Title

An ancestry-based approach for detecting interactions

Permalink

https://escholarship.org/uc/item/7kc0k081

Journal

Genetic Epidemiology, 42(1)

ISSN

0741-0395

Authors

Park, Danny S
Eskin, Itamar
Kang, Eun Yong
et al.

Publication Date

2018-02-01

DOI

10.1002/gepi.22087

Peer reviewed

Published in final edited form as:
Genet Epidemiol. 2018 February ; 42(1): 49–63. doi:10.1002/gepi.22087.

An Ancestry Based Approach for Detecting Interactions

Danny S. Park[1,*], **Itamar Eskin**[2], **Eun Yong Kang**[3], **Eric R. Gamazon**[4,5], **Celeste Eng**[6], **Christopher R. Gignoux**[1,7], **Joshua M. Galanter**[6], **Esteban Burchard**[1,6], **Chun J. Ye**[8], **Hugues Aschard**[9], **Eleazar Eskin**[3], **Eran Halperin**[2], and **Noah Zaitlen**[1,6,*]

[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]Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands [6]Department of Medicine. University of California San Francisco. San Francisco, CA [7]Department of Genetics. Stanford University. Palo Alto, CA [8]Institute of Human Genetics. University of California San Francisco. San Francisco, CA [9]Department of Epidemiology. Harvard School of Public Health. Boston, MA

Abstract

Background: Epistasis and gene-environment interactions are known to contribute significantly to variation of complex phenotypes in model organisms. However, their identification in human association studies remains challenging for myriad reasons. In the case of epistatic interactions, the large number of potential interacting sets of genes 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, defined as the proportion of ancestry derived from each ancestral population (e.g. the fraction of European/African ancestry in African Americans), in admixed populations. We applied our method to gene expression and methylation data from African American and Latino admixed individuals respectively, identifying nine interactions that were significant at $p < 5 \times$

[*]Corresponding Author sihyung@gmail.com, noah.zaitlen@ucsf.edu.

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, ERG, 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.

10^{-8}. We show that two of the interactions in methylation data replicate, and the remaining six are significantly enriched for low p-values ($p < 1.8 \times 10^{-6}$).

Conclusion: We show that genetic ancestry can be a useful proxy for unknown and unmeasured covariates in the search for interaction effects. These results have important implications for our understanding of the genetic architecture of complex traits.

Keywords

Gene-environment interaction; gene-gene interactions; admixture

II. Background

Genetic association studies in humans have focused primarily on the identification of additive single nucleotide polymorphism (SNP) effects through marginal tests of association. There is growing evidence that both epistatic and gene-environment ($G \times E$) interactions contribute significantly to phenotypic variation in humans and model organisms (Hemani et al., 2014; Jemal et al., 2011; Kang et al., 2014; Rouhani et al., 2014). 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 computational and multiple testing burden in the case of epistatic interactions, and the lack of consistently measured environmental covariates in the case of $G \times E$ interactions (Eichler et al., 2010; Manolio et al., 2009).

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 (Seldin, Pasaniuc, & Price, 2011) and for each admixed individual it is possible to accurately estimate genetic ancestry, the proportion of ancestry derived from each ancestral population (e.g. the fraction of European/African ancestry in African Americans) and commonly denoted as θ (Alexander, Novembre, & Lange, 2009). Ancestry has been previously leveraged to demonstrate that an array of environmental and biomedical covariates are correlated with θ (Burchard et al., 2003; Cheng et al., 2012; Choudhry et al., 2006; Florez et al., 2011; Kumar et al., 2013; 2010; Price et al., 2008; Reiner et al., 2007; Sanchez et al., 2010; Shaffer et al., 2007; Ziv et al., 2006) 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 differentiated between the ancestral populations, suggesting that θ may be effectively used as a proxy for detecting multi-way epistatic interactions. Therefore, we propose a new SNP by θ test of interaction in order to detect evidence of interactions in admixed populations. Detecting SNP by θ interactions promises to yield insights into the genetic architecture of disease risk and other complex traits.

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 our approach, the Ancestry Test of Interaction with Local Ancestry (AITL), is well-powered to detect $G \times E$ interactions when θ is correlated with the environmental covariates of interest and multi-way epistatic interactions. 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. However, the results of our simulations show that AITL is well powered to detect multi-way epistasis involving tens or hundreds of SNPs of small effects, not detectable by pairwise tests.

We first examined molecular phenotypes by applying our method to gene expression data from African Americans, as well as DNA methylation data from Latinos. Gene expression traits have previously been shown to have large-scale differences as a function of genetic ancestry (Price et al., 2008). Other molecular phenotypes, such as LDL levels, have also been shown to be associated with genetic ancestry (Fraser, Lam, Neumann, & Kobor, 2012; Galanter et al., 2017; Peralta et al., 2010; Price et al., 2008; Reiner et al., 2007; Spielman et al., 2007). For gene expression in particular, Price et al. (2008) showed that the effects of ancestry on expression are widespread and not restricted to a handful of genes. Additionally, molecular phenotypes are often used in deep phenotyping and Mendelian randomization studies and are thus directly relevant to elucidating disease biology(Delude, 2015; Vimaleswaran et al., 2013).

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. Two of the eight interactions associated with DNA methylation in the Latinos also replicated and the remaining six were enriched for low p-values ($p < 1.8 \times 10^{-6}$). To demonstrate that our approach works in larger data sets we also applied AITL to asthma case-control data from Latinos and observed well-calibrated test statistics. Together, these results provide evidence for the existence of interactions regulating expression and methylation and show that our approach is statistically sound.

III. Materials and Methods

Our approach is best illustrated with an example. First consider testing a SNP 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 1a). Now consider testing a SNP 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 (percentage of ancestry from population 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 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 1b).

Consider an admixed individual i who derives his or her genome from k ancestral populations. We denote individual i's global ancestry proportion as $\langle \theta_{i1}, \theta_{i2}, \ldots, \theta_{ik} \rangle$, where $\sum_{j=1}^{k} \theta_{ij} = 1$. The local ancestry of individual i at a SNP is denoted as $\gamma_{ia} \in \{0,1,2\}$ and is equal to the number of alleles from ancestry $a \in \{1 \ldots k\}$ inherited at that SNP. Current methods allow us to estimate ancestry directly from genotype data both globally and at specific SNPs (Alexander et al., 2009; Baran et al., 2012). We denote the genotype of an individual i at a given SNP as $g_i \in \{0,1,2\}$ and the corresponding phenotype as y_i.

In this work, we model continuous phenotypes in an additive linear regression framework. 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 $n \times v$ matrix of v covariates, and $\vec{\beta}$ is a $v \times 1$ 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:

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

Then the log likelihood is given by the following expression:

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

Let L_1 denote the likelihood under the alternative and L_0 the likelihood under the null. We can compute the log-likelihood ratio statistic (D) using a maximum likelihood approach:

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

where $\widehat{\vec{\beta}}_{L_1}$ and $\hat{\sigma}_{L_1}$ are the maximum likelihood estimators of $\vec{\beta}$ and $\boldsymbol{\sigma}$ under the alternative, respectively and $\widehat{\vec{\beta}}_{L_0}$ and $\hat{\sigma}_{L_0}$ are the maximum likelihood estimators (MLEs) of $\vec{\beta}$ and $\boldsymbol{\sigma}$ under the null, respectively.

We note that for a case-control phenotype we would use the following likelihood and log-likelihood ratio statistic, assuming a logistic regression model:

$$L = \prod_{i=1}^{n} \left[\frac{1}{1 + e^{-X_i \overrightarrow{\beta}}} \right]^{y_i} \left[1 - \frac{1}{1 + e^{-X_i \overrightarrow{\beta}}} \right]^{1 - y_i}$$

$$D = -2\left(\log L_1 - \log L_0\right) = -2\left(\sum_{i=1}^{n} -\log\left(1 + e^{-X_i \widehat{\overrightarrow{\beta}} L1}\right) + \sum_{i=1}^{n} \left(1 - y_i\right)\left(- X_i \widehat{\overrightarrow{\beta}}_{L_1}\right)\right) + 2 \left(\sum_{i=1}^{n} -\log\left(1 + e^{-X_i \widehat{\overrightarrow{\beta}} L0}\right) + \sum_{i=1}^{n} \left(1 - y_i\right)\left(-X_i \widehat{\overrightarrow{\beta}}_{L_0}\right)\right)$$

where X_i is the i-th row of the matrix $\boldsymbol{X}$, which correspond to the covariates of individual i and $y_i \in \{0,1\}$ is the phenotype of individual i.

For linear regression, the MLE of the effect sizes is $\widehat{\overrightarrow{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \overrightarrow{y}$, and the MLE of the error variance is $\widehat{\sigma}^2 = \frac{1}{n}\left(\overrightarrow{y} - \boldsymbol{X}\widehat{\overrightarrow{\beta}}\right)^2$. $\left(\widehat{\overrightarrow{\beta}}_{L_1}, \widehat{\sigma}^2_{L_1}\right)$ and $\left(\widehat{\overrightarrow{\beta}}_{L_0}, \widehat{\sigma}^2_{L_0}\right)$ 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 $\overrightarrow{g} = \left\langle g_1, \ldots, g_n \right\rangle$ be the vector of the individuals' genotypes at a given SNP, $\overrightarrow{\theta}_a = \left\langle \theta_{1a}, \ldots, \theta_{na} \right\rangle$ be the vector of their global ancestries for ancestry a, and $\overrightarrow{g} \times \overrightarrow{\theta}_a$ 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 $\left(\widehat{\beta}_{G \times \theta} \neq 0\right)$ against the null hypothesis $\left(\widehat{\beta}_{G \times \theta} = 0\right)$.

$$H_1 : \overrightarrow{y} = \widehat{\beta}_g \overrightarrow{g} + \widehat{\beta}_{G \times \theta}\left(\overrightarrow{g} \times \overrightarrow{\theta}_a\right) + \widehat{\beta}_\theta \overrightarrow{\theta}_a$$

$$H_0 : \overrightarrow{y} = \widehat{\beta}_g \overrightarrow{g} + \widehat{\beta}_\theta \overrightarrow{\theta}_a$$

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.

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 1c). 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 (Price et al., 2007; M. W. Smith et al., 2004).

Let $\vec{\gamma}_a = \langle \gamma_{1a}, \ldots, \gamma_{na} \rangle$ be the vector of local ancestry calls for all individuals for ancestry *a* and let $\vec{g} \times \vec{\gamma}_a$ be the interaction terms from piecewise multiplication of the two vectors. We use the following alternative and null hypotheses:

$$H_1 : \vec{y} = \hat{\beta}_g \vec{g} + \hat{\beta}_{G \times \theta} \left(\vec{g} \times \vec{\theta}_a \right) + \hat{\beta}_\theta \vec{\theta}_a + \hat{\beta}_\gamma \vec{\gamma}_a + \hat{\beta}_{G \times \gamma} \left(\vec{g} \times \vec{\gamma}_a \right)$$

$$H_0 : \vec{y} = \hat{\beta}_g \vec{g} + \hat{\beta}_\theta \vec{\theta}_a + \hat{\beta}_\gamma \vec{\gamma}_a + \hat{\beta}_{G \times \gamma} \left(\vec{g} \times \vec{\gamma}_a \right)$$

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*. All of these test statistics are straightforwardly modified to jointly incorporate several ancestries in the case of multi-way admixed populations. This can be done by adding in global and local ancestry terms for ancestries that are not collinear. For example, in Hispanic Latinos, this would be the European and Native American components of ancestry. Similarly, for K-way admixed populations, global and local ancestry components may be added into the analysis as long as ancestries included are not highly collinear.

Standard Pairwise Test of Interaction and Controlling Confounding in Admixed Populations

Here we present the standard approach for testing for interaction between two SNPs *s* and *j*. We use the following alternative and null hypotheses.

$$H_1 : \vec{y} = \hat{\beta}_s \vec{g}_s + \hat{\beta}_j \vec{g}_j + \hat{\beta}_{s \times j} \left(\vec{g}_s \times \vec{g}_j \right) + \hat{\beta}_\theta \vec{\theta}_a$$

$$H_0 : \vec{y} = \hat{\beta}_s \vec{g}_s + \hat{\beta}_j \vec{g}_j + \hat{\beta}_\theta \vec{\theta}_a$$

If AITL is significant for a given SNP *s*, then any SNP *j* tested for interaction with *s* may be biased if *j* is correlated with covariates that are also correlated with θ. We thus, propose the following alternative and null hypotheses:

$$H_1: \vec{y} = \hat{\beta}_s \vec{g}_s + \hat{\beta}_j \vec{g}_j + \hat{\beta}_{s \times j}\left(\vec{g}_s \times \vec{g}_j\right) + \hat{\beta}_\theta \vec{\theta}_a + \hat{\beta}_{s \times \theta}\left(\vec{g}_s \times \vec{\theta}_a\right)$$

$$H_0: \vec{y} = \hat{\beta}_s \vec{g}_s + \hat{\beta}_j \vec{g}_j + \hat{\beta}_\theta \vec{\theta}_a + \hat{\beta}_{s \times \theta}\left(\vec{g}_s \times \vec{\theta}_a\right)$$

We note that the utility of this test will require further investigation (see Discussion).

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 *i* with $\theta_{i1} > 1$ or $\theta_{i1} < 0$ were assigned a value of 1 or 0, respectively. We simulated phenotypes of individuals to investigate our method in four different scenarios: $G \times E$ interactions, pairwise epistatic interactions, multi-way epistatic interactions, and false positive interactions due to local differential tagging.

To simulate phenotypes under the situation of a $G \times E$ 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_{i1}) 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 population frequencies, p_1 and p_2, from the following beta distribution (Balding & Nichols, 1995).

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

where *p* is the ancestral population allele frequency 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 $\gamma_{i1} = 0$ or 1. Environmental covariates correlated with the proportion of ancestry from population 1, E_i, were generated for each individual *i* by drawing from a normal distribution $\mathcal{N}\left(\mu = \theta_{i1}, \sigma_E\right)$, where σ_E is the standard deviation of the environmental covariates. 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, $\mathcal{N}\left(\mu = \beta_{G \times E} \times g_{i1} \times E_i, \sigma = 1\right)$ for a given interaction effect size ($\beta_{G \times E}$). We show the relationship of this simulation framework with modeling environment as a linear function of ancestry (see *Environment as a Linear Function of Genetic Ancestry)*. In addition, we simulated the case in which *E* is a smooth (C^∞) function of genetic ancestry so that E can be approximated using a Taylor series expansion (see *Environment as a Smooth Function of θ)*.

To simulate phenotypes based on pairwise epistatic interactions, we simulated two SNPs. At both SNPs, we assigned the local ancestry values as described for the $G \times E$ 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 $p_1 = 0.95$ and $p_2 = 0.05$. Genotypes at the second SNP were drawn using the same approach described for $G \times E$. Using the simulated genotypes, phenotypes were drawn from a normal distribution, $\mathcal{N}\left(\mu = \beta_{G \times G} \times g_{i1} \times g_{i2}, \sigma = 1\right)$, where g_{is} is the genotype for individual i at the simulated SNP.

To simulate phenotypes based on multi-way epistatic interactions, we simulated a SNP s and m (independent) SNPs with pairwise interactions with s. Genotypes for individuals at SNP s were assigned assuming an allele frequency of 0.5 for both populations and drawing from two binomial trials. Genotypes at the m interacting SNPs were assigned in the same manner as the second SNP in the pairwise interaction simulations. Using the simulated genotypes, phenotypes were drawn from a normal distribution, $\mathcal{N}\left(\mu = \sum_{j=1}^{m} \beta_{s \times j}\left(g_{is} \times g_{ij}\right), \sigma = 1\right)$ where g_{ix} is the genotype for individual i at the simulated SNP x.

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, $\mathcal{N}\left(\mu = \beta_{Causal} \times g_{ic}, \sigma = 1\right)$, assuming no interaction and $\beta_{Causal} = 0.7$, where g_{ic} is the genotype of individual i at the causal variant c.

Environment as a Linear Function of Genetic Ancestry—Assume environment is a linear function of genetic ancestry, i.e $E = \alpha\theta + \delta$, where $\delta \sim \mathcal{N}\left(\mu_\delta, \sigma_\delta\right)$ and $\theta \sim Truncated\ \mathcal{N}\left(\mu_{\theta TR}, \sigma_{\theta TR}, a = 0, b = 1\right)$. This means $\theta \sim \mathcal{N}\left(\mu_\theta, \sigma_\theta\right)$, under the constraint that $0 < \theta < 1$. The assumption of a truncated normal distribution for θ is reasonable; indeed, we find no significant difference with the empirical distribution in our data (p=0.701 for GALA II Mexicans and p=0.216 for GALA II Puerto Ricans, Kolmogorov-Smirnov Test). Let $\pi_1 = \frac{a - \mu_\theta}{\sigma_\theta}$, $\pi_2 = \frac{b - \mu_\theta}{\sigma_\theta}$, $Z = \Phi(\pi_2) - \Phi(\pi_1)$, ϕ the probability density function of the standard normal, and Φ the cumulative distribution function of the standard normal. Then θ

will have mean $\mu_{\theta TR} = \mu_\theta + \frac{\phi(\pi_1) - \phi(\pi_2)}{Z}\sigma_\theta$ and variance $\sigma^2_{\theta TR} = \sigma^2_\theta\left[1 + \frac{\pi_1\phi(\pi_1) - \pi_2\phi(\pi_2)}{Z} - \left(\frac{\pi_1\phi(\pi_1) - \pi_2\phi(\pi_2)}{Z}\right)^2\right]$.

This implies E is random variable with mean given by E[E] = E[$\alpha\theta + \delta$] = αE[θ] + E[δ] = αE[θ] = $\alpha\mu_{\theta TR}$, and variance given by $\mathrm{var}[E] = \alpha^2\mathrm{var}[\theta] + \mathrm{var}[\delta] = \alpha^2\sigma^2_{\theta TR} + \sigma^2_\delta$. Thus our simulation framework captures the case where E is normally distributed with mean and variance given above.

Environment as a Smooth Function of θ—More generally, if the environment E is a smooth function (thus, also continuous) of θ, then E can be written as a Taylor series expansion[31]. Let $r \in \mathbb{Z}$ and $\alpha_r \in \mathbb{R}$. If environment is a function of ancestry, i.e.

$$E = f\left(\theta\right) = \alpha_0 + \sum_{r=1}^{\infty}\alpha_r\left(\theta - \mu_\theta\right)^r$$

Then the correlation between E and θ is given by

$$cor(E,\ \theta) = \ cor\left(\theta,\ \alpha_0\right) + \sum_{r=1}^{\infty}\alpha_r cor\left(\theta,\ \left(\theta - \mu_\theta\right)^r\right)$$

We show empirically that as $r \rightarrow \infty$, $cor(\theta,(\theta - \mu_\theta)^r) \rightarrow 0$ (see Supplementary Figure S6). This implies that $cor(E,\theta)$ is dominated by the first few terms of the summation.

Thus, we simulated an environmental factor of the form

$$E = f\left(\theta\right) = \alpha_0 + \sum_{r=1}^{2}\alpha_r\left(\theta - \mu_\theta\right)^r$$

where $\alpha_0 \sim \mathcal{N}(\mu = 0, \sigma)$. We varied σ from 0 to 1 in increments of 0.1 to allow for the correlation between E and θ to vary from 0 to 1. α_1 was set to 1 (which corresponds to the first derivative f'(μ_θ) being positive, indicating that the environment increases with ancestry at the mean) and α_2 was set to 0.1 (which corresponds to the second derivative f''(μ_θ) being positive, indicating the environment is concave upward at the mean) for all simulations. Note that although we made specific choices of these coefficients, this provides a general framework for modeling environment as a smooth function of ancestry.

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

Ancestry Inference

Global ancestry inference was done using ADMIXTURE (Alexander et al., 2009) and local ancestry inference was done using LAMP-LD (Baran et al., 2012). CEU and YRI from 1000 Genomes Phase 3 (McVean et al., 2012) were used as the European and African reference panels. For the Native American reference panels, 95 Native Americans genotyped on the Axiom LAT1 array were used (Drake et al., 2014).

Filtering for Related Individuals

All analyses in real data were filtered for related individuals due to the possibility of cryptic relatedness causing false positives. To filter for related individuals, we estimated kinship coefficients between all pairs of individuals using REAP (Thornton et al., 2012). We defined two individuals as related if they had a kinship coefficient greater than or equal to 0.025. For a pair of related individuals, we removed the one with a greater number of other individuals to whom he or she was related. In the case of a tie, we removed one of the pair at random.

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 (Du et al., 2010). 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.

IV. 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 have smaller effects in terms of the phenotypic variance they explain, 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 (Grundberg et al., 2012). However for most phenotypes, known interactions will explain a very small proportion of the phenotypic variance, mainly due to the fact that so few interactions have been identified and replicated (Aschard, Gusev, Brown, & Pasaniuc, 2015).

Power When Using θ as a Proxy for Highly Differentiated SNPs—To determine whether using θ as a proxy for 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 $\vec{\theta}_1$. 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 2a). 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 are available (Aschard et al., 2015). For the larger effect sizes we simulated, we do see power increasing as the difference between ancestral frequencies increases. The plots show that AITL has little power unless the effect was very strong. Figure 2b 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 since differentiated SNPs across the genome will be correlated with genetic ancestry. These simulations are important as other studies have suggested that higher order interactions may be important for some traits (De, Hu, Moore, & Gilbert-Diamond, 2015; Hemani et al., 2014; Ritchie et al., 2001). To evaluate the ability of θ to serve as a proxy for multiple (independent) differentiated SNPs, we simulated a scenario where a candidate SNP z had interactions with m SNPs (see Simulation Framework). For each interaction, we assumed a small interaction effect size ($\beta_{G\times G} = 0.025$), which would not be detectable using a pairwise approach, as we demonstrated in the pairwise simulation. Figure 3 shows that AITL is better powered to detect the existence of interactions than a pairwise approach in the presence of multiple interacting SNPs with a candidate SNP.

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 global ancestries in varying degrees (see Simulation Framework). Our simulation framework is similar to modeling E as a linear function of θ. Figure 4 shows the power of the AITL as a function of the Pearson correlation between $\overrightarrow{\theta}_1$ 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 $G \times E$ interaction is low whether one uses θ_1 or E. However, both tests are much better powered for effect sizes greater than or equal to 0.2, with the AITL's power being dependent on the level of correlation. We saw similar results when we simulated the case where E is a smooth function of θ (see *Environment as a Smooth Function of* θ and Supplementary Figure S7). Note that using θ as a proxy for E is equivalent to testing GxE in the presence of measurement error. Under the assumption of non-differential error with regard to the outcome (e.g. the correlation between θ and E is equal among cases and control) such a test is underpowered but has a controlled type I error rate under the null (Wong, Day, Luan, Chan, & Wareham, 2003).

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 genomic control λ_{GC}, the ratio of the observed median χ^2 over the expected median χ^2 under the null (Devlin & Roeder, 2004). 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 (Simon-Sanchez et al., 2006). 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 ($p \leq 5 \times 10^{-8}$). Here the marginal p-value is derived from standard linear regression of expression on genotype while controlling for global population structure. rs7585465 has a ‘C’ allele frequency of 0.218 in the Corriell data and appears to be differentiated between CEU and YRI with allele frequencies of 0.619 and 0.097 in the respective populations.

Given that the gene expression values come from LCLs (all cultured according to the same standards), the SNPs may be interacting with epigenetic alterations due to environmental exposures that have persisted since transformation into LCLs. This scenario is unlikely, and we believe that signals are driven by multi-way 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 many 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(Martin et al., 2014; Rouhani et al., 2014).

Although we believe the ERBB4 result to be representative of multi-way epistasis, we performed a standard pairwise interaction test (see Methods) to check for interaction between rs7585465 and other SNPs genome-wide. Interestingly, we found that the standard interaction test (see Methods) showed substantial departure from the null with a λ_{GC} equal to 1.8 (see Supplementary Figure S3). Since the interaction of rs7585465 by θ was significant, the pairwise interaction test statistics of rs7585465 by any SNP j can be inflated if j is correlated with θ. We found that including the original significant SNP by θ term in the null (see Methods) brought the λ_{GC} down to 1.05, and controlled for such scenarios in this dataset (See Supplementary Figure S3). As we had previously anticipated, identifying the exact interactions driving the SNP by θ interaction proved to be difficult. We found one borderline significant SNP (rs4839709, $p = 3.08 \times 10^{-7}$) but no interactions that passed genome-wide significance. These results are consistent with what we have observed in simulations, in which even though a standard pairwise interaction test is underpowered to detect interactions, AITL is able to identify the main locus involved in a multi-way interaction.

GALA II Case-Control—To determine if our method is biased in large structured genome-wide association study (GWAS) data, we applied AITL to case-control data from a study of asthmatic Latino individuals called the Genes-environments and Admixture in Latino Americans (GALA II) (Borrell et al., 2013). The dataset includes 1158 Mexicans and 1605 Puerto Ricans, which were analyzed separately. Case status was assigned to individuals if they were between the ages of 8 and 40 years with a physician-diagnosed asthma. Additionally, they had to have experienced 2 or more asthma related symptoms in the previous 2 years at the time of recruitment (Torgerson et al., 2012). In the Mexicans and Puerto Ricans there were 548 and 797 cases, respectively. In our analysis, we also included BMI, age, and sex as additional covariates. We observed well-calibrated statistics with a λ_{GC} equal to 1.00 and 0.98 in the Mexicans and Puerto Ricans, respectively (see Supplementary Figure S4). In contrast to the molecular phenotype data, searches for interactions in these phenotypes did not yield any findings passing genome-wide significance. This is consistent with previous disease studies that have failed to find many replicable interactions in disease studies (Aschard et al., 2015). In the data here, the lack of any findings may be due to the relatively small sample size or because the effects of the interactions are extremely small (if they exist for covariates correlated with θ_{EUR}).

GALA II Methylation Results—We searched for interactions in methylation data derived from a study of GALA II asthmatic Latino individuals (Borrell et al., 2013). 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 S5). We tested 128,794,325 methylation-SNP pairs, which result in a Bonferroni corrected p-value cutoff of 3.88×10^{-10}. However, this cutoff is extremely conservative given the tests are not independent. We therefore 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).

As we did for the gene expression data, we attempted to identify pairwise interactions involved in the methylation data results. For each genome-wide significant result, we performed a standard pairwise interaction test of all SNPs with the original SNP found to be significant with AITL. We were unable to identify any significant interactions after applying genomic control to the results. For all tests, we included the significant SNP by θ term (see Methods) in the null. For this dataset, unlike the gene expression data, we observed substantial remaining departure from the null (see Supplementary Table S2) even after including the original significant SNP by θ term, suggesting there may be other factors that need to be accounted for when testing for interactions in admixed populations. The results from our pairwise scan are what we would anticipate, given that in simulations only AITL (not the standard pairwise interaction test) was able to identify the main locus involved in the multi-way interaction.

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 = 6.25×10^{-3}. Two of the SNPs that AITL originally identified to have significant interaction with ancestry in Puerto Ricans, rs4312379 and rs8117083, replicated in the Mexicans (Table 2). They were also still highly significant after adjusting for a second component of ancestry as well (Table S1). Furthermore, there was a highly significant enrichment of low p-values in the replication study among the discovery results (permutation $p < 1 \times 10^{-4}$). Furthermore, 5 out of the 6 non-replicating results have a p-value less than 0.05 (binomial test $p < 1.8 \times 10^{-6}$). The results of the permutation and binomial test suggests that the interactions that did not replicate are likely to do so with bigger sample sizes. It is important to note that replicated interactions and the enrichment for low p-values do not necessarily indicate that the same genetic or environmental covariates are interacting with the genetic locus in both populations. The covariates correlated with θ_{EUR} in one population are not necessarily those correlated with θ_{EUR} in the other population. There may be correlations which exist in both populations but θ_{EUR} serves as a proxy for all such correlated covariates and therefore should not be necessarily viewed as a proxy for any specific one. Overall, our results from the GALA II (methylation) cohort suggest there are both genetic and environmental variables contributing to epistasis that have yet to be discovered in admixed individuals.

V. Discussion and Conclusions

For many disease architectures, interactions are believed to be a major component of missing heritability (Eichler et al., 2010). 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 $G \times \theta$ interactions can be leveraged to overcome some of the difficulties typically encountered when searching for interactions. The computational cost is minimal and has the same order of magnitude as running a standard GWAS.

One drawback of our method is that it does not identify which covariate is interacting with a genetic locus. Nevertheless, the approach can show whether an interaction effect exists in a given dataset and if it does exist, our method ensures that an underlying genetic or environmental covariate(s) is correlated with ancestry. Additionally, in the case where there is no marginal effect, our approach identifies new loci and shows that the genetic locus influences the phenotype and exerts its effects through interactions, which has important implications for the genetic architecture of the phenotype. The relative contribution of additive and non-additive genetic effects to variability in molecular phenotypes and disease risk is an important area of investigation, and our approach provides a direct test for detecting non-additive contributions (Powell et al., 2013). Also, if the SNP that is being tested for interaction with ancestry is a perfect ancestry informative marker (i.e. is fixed for the major allele in one population and is fixed for the minor allele in the other population), potential multicollinearity may arise between genotype and local ancestry when fitting the AITL model. In this scenario, our approach is unable to disentangle the signal from differential tagging and the interaction between genotype and ancestry. However, such cases are exceedingly rare and are not the case for any of the results presented here.

Environmental covariates are often not consistently measured across cohorts whereas genetic ancestry is nearly perfectly replicable. 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 (Keller, 2014).

Association testing for interaction effects involving continuous environmental exposures in the context of mixed-models remains an open problem. For binary environmental exposures, it has been shown that mixed-models control for population structure nominally better than including genetic ancestry (or principal components) as a covariate (Sul et al., 2016). Because it is unclear how mixed-models perform with continuous environmental exposures, especially those correlated with ancestry, in our analyses we took the standard approach of filtering related individuals and including ancestry as a covariate.

It has been shown that 2-step analyses may be more powerful for detecting interactions when exposures are binary (Hsu et al., 2012; Kooperberg & LeBlanc, 2008; Murcray, Lewinger, & Gauderman, 2008). However, these studies have primarily been done in a single homogeneous population, and the correct null distribution for the interaction effect must assume that the second stage procedure is independent of the marginal effect test statistic. In real data, using a 2-step approach in conjunction with AITL to test for interactions may be problematic because the interaction effect size will not necessarily be independent of the marginal effect size, as the allele frequency at any SNP will be a function of ancestry in an admixed population. Additionally, only 1 of the interaction results that we report here had a marginal effect ($p< 0.05$) and thus would have been missed by a 2-step approach. Thus, our approach can serve to complement or extend the frequently used 2-step procedure for detecting interaction effects.

Results from our multi-way epistasis simulation analyses and empirical data in 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. Our simulations also demonstrated that genetic ancestry can be a good proxy for an environmental covariate depending on the correlation between the two. However, it may be the case that there are multiple environmental factors interacting with a genetic locus, all of which are correlated with θ in differing degrees and effect sizes. Such a situation would mirror what we saw in our multi-way $G \times G$ simulations where a single interaction may not be detectable by using a traditional $G \times E$ test, but because θ aggregates the effects of all interacting covariates, AITL would be able to detect it. There are also other contexts in which modeling SNP by θ may be useful, such as using variance components. For example, SNP by θ interaction terms can be used in a mixed-model framework to test for interaction effects because genetic ancestry is correlated with many genetic markers and environmental covariates (Yang et al., 2010). We note that in our simulations, we made a normality assumption about the distribution of environment, as has been done in the majority GWAS to date. As is the case with linear regression, if model assumptions such as normality are not met, this may induce false positives associations.

For some traits, there may be systematic differences between ancestral populations in the genetic effects on the trait. In admixed individuals with these ancestral populations, the effect of genetic variation on phenotype will be reflected in the correlation between phenotype and θ, thereby affecting epistatic and $G \times E$ interactions. It will be interesting to see how much of the phenotype-ancestry correlations are due to epistatic and $G \times E$ interactions.

In our analysis of real data, we discovered gene by θ interactions associated with some genes that have known interactions. The GALA II dataset consists of asthmatic cases and controls from the GALA II study (Nishimura et al., 2013). The interactions that we detected could be signal from environmental factors interacting with genetic risk factors. Smoking, in particular maternal smoking during pregnancy, has been shown to contribute to methylation status in this dataset (Galanter et al., 2017). Various other environmental factors have also been shown to differ between racial/ethnic groups and may contribute to these interactions (Nishimura et al., 2013). The detected interactions could also result from multiway epistasis which may be part of a wider gene network, as others have shown for other phenotypes (De et al., 2015; Ritchie et al., 2001). In the GALA II Mexicans, the interaction of rs925736 with ancestry was associated with the methylation of HDAC4, a known histone deacetylase (HDAC). In concert with DNA methylases, HDACs function to regulate gene expression by altering chromatin state (Z. D. Smith & Meissner, 2013). In Europeans, HDACs have been shown to be associated with lung function through direct genetic effects and through environmental interactions (Artigas et al., 2011; Liao, Lin, & Christiani, 2013). For the GALA II Puerto Ricans, rs17091085 showed an interaction associated with the methylation state of SERPINA6 (Table 2). Of note, interaction between birth weight and SERPINA6 has been previously associated with Hypothalamic-Pituitary-Adrenal axis function (Anderson et al., 2014). Further investigations of our interaction findings are thus warranted.

In the GALA II (methylation) dataset, two of the eight significant associations replicated and, in general, the results had an enrichment of low p-values in the replication dataset. However, we note that if the interactions detected by AITL are multi-way epistasis it is more likely that the results will replicate. This is because most SNPs differentiated in the Mexicans will still be differentiated in the Puerto Ricans, and thus will be correlated with θ. If the interactions detected by AITL are $G \times E$ interactions, then the interactions are less likely to replicate because the same environmental covariate(s) will need to be correlated with ancestry in both groups.

Another caveat is that the Mexicans and Puerto Ricans, though independent, are part of the same study and occasionally technical artifacts, such as issues with genotyping or measuring methylation, can affect downstream analyses of both populations. For our analyses, we have taken careful quality-control steps to ensure that this is not the case and there is no apparent inflation of test statistics as demonstrated by our values for genomic control. Future research of interactions using AITL should keep such caveats in mind.

We investigated in detail the potential of single SNP-SNP interactions driving the results that were found both in the gene expression and methylation datasets. As demonstrated by the wide range of λ_{GC} values, we observed that non-linear effects can cause substantial departure from the null when testing for pairwise SNP-SNP interactions (Table S2). This is especially true when testing for interaction between SNPs *s* and *j*, where *s* has a significant interaction with θ and *j* is correlated with covariates that are also correlated with θ. As we saw in the gene expression data, including the significant SNP by θ term can properly control for such situations, but its use in standard pairwise interaction tests warrants further investigation.

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 methodological work is required 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 for heritability estimation in admixed populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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

References

Alexander DH, Novembre J, & Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655–1664.19648217

Anderson LN, Briollais L, Atkinson HC, Marsh JA, Xu J, Connor KL, (2014). 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. PloS One, 9(4), e92957.24691024

Artigas MAS, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, (2011). Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nature Genetics, 43(11), 1082–1090.21946350

Aschard H, Gusev A, Brown R, & Pasaniuc B (2015). Leveraging local ancestry to detect gene-gene interactions in genome-wide data. BMC Genetics, 16(1), 1591.

Balding DJ, & Nichols RA (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica, 96(1–2), 3–12.7607457

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodríguez-Cintrón W, Chapela R, Ford JG, Avila PC, Rodriguez-Santana J, Burchard EGL, & Halperin E (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics, 28(10), 1359–1367.22495753

Borrell LN, Nguyen EA, Roth LA, Oh SS, Tcheurekdjian H, Sen S, (2013). Childhood Obesity and Asthma Control in the GALA II and SAGE II Studies. American Journal of Respiratory and Critical Care Medicine, 187(7), 697–702.23392439

Burchard EGL, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, (2003). The Importance of Race and Ethnic Background in Biomedical Research and Clinical Practice. New England Journal of Medicine, 348(12), 1170–1175.12646676

Cheng CY, Reich D, Haiman CA, Tandon A, Patterson N, Elizabeth S, (2012). African Ancestry and Its Correlation to Type 2 Diabetes in African Americans: A Genetic Admixture Analysis in Three U.S. Population Cohorts. PloS One, 7(3), e32840.22438884

Choudhry S, Burchard EGL, Borrell LN, Tang H, Gomez I, Naqvi M, (2006). Ancestry-Environment Interactions and Asthma Risk among Puerto Ricans. American Journal of Respiratory and Critical Care Medicine, 174(10), 1088–1093.16973984

De R, Hu T, Moore JH, & Gilbert-Diamond D (2015). Characterizing gene-gene interactions in a statistical epistasis network of twelve candidate genes for obesity. BioData Mining, 8(1), 1431.

Delude CM (2015). Deep phenotyping: The details of disease. Nature, 527(7576), S14–S15.26536218

Devlin B, & Roeder K (2004). Genomic Control for Association Studies. Biometrics, 55(4), 997–1004.

Drake KA, Torgerson DG, Gignoux CR, Galanter JM, Roth LA, Huntsman S, (2014). A genome-wide association study of bronchodilator response in Latinos implicates rare variants. Journal of Allergy and Clinical Immunology, 133(2), 370–378.e15.23992748

Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, & Lin SM (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics, 11(1), 587.21118553

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, & Nadeau JH (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics, 11(6), 446–450.

Florez JC, Price AL, Campbell D, Riba L, Yu F, Duque C, (2011). Strong Association of Socioeconomic Status and Genetic Ancestry in Latinos: Implications for Admixture Studies of Type 2 Diabetes In Remembering or Forgetting Mendel: Sickle Cell Anemia and Racial Politics in Brazil (pp. 137–153). Palgrave Macmillan US.

Fraser HB, Lam LL, Neumann SM, & Kobor MS (2012). Population-specificity of human DNA methylation. Genome Biology, 13(2), R8.22322129

Galanter JM, Gignoux CR, Oh SS, Torgerson D, Pino-Yanes M, Thakur N, (2017). Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. eLife, 6, 1655.

Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature Genetics, 44(10), 1084–1089.22941192

Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, Mcrae AF, (2014). Detection and replication of epistasis influencing transcription in humans. Nature, 508(7495), 249–253.24572353

Hsu L, Jiao S, Dai JY, Hutter C, Peters U, & Kooperberg C (2012). Powerful Cocktail Methods for Detecting Genome-Wide Gene-Environment Interaction. Genetic Epidemiology, 36(3), 183–194.22714933

Jemal A, Bray F, Center MM, Ferlay J, Ward E, & Forman D (2011). Global cancer statistics. CA: a Cancer Journal for Clinicians, 61(2), 69–90.21296855

Kang EY, Han B, Furlotte N, Joo JWJ, Shih D, Davis RC, (2014). Meta-Analysis Identifies Gene-by-Environment Interactions as Demonstrated in a Study of 4,965 Mice. PLoS Genetics, 10(1), e1004022.24415945

Keller MC (2014). Gene-by-Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. Biological Psychiatry, 75(1), 18–24.24135711

Kooperberg C, & LeBlanc M (2008). Increasing the power of identifying gene × gene interactions in genome‐wide association studies, 32(3), 255–263.

Kumar R, Nguyen EA, Roth LA, Oh SS, Gignoux CR, Huntsman S, (2013). Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: The Genes-environments and Admixture in Latino Asthmatics (GALA II) study. Journal of Allergy and Clinical Immunology, 132(4), 896–905.e1.23684070

Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, Colangelo L, (2010). Genetic Ancestry in Lung-Function Predictions. New England Journal of Medicine, 363(4), 321–330.20647190

Liao SY, Lin X, & Christiani DC (2013). Gene-environment interaction effects on lung function- a genome-wide association study within the Framingham heart study. Environmental Health, 12(1), 787.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–753.19812666

Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee MC, (2014). Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. PLoS Genetics, 10(8), e1004549.25121757

McVean GA, Altshuler Co-Chair DM, Durbin Co-Chair RM, Abecasis GAR, Bentley DR, Chakravarti A, (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56–65.23128226

Murcray CE, Lewinger JP, & Gauderman WJ (2008). Gene-Environment Interaction in Genome-Wide Association Studies. American Journal of Epidemiology, 169(2), 219–226.19022827

Nishimura KK, Galanter JM, Roth LA, Oh SS, Thakur N, Nguyen EA, (2013). Early-Life Air Pollution and Asthma Risk in Minority Children. The GALA II and SAGE II Studies. American Journal of Respiratory and Critical Care Medicine, 188(3), 309–318.23750510

Peralta CA, Risch N, Lin F, Shlipak MG, Reiner A, Ziv E, (2010). The Association of African Ancestry and Elevated Creatinine in the Coronary Artery Risk Development in Young Adults (CARDIA) Study. American Journal of Nephrology, 31(3), 202–208.20029176

Powell JE, Henders AK, Mcrae AF, Kim J, Hemani G, Martin NG, (2013). Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. PLoS Genetics, 9(5), e1003502.23696747

Price AL, Patterson N, Hancks DC, Myers S, Reich D, Cheung VG, & Spielman RS (2008). Effects of cis and trans Genetic Ancestry on Gene Expression in African Americans. PLoS Genetics, 4(12), e1000294.19057673

Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, (2007). A Genomewide Admixture Map for Latino Populations. The American Journal of Human Genetics, 80(6), 1024–1036.17503322

Reiner AP, Carlson CS, Ziv E, Iribarren C, Jaquish CE, & Nickerson DA (2007). Genetic ancestry, population sub-structure, and cardiovascular disease-related traits among African-American participants in the CARDIA Study. Human Genetics, 121(5), 565–575.17356887

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, & Moore JH (2001). Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. The American Journal of Human Genetics, 69(1), 138–147.11404819

Rouhani F, Kumasaka N, de Brito MC, Bradley A, Vallier L, & Gaffney D (2014). Genetic Background Drives Transcriptional Variation in Human Induced Pluripotent Stem Cells. PLoS Genetics, 10(6), e1004432.24901476

Sanchez E, Webb RD, Rasmussen A, Kelly JA, Riba L, Kaufman KM, (2010). Genetically determined Amerindian ancestry correlates with increased frequency of risk alleles for systemic lupus erythematosus. Arthritis & Rheumatism, 62(12), 3722–3729.20848568

Seldin MF, Pasaniuc B, & Price AL (2011). New approaches to disease mapping in admixed populations. Nature Reviews Genetics, 12(8), 523–528.

Shaffer JR, Kammerer CM, Reich D, McDonald G, Patterson N, Goodpaster B, (2007). Genetic markers for ancestry are correlated with body composition traits in older African Americans. Osteoporosis International, 18(6), 733–741.17235662

Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, (2006). Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Human Molecular Genetics, 16(1), 1–14.17116639

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, (2004). A High-Density Admixture Map for Disease Gene Discovery in African Americans. The American Journal of Human Genetics, 74(5), 1001–1013.15088270

Smith ZD, & Meissner A (2013). DNA methylation: roles in mammalian development. Nature Reviews Genetics, 14(3), 204–220.

Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, & Cheung VG (2007). Common genetic variants account for differences in gene expression among ethnic groups. Nature Genetics, 39(2), 226–231.17206142

Sul JH, Bilow M, Yang WY, Kostem E, Furlotte N, He D, & Eskin E (2016). Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models. PLoS Genetics, 12(3), e1005849.26943367

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, & Risch N (2012). Estimating Kinship in Admixed Populations. The American Journal of Human Genetics, 91(1), 122–138.22748210

Torgerson DG, Gignoux CR, Galanter JM, Drake KA, Roth LA, Eng C, (2012). Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. Journal of Allergy and Clinical Immunology, 130(1), 76–82.e12.22502797

Vimaleswaran KS, Berry DJ, Lu C, Tikkanen E, Pilz S, Hiraki LT, (2013). Causal Relationship between Obesity and Vitamin D Status: Bi-Directional Mendelian Randomization Analysis of Multiple Cohorts. PLoS Med, 10(2), e1001383.23393431

Wong MY, Day NE, Luan JA, Chan KP, & Wareham NJ (2003). The detection of gene–environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? International Journal of Epidemiology, 32(1), 51–57.12690008

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, (2010). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7), 565–569.20562875

Ziv E, John EM, Choudhry S, Kho J, Lorizio W, Perez-Stable EJ, & Burchard EG (2006). Genetic Ancestry and Risk Factors for Breast Cancer among Latinas in the San Francisco Bay Area. Cancer Epidemiology Biomarkers & Prevention, 15(10), 1878–1885.

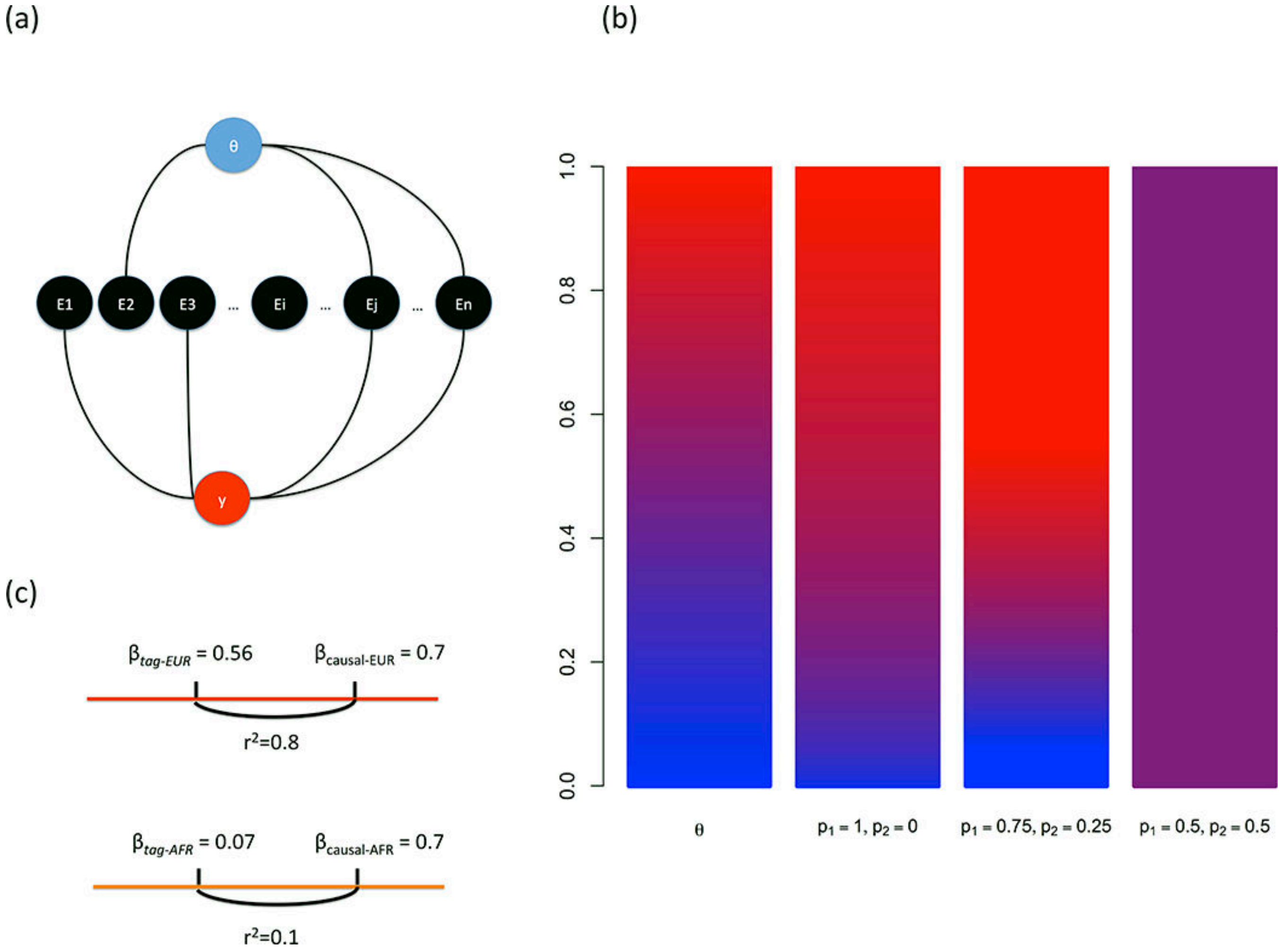

Figure 1.

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).

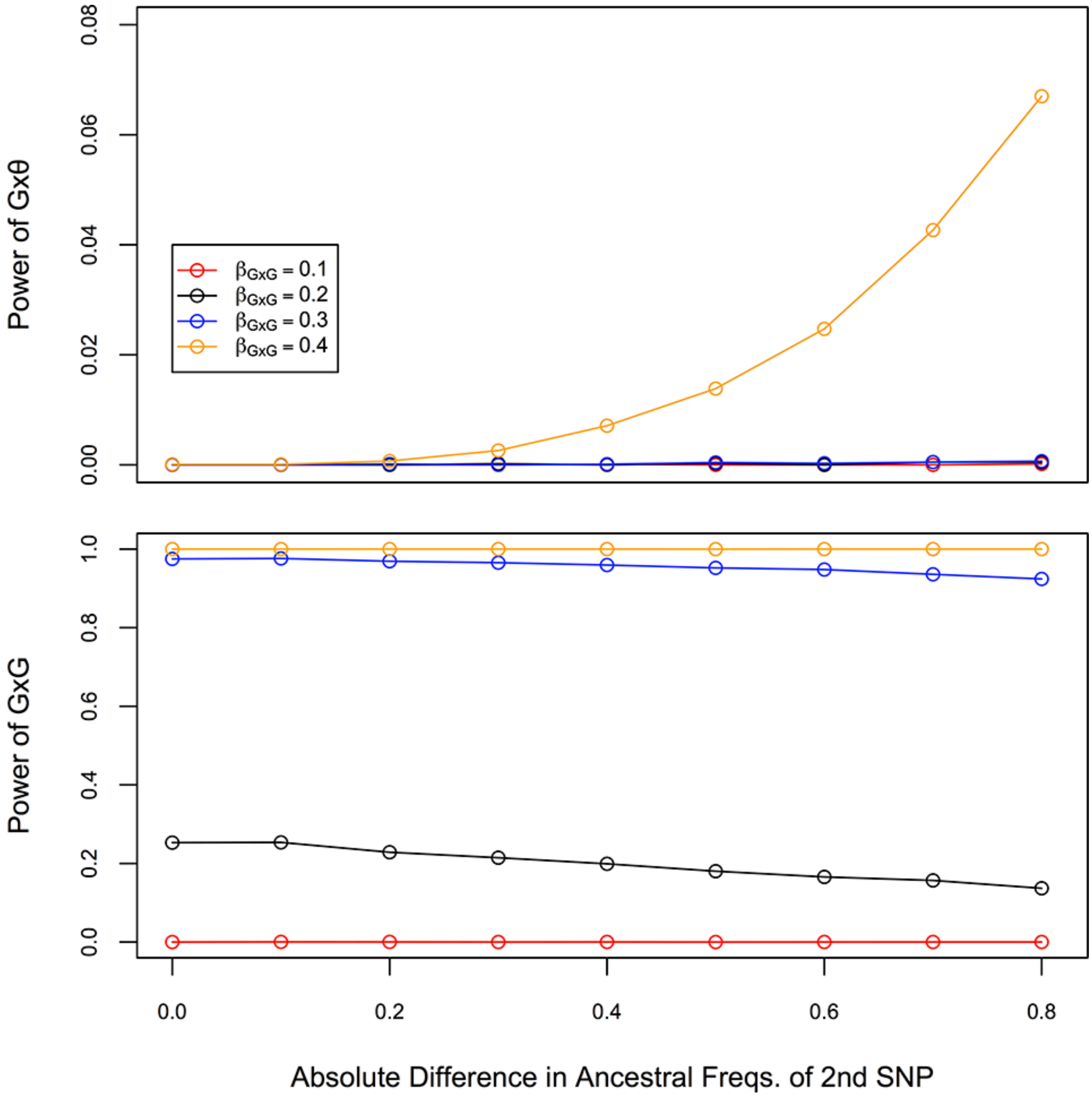

Figure 2.

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.

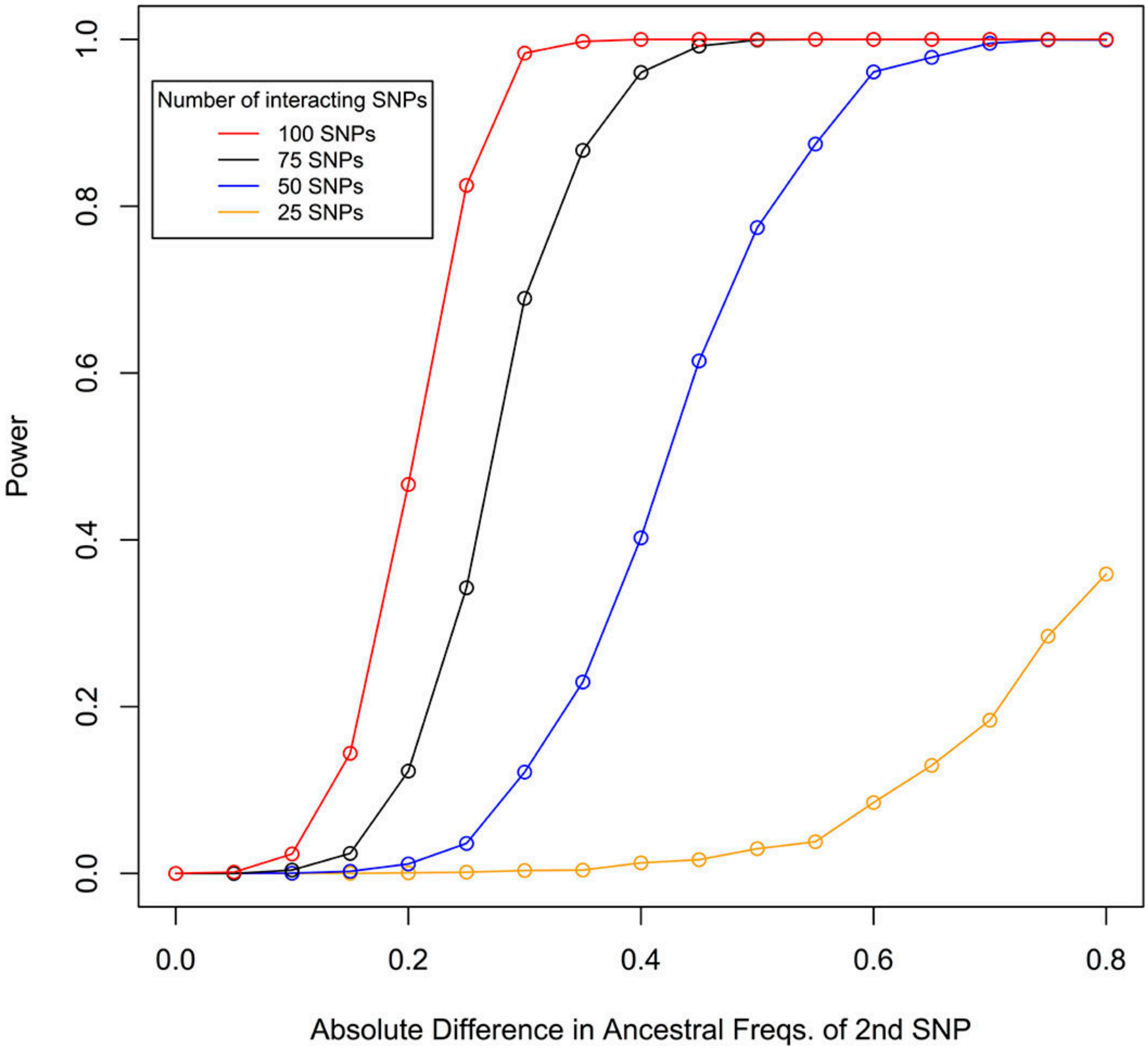

Figure 3.

Power Plots for Multi-way Pairwise Interaction Simulations Power of testing $G \times \theta$ as a function of the difference in the ancestral allele frequencies for multiple interacting SNPs.

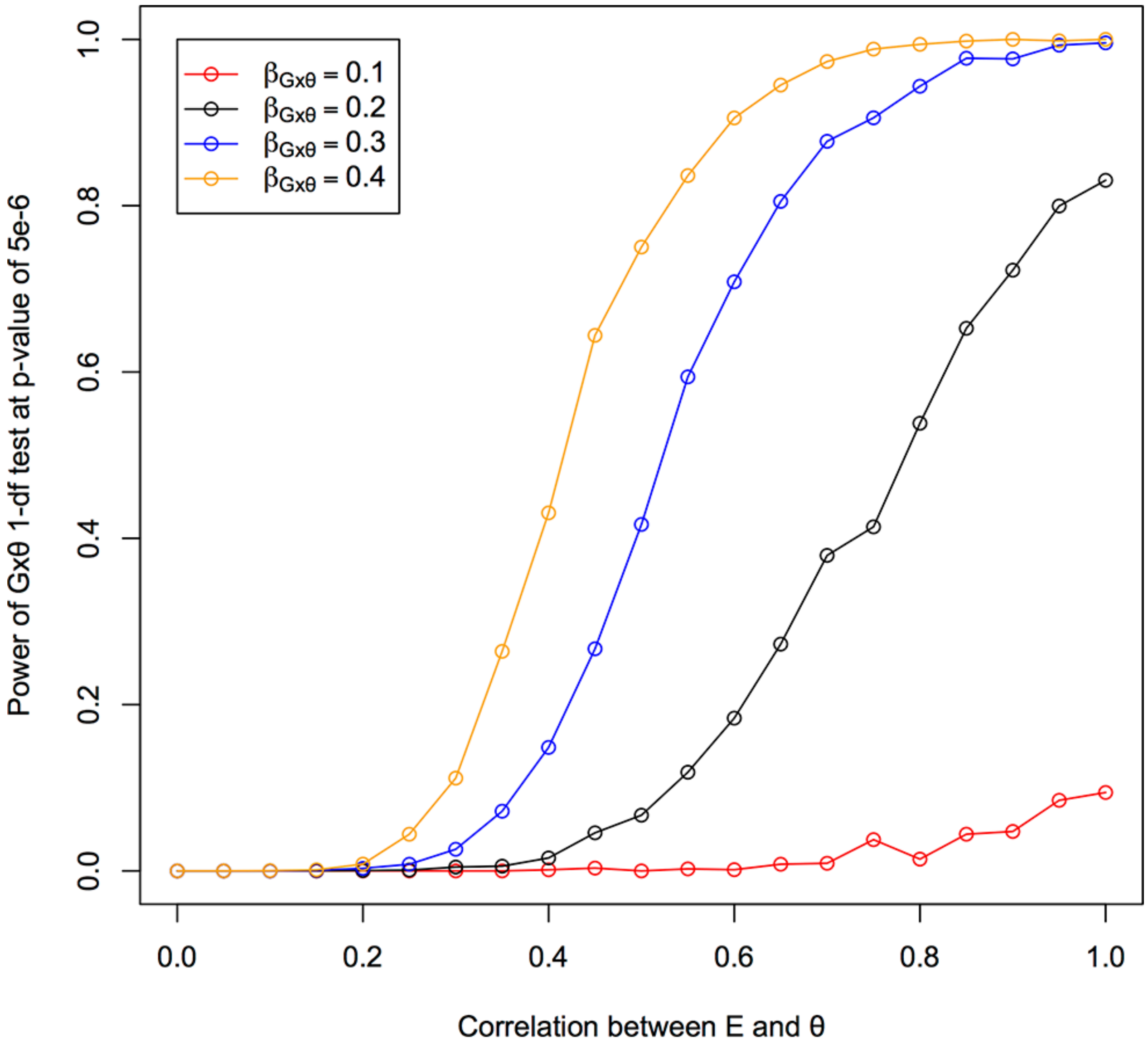

Figure 4.

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.

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=0.149
		μ=0.043,
		σ=0.025
			μ=0.561,
			σ=0.159

GALA II PR	μ=0.641,
	σ=0.094
		μ=0.246,
		σ=0.101
			μ=0.113
			σ=0.024

Mean and variance of the global ancestry distributions for each dataset. EUR, AFR, and NAM refer to European, African, and Native American ancestry respectively.

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.