing two ancestral frequencies, p_1 and p_2, from the following beta distribution[35].

$$p_1, p_2 \sim Beta\left(\frac{p(1-F_{ST})}{F_{ST}}, \frac{(1-p)(1-F_{ST})}{F_{ST}}\right)$$

where p is the underlying MAF in the entire population and is set to 0.2. Genotypes were drawn using a binomial trial for each local ancestry haplotype with the probability of success equal to p_1 or p_2 for values of $\gamma_{ai} = 0$ or 1, respectively. Environmental covariates correlated with θ_1, E_i, were generated for each individual *i* by drawing from a normal distribution $\mathcal{N}(\mu = \theta_{i1}, \sigma_E)$. σ_E was varied from 0 to 5 in increments of 0.005 to create E_i's that were correlated with individuals' global ancestries in varying degrees. We generated phenotypes for individuals assuming only an interaction effect by drawing from a normal distribution, $\mathcal{N}(\mu = \beta_{G\times E} \times \; g_{i1} \times E_i, \sigma = 1)$ for a given interaction effect size $(\beta_{G\times E})$.

To simulate phenotypes based on gene-gene interactions, we simulated two SNPs. At both SNPs, we assigned the local ancestry values as described for the gene-environment case. We assigned genotypes for individuals at the first SNP assuming an allele frequency of 0.5 for both populations and drawing from two binomial trials. We assigned genotypes at the second SNP over a wide range of ancestry specific allele frequencies to simulate different levels of SNP differentiation. Ancestry specific allele frequencies were initially $p_1 = p_2 = 0.5$ and iteratively increasing p_1 by 0.005 while simultaneously decreasing p_2 by 0.005 until p_1 = 0.05 and p_2 = 0.95. Genotypes at the second SNP were drawn using the same approach described for gene-environment. Using the simulated genotypes, phenotypes were drawn from a normal distribution, $\mathcal{N}(\mu = \beta_{G\times G} \times \; g_{i1} \times g_{i2}, \sigma = 1)$, where g_{is} is the genotype for individual *i* at the simulated SNP *s*.

To simulate the scenario of differential LD on different ancestral backgrounds leading to false positives, we simulated phenotypes based on a single causal SNP that was tagged by another SNP. At both SNPs, local ancestries were assigned as described previously and genotypes were drawn using ancestry specific allele frequencies. Ancestral allele frequencies were assigned such that the average r^2 between the causal and tag SNP was 0.272 on the background of ancestral population 1 and 0.024 on the background of ancestral population 2. Thus, the tag SNP was only a tag on the population1 background and not on the population 2 background. Phenotypes were drawn from a normal distribution, $\mathcal{N}(\mu = \beta_{Causal} \times g_{ic}, \sigma = 1)$, assuming no interaction and $\beta_{Causal} = 0.7$, where g_{ic} is the genotype of individual *i* at the causal variant.

We implemented our approach in an R package (GxTheta), which is available for download at http://www.scandb.org/newinterface/GxTheta.html

Data Normalization

Gene Expression Normalization

Gene expression data (see Results) were first standardized for each gene such that mean expression was 0 and variance was 1. We then computed a covariance matrix of individual's expression values and performed PCA on the covariance matrix. Residuals were computed for all expression values by adjusting for the top 10 principal components and the mean for each gene was added back to the residuals. Due to the high dynamic range of gene expression compared to methylation we conservatively chose to additionally perform quantile normalization. We then sorted the gene expression residuals and used the quantiles of their rank order to draw new expression values from a normal distribution, $\mathcal{N}(\mu = 0, \sigma = 1)$, by using the inverse cumulative density function[24,25].

Methylation Data Normalization

Raw methylation values (see Results) were first normalized using Illumina's control probe scaling procedures. All probes with median methylation less than 1% or greater than 99% were removed and the remaining probes were logit-transformed as previously described[36]. To control for extreme outliers, we truncated the distribution of methylation values. For a given probe, we first computed the mean and standard deviation of the methylation values. We then set any methylation values deviating more than 2.58 standard deviations from the mean to the methylation value corresponding to the 99.5th quantile.

Availability of Supporting Data

The Coriell data is available from dbGAP under accession number phs000211.v1.p1. The GALA and SAGE data is available by emailing the study organizers at https://pharm.ucsf.edu/gala/contact.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

DSP, IE, EK, EE, EH and NZ designed research. DSP, IE, EK, and NZ performed research. DSP, IE, EK, EE, CE, CRG, JMG, EG, HA, CJY, EE, EH, and NZ contributed new reagents/analytic tools. DSP, ERG, and NZ wrote the manuscript. All authors read and approved the final manuscript.

Description of Additional Data Files

The following data are available with the online version of this paper. The Supplemental contains QQ-plots for the simulations and real analyses performed as well as a table containing p-values for the 2-component ancestry analysis of the GALA methylation data.

Acknowledgements

We would like to thank Lancelote Leong for his helpful manuscript comments.

References

1. Hemani G, Shakhbazov K, Westra H-J, Esko T, Henders AK, Mcrae AF, et al. Detection and replication of epistasis influencing transcription in humans. Nature. Nature Publishing Group; 2014 Apr 10;508(7495):249–53.

2. Rouhani F, Kumasaka N, de Brito MC, Bradley A, Vallier L, Gaffney D. Genetic Background Drives Transcriptional Variation in Human Induced Pluripotent Stem Cells. Gibson G, editor. PLoS Genet. 2014;10(6):e1004432.

3. Kang EY, Han B, Furlotte N, Joo JWJ, Shih D, Davis RC, et al. Meta-Analysis Identifies Gene-by-Environment Interactions as Demonstrated in a Study of 4,965 Mice. Gibson G, editor. PLoS Genet. Public Library of Science; 2014 Jan 9;10(1):e1004022.

4. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011 Mar;61(2):69–90.

5. Lee M, Raj T, Castillo IW. ImmVar Project: Genetic architecture of leukocyte gene expression in healthy humans. JOURNAL OF ...; 2012.

6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. Nature Publishing Group; 2009 Oct 8;461(7265):747–53.

7. Eichler EE, Flint J, Gibson G, Kong A, Leal SM. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews .... 2010.

8. Seldin MF, Pasaniuc B, Price AL. New approaches to disease mapping in admixed populations. Nature Reviews Genetics. Nature Publishing Group; 2011 Aug 1;12(8):523–8.

9. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. Cold Spring Harbor Lab; 2009 Sep 1;19(9):1655–64.

10. Choudhry S, Burchard EG, Borrell LN, Tang H, Gomez I, Naqvi M, et al. Ancestry–Environment Interactions and Asthma Risk among Puerto Ricans. Am J Respir Crit Care Med. American Thoracic Society; 2012 Dec 20;174(10):1088–93.

11. Karter AJ, Ferrara A, Liu JY, Moffet HH, Ackerson LM, Selby JV. Ethnic Disparities in Diabetic Complications in an Insured Population. JAMA. American Medical Association; 2002 May 15;287(19):2519–27.

12. Burchard EG, Ziv E, Coyle N, Gomez SL. The importance of race and ethnic background in biomedical research and clinical practice. New England Journal … [Internet]. 2003. Available from: http://rds.epi-ucsf.org/ticr/syllabus/courses/23/2012/03/29/Lecture/readings/The%20Importance%20of%20Race%20%26%20Ethnicity%20in%20Biomedical%20Research%20and%20Clinical%20Practice..pdf

13. Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, Colangelo L, et al. Genetic Ancestry in Lung-Function Predictions. N Engl J Med. 2010 Jul 22;363(4):321–30.

14. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. Nature Publishing Group; 2007 Jun 7;447(7145):661–78.

15. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. Nature Publishing Group; 2012 Oct 1;44(10):1084–9.

16. Devlin B, Roeder K. Genomic Control for Association Studies. Biometrics [Internet]. Blackwell Publishing Ltd; 2004 May 25;55(4):997–1004. Available from: http://doi.wiley.com/10.1111/j.0006-341X.1999.00997.x

17. Simon-Sanchez J, Scholz S, Fung H-C, Matarin M, Hernandez D, Gibbs JR, et al. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet. Oxford University Press; 2007 Jan 1;16(1):1–14.

18. Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee M-C, et al. Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. Gibson G, editor. PLoS Genet. Public Library of Science; 2014 Aug 14;10(8):e1004549.

19. Borrell LN, Nguyen EA, Roth LA, Oh SS, Tcheurekdjian H, Sen S, et al. Childhood Obesity and Asthma Control in the GALA II and SAGE II Studies. dx.doi.org. American Thoracic Society; 2013. 6 p.

20. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics. Nature Publishing Group; 2010 Jun 1;11(6):446–50.

21. Keller MC. Gene × Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. Biological Psychiatry. 2014 Jan;75(1):18–24.

22. Zaitlen N, Pasaniuc B, Sankararaman S, Bhatia G, Zhang J, Gusev A, et al. Leveraging population admixture to characterize the heritability of complex traits. Nat Genet. Nature Publishing Group; 2014 Dec 1;46(12):1356–62.

23. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jun 20;42(7):565–9.

24. Huang YZ, Won S, Ali DW, Wang Q, Tanowitz M, Du QS, et al. Regulation of Neuregulin Signaling by PSD-95 Interacting with ErbB4 at CNS Synapses. Neuron. Elsevier; 2000 Jan 5;26(2):443–55.

25. Georgieva L, Moskvina V, Peirce T, Norton N, Bray NJ, Jones L, et al. Convergent evidence that oligodendrocyte lineage transcription factor 2 (OLIG2) and interacting genes influence susceptibility to schizophrenia. PNAS. National Acad Sciences; 2006 Aug 15;103(33):12469–74.

26. Wang Y, Liu X, Robbins K, Rekaya R. AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. BMC Research Notes. BioMed Central Ltd; 2010 Apr 28;3(1):117.

27. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nature Reviews Genetics. Nature Publishing Group; 2013 Mar 1;14(3):204–20.

28. Artigas MS, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet. Nature Publishing Group; 2011 Nov 1;43(11):1082–90.

29. Liao SY, Lin X, Christiani DC. Gene-environment interaction effects on lung function-a genome-wide association study within the Framingham heart study. Environ Health. 2013.

30. Anderson LN, Briollais L, Atkinson HC, Marsh JA, Xu J, Connor KL, et al. Investigation of Genetic Variants, Birthweight and Hypothalamic-Pituitary-Adrenal Axis Function Suggests a Genetic Variant in the SERPINA6 Gene Is Associated with Corticosteroid Binding Globulin in the Western Australia Pregnancy Cohort (Raine) Study. Hsu Y-H, editor. PLoS ONE. Public Library of Science; 2014 Apr 1;9(4):e92957.

31. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics. Oxford University Press; 2012 May 15;28(10):1359–67.

32. Sankararaman S, Sridhar S, Kimmel G. Estimating local ancestry in admixed populations. The American Journal of …. 2008.

33. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, et al. A Genomewide Admixture Map for Latino Populations. The American Journal of Human Genetics. 2007 Jun;80(6):1024–36.

34. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, et al. A High-Density Admixture Map for Disease Gene Discovery in African Americans. The American Journal of Human Genetics. 2004 May;74(5):1001–13.

35. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Human Identification: The Use of DNA Markers. 1995.

36. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC

Bioinformatics. BioMed Central Ltd; 2010 Nov 30;11(1):587.

Figure Legends

Figure 1. Power Plots for Pairwise Interaction Simulations.
Power of testing $G \times \theta$ (a) versus testing pairwise SNPs directly (b) as a function of the difference in the ancestral allele frequencies at a differentiated SNP.

Figure 2. Power Plots for $G \times E$ Interaction Simulations.
Power of testing $G \times \theta$ as a function of the correlation between an environmental covariate and genetic ancestry.

Figure 3. Examples of How Genetic Ancestry Can Be A Proxy for Interacting Covariates.
(a) Model of how genetic ancestry θ can be correlated with various environmental exposures, some of which affect a phenotype. (b) Example of how the correlation between the probability of an AA genotype (bars 2-4) and values of θ (bar 1) increase with higher levels of SNP allele frequency differentiation. In this plot p_1 and p_2 denote the allele frequency of allele A in ancestral populations 1 and 2 respectively. (c) Example of how effect sizes at a tag-SNP may differ due to differential LD on distinct ancestral backgrounds (here, EUR and AFR).

Tables

Table 1. Distribution of Ancestry in Coriell and GALA II.

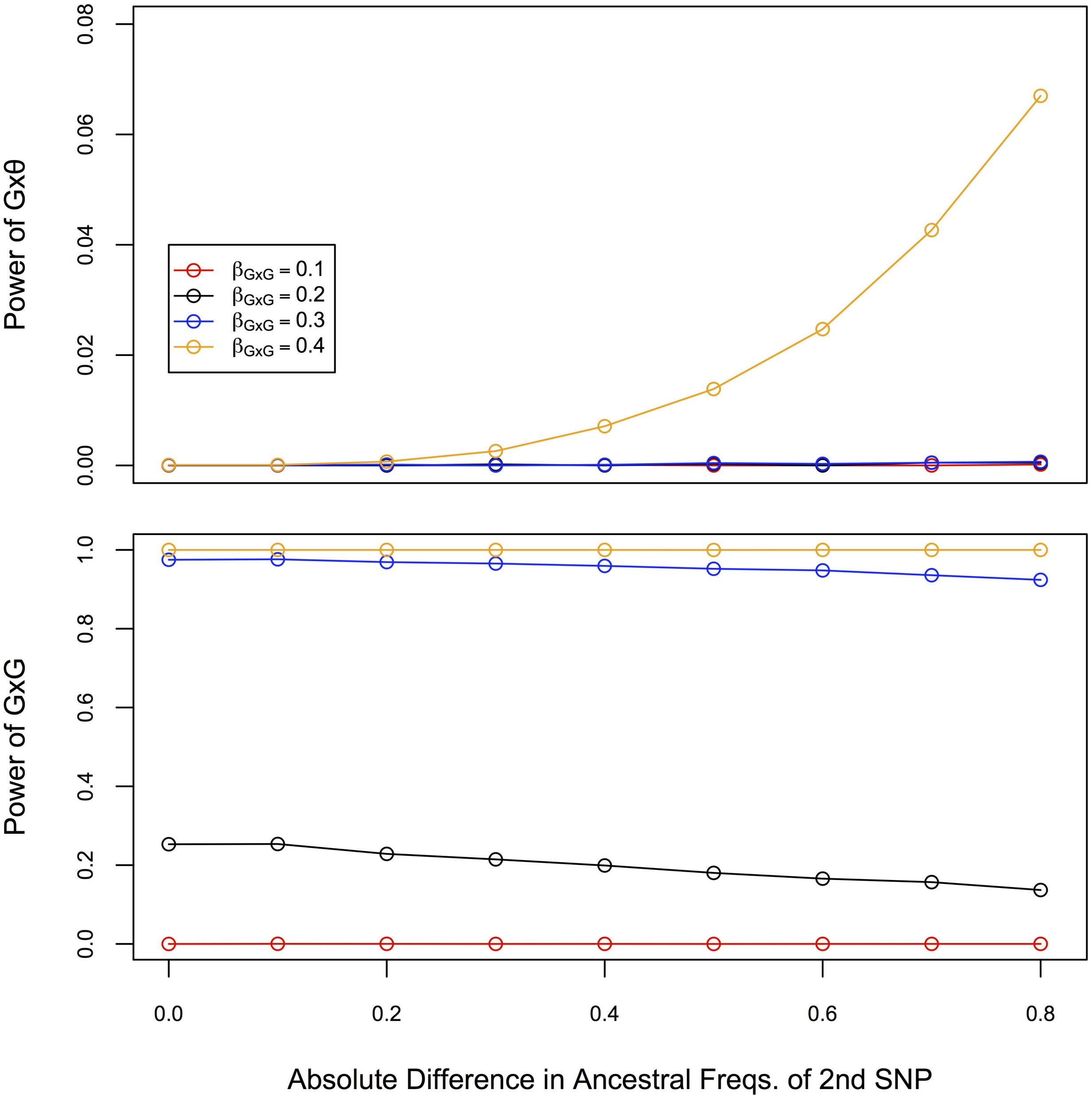
Dataset	θ_{EUR}	θ_{AFR}	θ_{NAM}
Coriell	μ=0.212, σ=0.021	μ=0.788, σ=0.021	NA
GALA II MX	μ=0.396, σ=0.022	μ=0.043, σ=0.001	μ=0.561, σ=0.025
GALA II PR	μ=0.464, σ=0.008	μ=0.241, σ=0.009	μ=0.113 σ=0.001

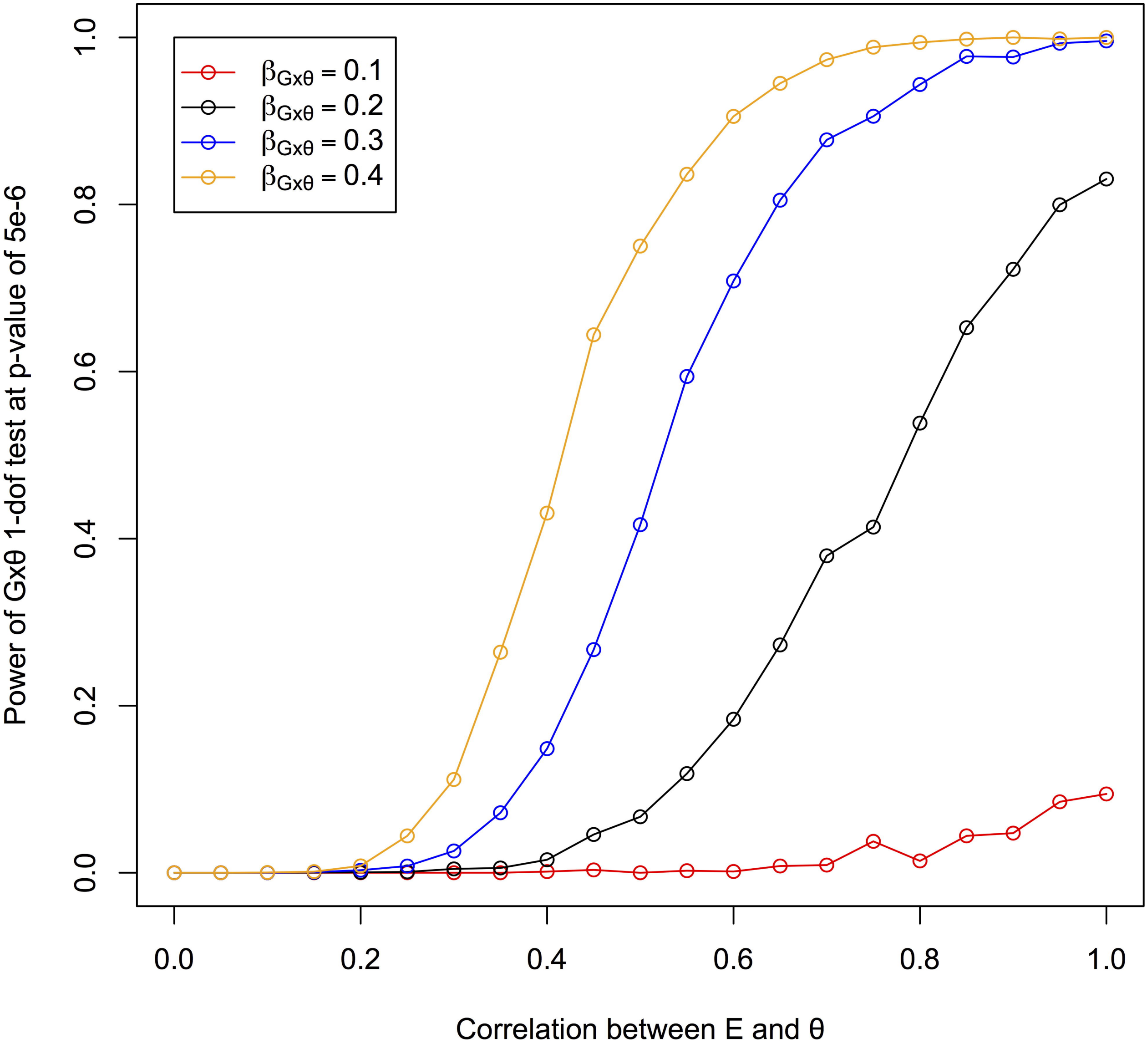
Mean and variance of the global ancestry distributions for each dataset.

Table 2. GALA II DNA Methylation Analysis Results.

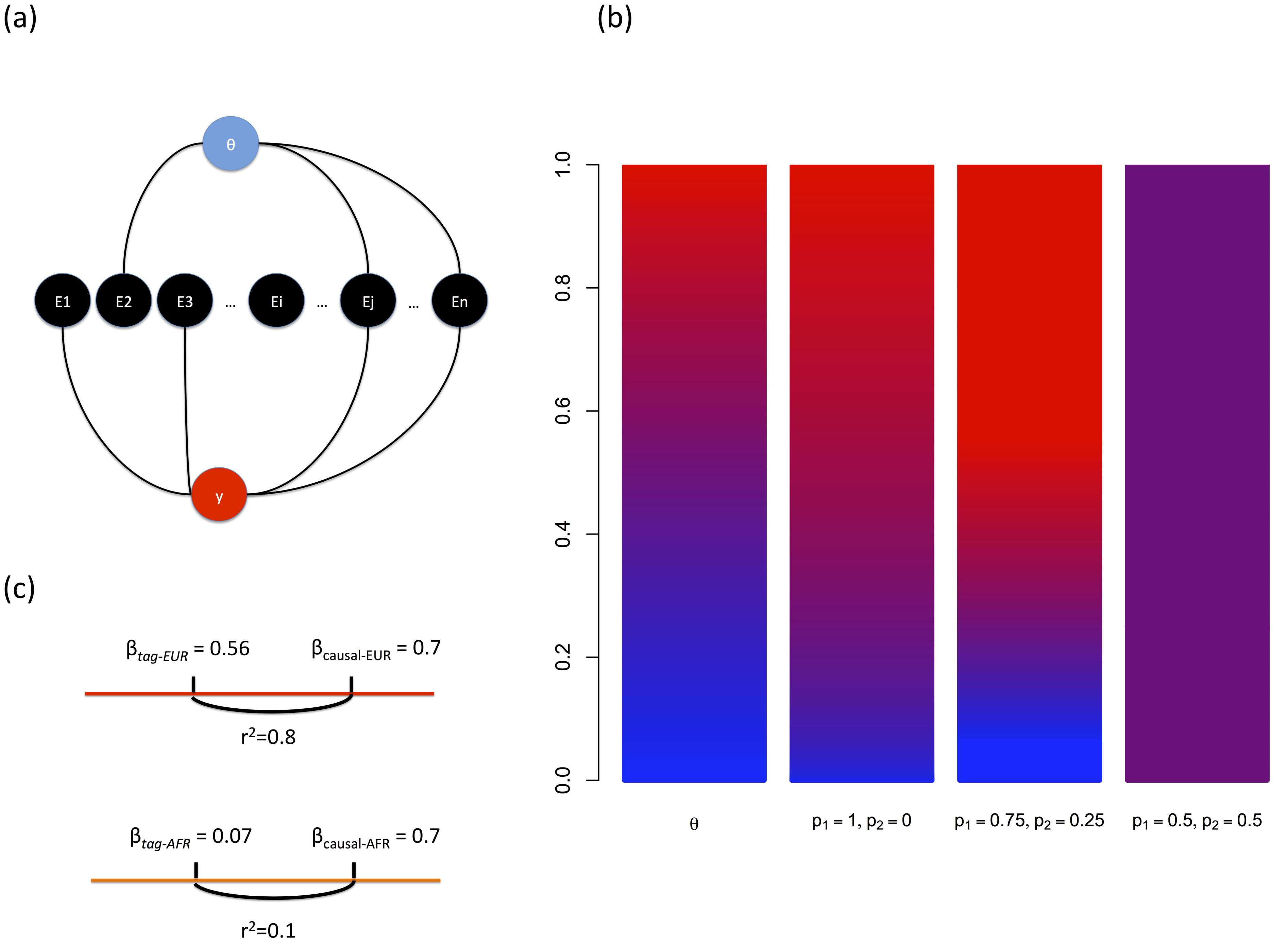
GALA II Population	Probe Gene	Probe ID	rsid	Distance of SNP to Probe	Marginal p-value	AITL p-value	AITL Replication p-value
MX	CNFN	cg14327995	rs16975986	280795	2.49E-09	5.69E-09	9.27E-03
MX	C11orf95	cg16678159	rs7106153	249768	2.58E-01	2.52E-08	9.39E-02
MX	NA	cg05697734	rs1560919	13711	1.14E-01	2.21E-08	8.18E-03
MX	TNK2	cg01792640	rs67217828	278866	4.49E-01	6.38E-09	1.43E-02
MX	HDAC4	cg06533788	rs925736	9548	4.51E-01	3.09E-09	2.80E-02
PR	NA	cg07436864*	rs8117083	31813	7.46E-02	1.34E-09	5.34E-03
PR	NA	cg16803083*	rs4312379	63847	3.69E-01	2.29E-08	2.31E-04
PR	SERPINA6	cg10025865	rs17091085	247796	6.83E-01	2.97E-08	8.05E-03

P-values for AITL applied to the methylation data in the GALA II Latinos. MX and PR denote Mexicans and Puerto Ricans respectively in the GALA II population columns. The probe gene column shows the gene that the methylation probe lies in. The marginal column is the p-value for standard linear regression of methylation on genotype while controlling for population structure. * indicates results that replicated between the Mexicans and Puerto Ricans.


Power of Gxθ
0.00
0.02
0.04
0.06
0.08
β_{GxG} = 0.1
β_{GxG} = 0.2
β_{GxG} = 0.3
β_{GxG} = 0.4
Power of GxG
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Absolute Difference in Ancestral Freqs. of 2nd SNP



$\beta_{Gx\theta} = 0.1$
$\beta_{Gx\theta} = 0.2$
$\beta_{Gx\theta} = 0.3$
$\beta_{Gx\theta} = 0.4$
Power of Gxθ 1-dof test at p-value of 5e-6
0.0
0.2
0.4
0.6
0.8
1.0
Correlation between E and θ
0.0
0.2
0.4
0.6
0.8
1.0


(a)
θ
E1
E2
E3
...
Ei
...
Ej
...
En
y
(b)
1.0
0.8
0.6
0.4
0.2
0.0
θ
p₁ = 1, p₂ = 0
p₁ = 0.75, p₂ = 0.25
p₁ = 0.5, p₂ = 0.5
(c)
βtag-EUR = 0.56
βcausal-EUR = 0.7
r²=0.8
βtag-AFR = 0.07
βcausal-AFR = 0.7
r²=0.1