Published in final edited form as:

*Genet Epidemiol*. 2018 March ; 42(2): 214–229. doi:10.1002/gepi.22103.

# Properties of Global and Local Ancestry Adjustments in Genetic Association Tests in Admixed Populations

**Eden R. Martin**[1,2], **Ilker Tunc**[3], **Zhi Liu**[4], **Susan H. Slifer**[1], **Ashley H. Beecham**[2], and **Gary W. Beecham**[1,2]

[1]John P. Hussman Institute for Human Genetics, University of Miami, Miller School of Medicine, Miami, FL

[2]John T. MacDonald Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL

[3]Bioinformatics and Systems Biology, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda MD

[4]Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

## Abstract

Population substructure can lead to confounding in tests for genetic association, and failure to adjust properly can result in spurious findings. Here we address this issue of confounding by considering the impact of global ancestry (average ancestry across the genome) and local ancestry (ancestry at a specific chromosomal location) on regression parameters and relative power in ancestry adjusted and unadjusted models. We examine theoretical expectations under different scenarios for population substructure; applying different regression models, verifying and generalizing using simulations, and exploring the findings in real-world admixed populations. We show that admixture does not lead to confounding when the trait locus is tested directly in a single admixed population. However, if there is more complex population structure or a marker locus in linkage disequilibrium (LD) with the trait locus is tested, both global and local ancestry can be confounders. Additionally, we show the genotype parameters of adjusted and unadjusted models all provide tests for LD between the marker and trait locus, but in different contexts. The local-ancestry adjusted model tests for LD in the ancestral populations; while tests using the unadjusted and the global-ancestry adjusted models depend on LD in the admixed population(s), which may be enriched due to different ancestral allele frequencies. Practically, this implies that global-ancestry adjustment should be used for screening, but local-ancestry adjustment may better inform fine-mapping and provide better effect estimates at trait loci.

### Keywords

African American; bias; confounding; Hispanic; linear regression; linkage disequilibrium; population stratification

**Corresponding Author:** Eden R. Martin, PhD, 1501 NW 10th Ave, Biomedical Research Building (BRB), Room 305, University of Miami, Miller School of Medicine, Miami, FL 33136, Emartin1@med.miami.edu.

## 1. INTRODUCTION

It is well known that ancestry differences among individuals can lead to confounding in tests of association between genes and a phenotype (Chakraborty & Weiss, 1988). Allele frequencies may differ depending on ancestry, as may trait values or disease prevalence, and this can lead to spurious associations if not properly accounted for in analyses. For this reason, it has become standard practice for genetic association tests to adjust for covariates that capture the background ancestry of study subjects. Prior to the genome-wide association study (GWAS) era, studies often relied on self-reported race and ethnicity. With the availability of dense, genome-wide genotyping, it became feasible to estimate ancestry from genetic markers, giving an assessment of ancestry that reflects the true genetic ancestry of the individual, rather than assessment that reflects cultural or societal perceptions (Burnett et al., 2006; Pasaniuc et al., 2011; Rosenberg, Li, Ward, & Pritchard, 2003). Often GWAS use principal component analysis to model ancestry differences among study participants (Price et al., 2006), though other methods are available (Alexander, Novembre, & Lange, 2009; Alexander & Lange, 2011; Cox & Cox, 2001; Falush, Stephens, & Pritchard, 2003; Hubisz, Falush, Stephens, & Pritchard, 2009; Pritchard, Stephens, & Donnelly, 2000; Tang, Peng, Wang, & Risch, 2005). Under many scenarios, the top principal components (PCs) from genome-wide array data are highly correlated with underlying ancestry. Thus, such PCs provide surrogate measures of “global” ancestry—that is, a picture of an individual’s average ancestral origin across the genome—and can be used as covariates in regression models for association analysis (Adeyemo et al., 2015; Armstrong et al., 2014; Beecham et al., 2014; Cruchaga et al., 2013; Melton et al., 2013; Naj et al., 2011; Nalls et al., 2014). However, global ancestry is not always representative of ancestry at individual genomic loci. The genome of admixed individuals is composed of stretches of DNA from different ancestral origins that are not always consistent with the global measure. This “local” ancestry can be estimated using a variety of statistical methods (Baran et al., 2012; Maples, Gravel, Kenny, & Bustamante, 2013) and offers a more fine-scale measure of genetic ancestry. However, outside of admixture mapping, few GWAS have used local ancestry in their association analyses (Baran et al., 2012; Pino-Yanes et al., 2015).

Recently some studies have begun to examine the properties of models adjusted for local ancestry and compare the behavior of hypothesis tests (e.g., power and Type I error) using local- and global-ancestry adjusted models. Liu et al (Liu, Lewinger, Gilliland, Gauderman, & Conti, 2013) argued that global-ancestry adjustment is sufficient to control Type I error, but local-ancestry adjustment can improve power when the LD in the ancestral populations and the correlation between genotype and ancestry in the admixed population are in the opposite direction. Zhang and Stram (2014) showed that adjusting for global ancestry is largely sufficient to control Type I error in their simulated model, but found that adjusting for local ancestry generally leads to lower power compared to adjusting for global ancestry, particularly for markers with large ancestral allele frequency differences. Wang *et al* (2011) pointed out that adjustment for global ancestry may not be sufficient to control Type I error when forces such as selection may be acting to create a local-ancestry effect near the test locus.

Here we seek to expand on these studies and develop a comprehensive picture of the impact of global- and local-ancestry adjustment in regression analysis for genetic studies in admixed populations. This includes a taking a formal look at statistical confounding, the forms of regression parameters under different adjustments, and approximations for relative power. Formulating these properties in terms of genetic parameters (e.g., allele frequencies, coefficients of linkage disequilibrium (LD), admixture proportions) under specific models of population structure allows us both to address questions of when and how to adjust for ancestry and to better interpret hypothesis tests and regression parameter estimates. Throughout, we consider both a single admixed population and a stratified admixed population (composed of two non-intermating admixed subpopulations), under a quantitative genetic model. We use simulations to validate our theoretical results and illustrate findings with numerical examples. Finally, because we find that LD in the ancestral populations and admixed populations are key parameters driving differences in regression models, we use genome-wide single-nucleotide polymorphism (SNP) data generated in samples from the Genomic Origins and Admixture in Latinos (GOAL) study and other public data to examine LD in admixed and non-admixed samples; thereby placing our findings in the context of contemporary admixed populations.

## 2. METHODS

### 2.1 Confounding in linear regression models

The definition of a confounding variable is one that is related to both the outcome (dependent variable) and the predictor (independent variable). Consider the adjusted linear model:

$$E(Y|X_1, X_2)=\beta_0+\beta_1 X_1+\beta_2 X_2, \quad (1)$$

where $Y$ is the outcome variable (a continuous quantitative variable), $X_1$ is the predictor variable being tested and $X_2$ is a potential confounder for which we are adjusting. Formally, $X_2$ is a confounder if and only if both of the following conditions hold:

**i.**

$$\rho_{12}^2 \neq 0$$

**ii.**

$$\rho_{Y2,1}^2 \neq 0 (\text{equivalent} \quad \text{to } \beta_2 \neq 0), \quad (2)$$

where $\rho_{12}^2$ is the squared correlation coefficient between $X_1$ and $X_2$, and

$\rho_{Y2,1}^2=\frac{(\rho_{Y2}-\rho_{Y1}\rho_{12})^2}{(1-\rho_{Y1}^2)(1-\rho_{12}^2)}$ is the partial squared correlation coefficient of $Y$ and $X_2$ with $X_1$ fixed (Robinson & Jewell, 1991). This means that $X_2$ is a confounder if it is correlated with both the predictor ($X_1$) and with the outcome $Y$ after removing the effect of the predictor. Note that condition (ii) is not the same as saying that $X_2$ and $Y$ are correlated. They may in fact be correlated, but if that correlation is explained solely by the relationship between $X_1$ and $X_2$ and the effect of $X_1$ on $Y$, then $X_2$ is not a true confounder.

If the conditions in (2) both hold, so that $X_2$ is a confounder, then we must include $X_2$ in the model to ensure valid inference about the effect of $X_1$ on $Y$; otherwise, adjustment for $X_2$ is not necessary to obtain valid inference about the effect of $X_1$ on $Y$. There are cases, however in which it may still be desirable with respect to power to adjust for a non-confounding variable (Robinson & Jewell, 1991). Specifically, suppose we consider testing the null hypothesis $\beta_1 = 0$. If condition (ii) holds but (i) does not, then adjusting for $X_2$ will increase power relative to power in the unadjusted model. If condition (i) holds but (ii) does not, then the adjusting for $X_2$ will decrease power relative to the unadjusted model. If neither condition holds, then the powers of the adjusted and unadjusted models are largely equivalent. This leads to the general conclusion for linear regression that adjusting for a covariate that is correlated with outcome (after removing the effect of variable of interest) is desirable with respect to power. However, there is no benefit to adjusting for a variable that is not correlated with outcome; and furthermore, such adjustment is undesirable when the covariate is also correlated with the predictor of interest.

### 2.2 Regression parameters

In the multiple regression equation shown above (1), the unstandardized regression coefficient $\beta_1$ takes the following form:

$$\beta_1 = \sqrt{\frac{V_Y}{V_1}} \left( \frac{\rho_{Y1} - \rho_{Y2}\rho_{12}}{1 - \rho_{12}^2} \right), \quad (3)$$

where $V_Y$ and $V_1$ are the variances of $Y$ and $X_1$ respectively, and the other terms are correlation coefficients. If $X_2$ is not a confounder (i.e., either condition i or ii of (2) fail), then the regression coefficient is equivalent to the coefficient from the univariate model with $X_1$ alone:

$$\beta_1 = \frac{C_{Y1}}{V_1}, \quad (4)$$

where $C_{Y1}$ is the covariance between $Y$ and $X_1$. However, if $X_2$ is a confounder, the equivalence no longer holds. Thus, adjusting by a confounding variable changes the interpretation of $\beta_1$ in the adjusted model compared to the coefficient in the unadjusted model. Understanding how the form of the regression coefficient being tested changes in terms of population genetic parameters, in unadjusted models or with covariate adjustment in the presence of confounding, is key to interpreting hypothesis tests. Failure to include confounding variables in the model can lead to what is known as specification bias, when parameters do not correctly model the desired effect because their relationship is confounded by an important omitted variable.

## 2.3 Asymptotic relative precision (ARP) and power

To compare the performance of estimators from adjusted and unadjusted regression models, we examined the asymptotic relative precision (ARP) of the models. This is the ratio of the variance of the estimate of the regression parameter ($\hat{\beta}_{UN}$) in an unadjusted model containing only $X_1$ to the variance of the parameter estimate $\hat{\beta}_1$ in the adjusted model (1):

$$\mathrm{ARP}(\hat{\beta}_1, \hat{\beta}_{\mathrm{UN}}) = \frac{\mathrm{Var}(\hat{\beta}_{\mathrm{UN}})}{\mathrm{Var}(\hat{\beta}_1)} = \frac{1-\rho_{12}^2}{1-\rho_{Y2,1}^2}. \quad (5)$$

When ARP($\hat{\beta}_1$, $\hat{\beta}_{UN}$) > 1, the adjusted model leads to greater precision (smaller variance for beta estimate); when ARP($\hat{\beta}_1$, $\hat{\beta}_{UN}$) < 1, the unadjusted model leads to greater precision. Precision is a measure of variance of the beta estimator. When the expected values of the beta estimators in the adjusted and unadjusted models are the same (which happens when there is no confounding), then ARP is related directly to relative power. Specifically in this case, ARP is equivalent to Asymptotic Relative Efficiency (ARE), which represents the ratio of required sample sizes for the tests to have equivalent power asymptotically (Serfling, 2009). We note that these statements are about the large-sample (asymptotic) behavior of tests, and conclusions may not hold in small samples.

## 2.4 Population and genetic models

We examined the correlation coefficients (2), regression parameters (3), and ARP (5) in the context of two different genetic models of population structure: (1) a single admixed population with two source ancestral populations (e.g., African Americans); and (2) a stratified admixed population composed of two subpopulations, each of which is itself admixed from two source ancestral populations (e.g., Haitians and Dominicans in Hispañola). We considered a quantitative trait locus (QTL) and "marker" loci with varying levels of LD dependent on recombination and population dynamics. Table I defines the notation used in the following sections.

**2.4.1 Single admixed population model—**For this model, we assume that there is initial mixing of two diploid ancestral populations with proportion $q$ of the chromosomes originating from ancestral population 1. Following the initial mixing, we assume that there has been random mating within the admixed population for some number of generations, but for simplicity no additional migration. Suppose that each haplotype of an individual is broken into $K$ chromosomes, each with $m$ distinct biallelic loci (total $M = mK$ loci). Here we use the term "haplotype" to refer to the entire set of chromosomes inherited from one parent (parental gamete). We define the recombination probability between any two loci $j$ and $k$ on the same chromosome, $r_{jk}$, as the chance that there has been some crossover event between the two loci since the initial mixing. For any pair of loci on different chromosomes (on the same haplotype), $r_u$ is the probability that the gamete has recombined (or assorted with a non-gametic allele) since the initial admixture. For a single generation, $r_u$ is expect to be ½. In general, the probabilities $r_{jk}$ and $r_u$ are functions of the crossover rate between a pair of loci and the number of generations since initial admixture.

In the admixed population, at each locus, the alleles of the two haplotypes each descend from ancestral population 1 or 0. For the *h*th haplotype at the *j*th locus, we define

$$A_{\mathrm{hj}}= \begin{array}{l} 1 \text{ if the jth locus descends from ancestral population 1} \\ 0 \text{ if the jth locus descends from ancestral population 0.} \end{array} \quad (6)$$

Initially the $A_{hj}$ are the same for all *j*, but recombination begins to break down the correlation across the genome. Averaging over the two haplotypes within an individual, $A_{\cdot j} = (A_{1j} + A_{2j})/2$, gives a measure of local ancestry for the *j*th locus for the individual. As a global measure of ancestry, we use the average of local ancestry measures across all loci (and both haplotypes):

$$\overline{A}=\frac{\sum_{j=1}^{M} A_{\cdot j}}{M}, \quad (7)$$

where *M* is the total number of loci across the genome.

We suppose that there is a biallelic QTL with alleles $T_1$ and $T_2$, where $p_i$ =the frequency of allele $T_1$ in ancestral population *i*, for *i* = 0,1. For the trait model, we assume allele effects at the QTL are additive so that the trait has the following conditional distribution:

$$Y\sim\begin{cases} N\left(a,\sigma^2\right) & \text{if } T_1T_1 \\ N(0,\sigma^2) & \text{if } T_1T_2 \\ N(-a,\sigma^2) & \text{if } T_2T_2 \end{cases}. \quad (8)$$

We define the genotypic random variable for the QTL:

$$G=\begin{cases} 1 & \text{if } T_1T_1 \\ 0 & \text{if } T_1T_2 \\ -1 & \text{if } T_2T_2 \end{cases}. \quad (9)$$

We also consider a marker locus with two alleles ($L_1$ and $L_2$), where $p_{Li}$ =the frequency of allele $L_1$ in ancestral population *i*, for *i* = 0,1, and define the genotypic random variable:

$$L=\begin{cases} 1 & \text{if } L_1L_1 \\ 0 & \text{if } L_1L_2 \\ -1 & \text{if } L_2L_2 \end{cases}. \quad (10)$$

The LD coefficient between alleles at the QTL ($T_1$) and marker ($L_1$) in ancestral population *i* is defined as follows:

$$D_i = P_{\mathrm{TLi}} - p_{\mathrm{Li}} p_i, \quad (11)$$

where $P_{TLi}$ is the frequency of the haplotype carrying $T_1$ and $L_1$ in ancestral population $i$. The LD in the admixed population depends on the LD in ancestral populations, the mixing proportion ($q$), the ancestral allele frequency differences at the QTL ($\Delta = p_1 - p_0$) and marker ($\Delta_L = p_{L1} - p_{L0}$), and probability that there has been a recombination event between QTL and marker ($r_{gl}$):

$$D^* = (1 - r_{\mathrm{gl}})\,(W + q(1-q)\Delta\Delta_L), \quad (12)$$

where $W = qD_1 + (1 - q)D_0$ is the weighted average of LD coefficients in the ancestral populations.

**2.4.2 Stratified admixed population model—**For this model, we assume a population stratified into two admixed subpopulations: subpopulation 0 and 1 (Appendix B.3 extends the model to >2 subpopulations). Let $Q$ be the probability that a random individual sampled from the whole population is from subpopulation 0. Suppose that for subpopulation $s$ ($s$ = 0,1), each locus is sampled from one of the two ancestral populations, and let $q_s$ be the probability a locus is from ancestral population 1. For the QTL, we assume Hardy-Weinberg equilibrium (HWE) within each subpopulation, but not necessarily in the overall population (e.g., no intermating between subpopulations). Let subpopulation membership for an individual be denoted by the random variable,

$$S = \begin{array}{l} 1 \text{ if the individual is in subpopulation } 1 \\ 0 \text{ if the individual is in subpopulation } 0. \end{array}$$

We let the trait within subpopulation 0 be defined as above in the single admixed population (8), but for subpopulation 1, assume the following trait model:

$$Y \sim \begin{cases} N(c+a, \sigma^2) & \text{if } T_1T_1 \\ N(c, \sigma^2) & \text{if } T_1T_2 \\ N(c-a, \sigma^2) & \text{if } T_2T_2 \end{cases} . \quad (13)$$

That is, both subpopulations have the same additive genetic effect ($a$), but subpopulation 1 has its mean trait value increased by a constant, $c$. This model would be reasonable, for example, if there was an independent environmental factor that increased the baseline value of the trait in subpopulation 1 (or lowered it in population 0) or independent genes that increased/decreased the baseline trait value.

Genotypic variables ($G$ and $L$) for the trait locus and marker loci are defined as above (9 and 10). For marker loci in the stratified admixed population we must consider LD between marker and trait loci in the ancestral populations, $D_0$ and $D_1$, and LD in the admixed subpopulations, $s$=0 and 1 (similar to equation 12):

$$D_s^{*}=(1-r_{\mathrm{gl}})\,(W_s+q_s(1-q_s)\Delta\Delta_L),$$

where $W_s = q_sD_1 + (1 - q_s)D_0$, the weighted disequilibrium coefficient in subpopulation $s$. The LD coefficient for the stratified population as a whole depends on LD in the ancestral populations ( $W^* = QW_1 + (1 - Q)\,W_0$), the ancestral allele frequency differences (Δ and $\Delta_L$), and functions of both the mixing parameters and relative size of the subpopulations ($\omega = Qq_1(1 - q_1) + (1 - Q)q_0(1 - q_0)$ and $\delta = Q(1 - Q)(q_1 - q_0)^2$). The LD coefficient in the stratified population can be written as follows (see Appendix B.2.1):

$$D^{**}=(1-r_{\mathrm{gl}})\,(W^{*}+\omega\Delta\Delta_L)+\delta\Delta\Delta_L. \quad (14)$$

We see that like $D^*$, the coefficient $D^{**}$ also depends on the recombination probability, $r_{gl}$, and decreases with increasing recombination; however, the term $\delta\Delta\Delta_L$ represents LD that results from the stratification (as a result of non-intermating between subpopulations) and is not influenced by recombination within the subpopulations.

## 2.5 Simulations

We performed simulations to test our theoretical conclusions using simuPOP (Peng & Kimmel, 2005), a forward-time population genetics simulation program. Genotype data and phenotype data were generated for single admixed and stratified admixed populations; then data were analyzed with regression models using different ancestry adjustments. Details of simulations are provided in Supplemental Material.

## 2.6 Analysis of LD in admixed and non-admixed datasets

**2.6.1 Datasets—**We estimated LD in seven different admixed datasets and three non-admixed “ancestral” datasets. Each of these would be comparable to our single-admixed population model but with three-way admixture rather than the simpler two-way admixture considered in our theory. Datasets included primarily family trios and parent-child pairs. The admixed datasets include samples from the GOAL study (Moreno-Estrada et al., 2013) sampled in South Florida with origins from five Caribbean countries: Colombia (39 individuals from 18 families), Honduras (24 individuals from 8 families), Cuba (60 individuals from 20 families), Puerto Rico (48 individuals from 16 families), Dominican Republic (27 individuals from 9 families). We also included four datasets from the 1000-Genomes Project (1000 Genomes Project Consortium et al., 2015): Mexican (MXL 64 individuals) and African American (ASW 61 individuals) admixed datasets, and two non-admixed datasets of European (CEU 99 individuals) and West African (YRI 107 individuals) individuals. Finally, a third ancestral population was included: 56 Native American samples from the Human Genome Diversity Project (includes Colombians, Karitiana, Maya, and Pima) (Cann et al., 2002). GWAS array data (1000 Genomes Project Consortium et al., 2015; Moreno-Estrada et al., 2013) were merged across datasets and 500 SNPs were selected on chromosome 20 to provide a range of intermarker distances. To represent unlinked markers, 500 SNPs were also selected on chromosome 21. SNPs were not selected on the basis of allele frequency. Genotypes from these SNPs were used for LD calculations.

**2.6.2 Estimation of disequilibrium coefficient—**Plink version 1.07 was used to estimate haplotype frequencies in the nine different datasets (Purcell et al., 2007). First, all founders were phased by Plink using the E-M algorithm; then all descendants of these founders were phased given the set of possible parental phases and assuming random-mating. Haplotype frequencies were then estimated from phased data and the LD coefficient was computed from these estimates as the difference between haplotype frequency and the product of respective allele frequencies. Haplotypes were matched across datasets to ensure consistency of the sign of the LD coefficient. R version 3.0.1 was used for computations with haplotype frequencies and plots of LD. In addition, we computed an estimate of $W$ for each admixed population using previous estimates of mixing proportions (Bryc, Durand, Macpherson, Reich, & Mountain, 2015; Johnson et al., 2011; Moreno-Estrada et al., 2013) and LD estimates from the three ancestral datasets.

# 3. RESULTS

## 3.1 Observations based on theory

For each of the population models, we evaluated the conditions of confounding on measures of local and global ancestry, and examined the form of the regression parameters from unadjusted and adjusted regression models (Table II). We considered tests of genotype-phenotype association at the QTL itself and at marker locus, taking into account LD. Where appropriate we examined ARP to compare the relative power of tests with different adjustments.

### 3.1.1 Single admixed population

**3.1.1.1 Testing at the QTL:** Suppose that we have measured the trait and the QTL genotype. We first asked whether local and/or global measures of ancestry are confounders by evaluating the conditions i and ii in (2). The relevant correlations for the two conditions are derived in Appendix A.1. We found that for both local and global ancestry, the partial correlations required to evaluate condition 2ii are always $0(\rho^2_{\mathrm{YA},G}=0 \text{ and } \rho^2_{Y\overline{A},G}=0)$. This makes sense because at the QTL, once genotype is known, ancestry provides no additional information. It follows that neither local nor global ancestry are confounders for the relationship between the QTL genotype and trait in this scenario.

Since neither local nor global ancestry measures are confounders, the regression parameters in the unadjusted and both adjusted models should all take the same form. It is easy to show (see Appendix A.1.5) that the genotype-term regression parameters for the unadjusted, local-ancestry adjusted and global-ancestry adjusted models are $\beta_G=\beta^*_G=\beta^{'}_G=a$, respectively (Table III); thus all models correctly model the true genetic effect.

Even though ancestry is not a confounder, and thus not a necessary covariate, we can explore the consequence of adjustment on statistical power. Since each model correctly estimates the same parameter, we can use ARP as a measure of relative power (or more precisely, the relative sample size required to achieve equivalent power). Because the partial correlations are 0 for both the local and global ancestry measures, the ARPs comparing the estimates from the adjusted models to unadjusted models are the following:

$$\mathrm{ARP}(\hat{\beta}^{*}_{G}, \hat{\beta}_{G}) = 1 - \rho^{2}_{\mathrm{GA}} \ \text{ and } \ \mathrm{ARP}(\hat{\beta}^{'}_{G}, \hat{\beta}_{G}) = 1 - \rho^{2}_{G\overline{A}}.$$

These ARPs are clearly both always less than or equal to 1, and it follows that the test of the genetic effect in either ancestry adjusted model will always be less powerful than the test based on the unadjusted model.

We can further describe the relationship between $\mathrm{ARP}(\hat{\beta}^{*}_{G}, \hat{\beta}_{G})$ and $\mathrm{ARP}(\hat{\beta}^{'}_{G}, \hat{\beta}_{G})$, by noticing that $\rho^{2}_{G\overline{A}} = \left(\frac{\varphi_g^2}{\varphi}\right)\rho^{2}_{\mathrm{GA}}$, where $\varphi = 1 - \left(\frac{K-1}{K}\right) r_{U} - \frac{1}{\mathrm{Mm}} \sum_{j=1}^{m} \sum_{k \neq j}^{m} r_{\mathrm{jk}}$ and $\varphi_g = 1 - \left(\frac{K-1}{K}\right) r_{U} - \frac{1}{M} \sum_{j \neq g}^{m} r_{\mathrm{gj}}$ (see Appendix A.1.5). The first equation, $\boldsymbol{\varphi}$, involves the sum over all pairwise recombination probabilities and the second equation, $\boldsymbol{\varphi_g}$, involves the sum of recombination probabilities between each locus and the QTL (locus *g*). We show in Appendix C and Supplemental Figure 2 that $1 \geq \left(\frac{\varphi_g^2}{\varphi}\right) \geq \frac{1}{M}$; the ratio is largest when there is little recombination and smallest when there is lots of recombination. This gives the following relationship: $\mathrm{ARP}(\hat{\beta}^{*}_{G}, \hat{\beta}_{G}) \leq \mathrm{ARP}(\hat{\beta}^{'}_{G}, \hat{\beta}_{G}) \leq 1$, meaning the unadjusted model will have the most power, followed by the global adjusted model and then the local-adjusted model. At the extremes, when $r_{jk} = r_U = 0 \forall j,k$ (which would be expected immediately after the initial admixture event), then $\left(\frac{\varphi_g^2}{\varphi}\right) = 1$ and $\mathrm{ARP}(\hat{\beta}^{*}_{G}, \hat{\beta}_{G}) = \mathrm{ARP}(\hat{\beta}^{'}_{G}, \hat{\beta}_{G}) \leq 1$, which implies that adjusting for global or local ancestry would give equivalent power. This makes intuitive sense because in this case of no recombination, global ancestry and local ancestry are equivalent. For the other extreme, when $r_{jk} = r_U = 1 \forall j,k$, then $\left(\frac{\varphi_g^2}{\varphi}\right) = \frac{1}{M}$ and $\mathrm{ARP}(\hat{\beta}^{'}_{G}, \hat{\beta}_{G}) = 1 - \frac{1}{M}\rho^{2}_{\mathrm{GA}} \approx 1$, for large *M*. In this case, little is lost by using the global-adjusted model relative to the unadjusted model, but the local adjusted model loses power due to the correlation between QTL genotype and local ancestry.

We examined the magnitude of the adjusted models' power loss relative to the unadjusted model (Supplemental Figures 3 and 4) over a range of parameter values. We found that the power loss with the local-ancestry adjusted model can be substantial, particularly when allele frequencies are very different between the ancestral populations (Supplemental Figure 3). The global-ancestry adjusted model (Supplemental Figure 4) still loses power, but less power than the local-ancestry adjusted model. Additionally, we find that when recombination increases (Supplemental Figure 4B), $\mathrm{ARP}(\hat{\beta}^{'}_{G}, \hat{\beta}_{G})$ is very close to 1. This illustrates that as the amount of recombination across the genome increases (such as after several generations of random mating), the power of the global-ancestry adjusted model approaches the power of the unadjusted model.

**3.1.1.2 Testing at a marker:** For tests at a marker locus, we also evaluated the confounding conditions (2), but with respect to marker genotype, local ancestry at the marker, and global ancestry. The correlation between the marker genotype and local ancestry at the marker (required for condition 2i) is analogous to that derived for the QTL: $\rho^2_{\mathrm{LA}}=\frac{q(1-q)\Delta_L{}^2}{q(1-q)\Delta_L{}^2+\gamma_L}$, where $\gamma_L = qp_{L1}(1 - p_{L1}) + (1 - q)p_{L0}(1 - p_{L0})$. However, unlike the case of testing at the QTL, the partial correlation necessary to evaluate condition 2ii can be non-zero because it involves both the LD between QTL and marker alleles and the correlation between QTL and marker ancestries. Specifically, we can show that the numerator of $\rho^2_{\mathrm{YA},L}$ is proportional to $\Psi^2$, where $\Psi=\sqrt{2q(1-q)}a(1-r_{\mathrm{gl}})(\Delta\gamma_L-\Delta_L W)$ (see Appendix A.2.4 for full form).

As a validation, we can see that when there is complete LD in the ancestral populations (so that marker and QTL alleles are completely correlated), then $\Delta = \Delta_L$ and $W = \gamma_L$, and these correlations reduce to the equations that we derived for testing the QTL itself. In that special case, there is no confounding. In general the partial correlation is 0 (and local ancestry is not a confounder) when any one of the following conditions holds: (1) $a = 0$ or (2) $r_{gl} = 1$ or (3) $\Delta\gamma_L = \Delta_L W$. The last condition holds when $\Delta = 0$ and either $\Delta_L = 0$ or $W = 0$ (or if there is complete LD in the ancestral populations). If any of these conditions hold, local ancestry at the marker will not be a confounder. Alternatively the correlation between marker genotype and local ancestry will be 0 (which would also lead to no confounding) if $\Delta_L = 0$.

The correlation between marker genotype and global ancestry also has the same form shown before for the QTL:

$$\rho^2_{L\overline{A}}=\rho^2_{\mathrm{LA}}\left(\frac{\varphi_l^2}{\varphi}\right),$$

where $\varphi_l=1-(\frac{K-1}{K})r_U-\frac{1}{M}\sum_{j\neq l}^{m}r_{\mathrm{lj}}$ and $\varphi$ is as defined previously. The numerator of $\rho^2_{Y\overline{A},L}$ is proportional to $(\Psi + \psi)^2$, where $\psi=2\mathrm{aq}(1-q)[D^*\Delta_L(1-\varphi_l)-p^*_L(1-p^*_L)\Delta(1-\varphi_g)]$ (see Appendix A.2.4). If there has been little recombination, $\varphi_l$ and $\varphi_g$ will both be close to 1 and the numerator will look like $\Psi^2$ as with local ancestry. In general, the partial correlation is 0 if either of the following conditions hold: (1) $a = 0$ or (2) $\Delta = 0$ and either $\Delta_L = 0$ or $W = 0$. The correlation between marker genotype and local ancestry is 0 if $\Delta_L = 0$. Taken together, this means global ancestry will not be a confounder if either: (1) $a = 0$ or (2) $\Delta_L = 0$ or (3) $\Delta = 0$ and $W = 0$. Unlike adjusting for local ancestry, this partial correlation is not necessarily 0 when $r_{gl} = 1$. These observations show that when we are testing at a marker locus (not in perfect LD with the QTL), both local ancestry at the marker and global ancestry can be confounders, and adjusting may be required to obtain valid inference.

We can gain more insight by examining the regression parameters for the unadjusted and adjusted models (Table III). For the unadjusted model, we can show (see Appendix A.2.5) that $\beta_L$ is a function of the LD between alleles $T_1$ at the QTL and $L_1$ at the marker locus in the admixed population, $D^*$ (equation 12). Specifically, we see (Table III) that the genotype coefficient in the unadjusted model is 0 if either there is no genetic effect ($a = 0$) or there is no LD in the admixed population ($D^* = 0$). It is well known the admixture can generate LD even in the absence of LD in the ancestral populations. LD in the admixed population depends on the LD in the ancestral populations ($W$), the difference between allele frequencies in the ancestral populations (Δ and $\Delta_L$), as well as recombination between the QTL and marker ($r_{gl}$). We expect no LD ($D^* = 0$) if either (1) $r_{gl} = 1$ or (2) W = 0 and either Δ = 0 or $\Delta_L = 0$. Either of these conditions or the condition that $a = 0$ satisfies the criteria for no confounding by local ancestry discussed above. Therefore, we can conclude that the unadjusted model will provide a valid test of the null hypothesis of no genetic effect ($a = 0$) or no LD in the admixed population ($D^* = 0$).

For the local-ancestry adjusted model, we can show that $\beta_L^*$ is a function of genetic effect ($a$), LD in the ancestral populations ($W$) and recombination between the marker and QTL ($r_{gl}$) (see Appendix A.2.5). Specifically we see (Table III) that $\beta_L^*=0$ if $a = 0$, $W = 0$ or $r_{gl} = 1$; thus the model adjusted for local ancestry provides a valid test of either no genetic effect ($a = 0$) or no LD in the ancestral populations ($W = 0$). Since the parameter is also 0 when there has been recombination between the marker and trait loci, we can view this test as having power only for markers linked to the QTL (or for very recent admixture).

The genotype coefficient for the global-ancestry adjusted model is shown in Table III (see Appendix A.2.5 for derivation). We see that, in general, $\beta_L^{'}=0$ if either $a = 0$ or $D^* = 0$ and either Δ = 0 or $\Delta_L = 0$. This means that a test of genotypic effect using the model adjusted by global ancestry is valid as a test of no genetic effect, but is not generally a valid test of $D^* = 0$ (unlike the unadjusted model) or $W = 0$ (unlike the local-ancestry adjusted model). In Appendix C and Supplemental Figure 2, we show the behavior of the functions $\frac{\varphi_g\varphi_l}{\varphi}$ and $\frac{\varphi_l^2}{\varphi}$, which appear in the numerator and denominator of $\beta_L^{'}$, as a function of the amount of pairwise recombination. We see that when of the amount of recombination increases (e.g., Supp Fig 2B as $s$ increases), the functions will both approach 0, and then $\beta_L^{'}$ will approximate the parameter from the unadjusted model. On the other extreme, when there has been little recombination (e.g., Supp Fig 2D for small $s$), both $\frac{\varphi_g\varphi_l}{\varphi}$ and $\frac{\varphi_l^2}{\varphi}$ will be close to 1, and then $\beta_L^{'}$ will approximate parameter from the local-ancestry adjusted model.

In general power comparisons between the different models using ARP are not appropriate since the genotype regression parameters differ for the different models. However, relative rejection rates can be addressed with simulations, results following.

### 3.1.2 Stratified admixed population

**3.1.2.1 Testing at the QTL:** In the stratified population model, the two admixed populations may have different ancestry proportions ($q_1$, $q_0$) and different phenotypic means (shifted by a constant $c$). For the linear model adjusted for local ancestry at the QTL, the squared correlation coefficients relevant to assess confounding by local ancestry are as follows (see Appendix B.1.4):

$$\rho^2_{\mathrm{GA}}=\frac{(\omega+2\delta)\Delta^2}{(\omega+2\delta)\Delta^2+\gamma'}$$

$$\rho^2_{\mathrm{YA},G}=\frac{2c^2Q(1-Q)\delta\gamma'}{2c^2Q(1-Q)\delta\gamma'+((\omega+2\delta)\Delta^2+\gamma')(c^2Q(1-Q)\omega+(\omega+2\delta)\sigma^2)},$$

where $\Delta = p_1 - p_0$ as defined before, $\gamma' = \alpha p_1(1 - p_1) + (1 - \alpha)p_0(1 - p_0)$, $\delta = Q(1 - Q)(q_1 - q_0)^2$, and $\omega = Qq_1(1 - q_1) + (1 - Q)q_0(1 - q_0)$.

As in the single admixed population, $\rho^2_{\mathrm{GA}}=0$ if the allele frequencies at the QTL are the same in the two ancestral populations ($\Delta = 0$). It is also 0 if the mixing proportions are the same in the subpopulations, $q_1 = q_0$. Unlike the single admixed population model, we see that the partial correlation is non-zero in the stratified population when there are different ancestry proportions between the subpopulations ($q_1 \neq q_0$) and different phenotypic means ($c \neq 0$). Thus in general, if there are differences between ancestral allele frequencies at the QTL and differences in ancestry proportions and phenotypic means in the subpopulations, then local ancestry will be a confounder.

For global ancestry, the correlations to assess confounding are the following (see Appendix B.1.4 for details):

$$\rho^2_{G\overline{A}}=\frac{(\omega\varphi_g+2\delta)^2\Delta^2}{[(\omega+2\delta)\Delta^2+\gamma'](\omega\varphi+2\delta)}$$

$$\mathrm{numerator}(\rho^2_{Y\overline{A},G})\propto 2Q(1-Q)c^2\delta[\omega\Delta^2(1-\varphi_g)+\gamma']^2.$$

where $\varphi$ and $\varphi_g$ are functions of the recombination probabilities defined previously. The conclusions are the same as for local ancestry: if there are differences between ancestral allele frequencies at the QTL ($\Delta \neq 0$) and differences in ancestry proportions ($\delta \neq 0$) and phenotypic means ($c \neq 0$) in the subpopulations, then global ancestry will be a confounder.

The parameter estimates for the genotypic term in the unadjusted and ancestry adjusted models are shown in Table III (see Appendix B.1.5 for derivations). For the unadjusted model, we see that the regression parameter does not generally provide an estimate of the true genetic effect of the QTL, $a$ (except in the case $c = 0$ or $\Delta = 0$ or $q_1 = q_0$). On the other hand, the model adjusted by local ancestry does correctly model the true genetic effect. Like the unadjusted model, the regression parameter model adjusted by global ancestry, $\beta_G^{'}$, does not estimate the genetic effect, $a$, when $c \neq 0$, $\Delta \neq 0$ and $q_1 \neq q_0$. However, this bias also depends on pairwise recombination through $\varphi$ and $\varphi_g$. Supplemental Figure 1 shows that $0 < \varphi \leq 1$ and the difference $\varphi - \varphi_g$ approaches 0 as recombination increases or decreases (and tends to be small in general). Therefore, unless $c$ is very large relative to $a$, the bias is expected to be small.

Although, subpopulation membership is often unknown, it is interesting to consider the properties of the model adjusted for membership. For the regression model adjusted for subpopulation membership (Table II), it is not hard to show (see Appendix B.1.5) that, like the local-ancestry adjusted model, using membership as a covariate also correctly models the true genetic effect: $\beta_G^{''}=a$ (Table III).

Since adjusting for local ancestry or subpopulation membership both provide estimates of the genetic effect, $a$, we can use ARP to compare power. Using the correlations derived in Appendix B.1.3 we have,

$$\mathrm{ARP}(\hat{\beta}_G^{*}, \hat{\beta}_G^{''})=\left(\frac{\gamma'\sigma^2(\omega+2\delta)}{\gamma'\sigma^2(\omega+2\delta)+\Delta^2\omega\sigma^2(\omega+2\delta)+c^2Q(1-Q)\omega(\omega\Delta^2+\gamma')}\right).$$

All parameters in the second and third terms of the denominator are positive, and consequently $\mathrm{ARP}(\hat{\beta}_G^{*}, \hat{\beta}_G^{''}) \leq 1$. Thus the model adjusting for subpopulation membership is at least as powerful as the local-ancestry adjusted. Supplemental Figure 5 shows examples of $\mathrm{ARP}(\hat{\beta}_G^{*}, \hat{\beta}_G^{''})$ for various parameter values. Power of the tests in the two models is most similar ( $\mathrm{ARP}(\hat{\beta}_G^{*}, \hat{\beta}_G^{''})$ closest to 1) when $q_0$ and $q_1$ are most different (Supplemental Figure 5, red line $q_1 = 1$). Though not shown in this figure, when $|q_1 - q_0| = 1$, we have $\omega = 0$, and it can be seen from the formula above that $\mathrm{ARP}(\hat{\beta}_G^{*}, \hat{\beta}_G^{''})=1$. This makes sense because when local ancestries come from distinct ancestral populations in the strata, the local ancestry variable will be the same as the subpopulation membership variable. On the other hand, when $q_1 - q_0 = 0$ (Supplemental Figure 5, purple line $q_1 = 0.65$), local ancestry gives no information about subpopulation membership, and we see the greatest loss in power for the local-ancestry adjusted model relative to the membership-adjusted model. As with the single admixed population, for each value of $q_j$, the loss in power using local ancestry instead of subpopulation membership increases with more divergent ancestral allele frequencies ($|\Delta|$). In Supplemental Figure 5, we also see that the ARPs increase (and hence the powers of the local ancestry and membership adjusted models become more similar) if the difference in trait means between populations is smaller (e.g., $c = 1$ vs $c = 2$).

**3.1.2.2 Testing at a Marker:** To assess confounding by local ancestry when testing at a marker, we derive the following correlation coefficients (see Appendix B.2.4):

$$\rho^2_{\mathrm{LA}}=\frac{(\omega+2\delta){\Delta_L}^2}{(\omega+2\delta){\Delta_L}^2+\gamma'_L}$$

$$\text{numerator}\left(\rho^2_{\mathrm{YA},L}\right)\propto\Phi^2,$$

where $\Phi=a\{(1-r_{\mathrm{gl}})(\omega+2\delta)(\Delta\gamma'_L-W^*\Delta_L)+2\delta\Delta r_{\mathrm{gl}}\gamma'_L\}+\mathrm{cQ}(1-Q)(q_1-q_0)\gamma'_L$. Similar to the case of testing at the QTL, we see that local ancestry will be a confounder if there are differences between ancestral allele frequencies at the marker ($\Delta_L \neq 0$), differences in ancestry proportions ($q_1 \neq q_0$) and phenotypic means in the subpopulations ($c \neq 0$). However, when testing at a marker, the partial correlation $\rho^2_{\mathrm{YA},L}$ can also be non-zero even when $q_1 = q_0$ or $c = 0$ if $a \neq 0$.

For global ancestry, the relevant correlation coefficients are derived in Appendix B.2.4:

$$\rho^2_{L\overline{A}}=\frac{(\varphi_l\omega+2\delta)^2\Delta^2_L}{[(\omega+2\delta)\Delta^2_L+\gamma'_L][\omega\varphi+2\delta]}$$

$$\text{numerator}\left(\rho^2_{Y\overline{A},L}\right)\propto(\Phi+\Upsilon)^2,$$

where

$$\begin{aligned}\Upsilon=2\omega\,\{a\{(1-r_{\mathrm{gl}})(W^*\Delta_L(1\\ -\varphi_l)-\Delta\gamma'_L)\\ +\Delta\gamma'_L\varphi_g\\ +\Delta\Delta^2_L\{r_{\mathrm{gl}}(\omega\varphi_l+2\delta)+(\varphi_g-\varphi_l)(\omega+2\delta)\}\Big\}\\ -[\mathrm{cQ}(1\\ -Q)(q_1\\ -q_0)\Delta^2_L][(\varphi_l-1)]\Big\}.\end{aligned}$$

If there is complete LD between QTL and marker, then Y = 0 and the correlations are the same as derived for local ancestry. We see that like local ancestry, global ancestry can be a confounder in general.

Examining the regression coefficients from the unadjusted and adjusted models (Table III), we see that the coefficient for the unadjusted model does not estimate the true genetic effect, and will not provide a valid test of $a = 0$ if $c \neq 0$, $\Delta_L \neq 0$ and $q_0 \neq q_1$. Unlike the unadjusted model in the single admixed population examples, in stratified populations the unadjusted model does not generally provide a test of LD in the population as a whole either, $D^{**} \neq 0$ because the regression parameter can be non-zero even when $D^{**} = 0$ if $\Delta_L \neq 0$ and $q_0 \neq q_1$. Therefore, when testing a marker locus, the unadjusted model generally provides neither a test for no genetic effect or of no LD in the stratified population.

For the model adjusted for local ancestry, the form of the regression parameter $\beta_L^*$ is similar to that in the single admixed population (Table III). We see that $\beta_L^*=0$ if $a = 0$ or $W^* = 0$ or $r_{gl} = 0$; thus the model adjusted for local ancestry provides a valid test of either no genetic effect or no LD in the ancestral populations (independent of the value of $c$). Notably it is also possible for the regression parameter to be 0 even when the LD coefficients in the ancestral populations are non-zero if the LD in the ancestral populations is in opposite directions such that $QW_0 = (1 - Q)W_1$, or equivalently $\alpha D_0 = (1 - \alpha)D_1$.

As for the unadjusted model, the coefficient for the model adjusted for global ancestry also has a bias that depends on c, and so does not generally provide a test for no genetic effect ($a = 0$). The bias term will be 0 if $\mathrm{c}\mathrm{Q}(1 - Q)(q_1 - q_0)\omega\Delta_L\frac{(\varphi-\varphi_l)}{(\omega\varphi+2\delta)}=0$, which happens if $c = 0$, $\Delta_L = 0$, $q_1 = q_0$, or $\frac{(\varphi-\varphi_l)}{(\omega\varphi+2\delta)}=0$. We show in Supplemental Figure 1 that $\varphi$ and $\varphi_l$ approach 0 as the amount of recombination increases and approach 1 as recombination decreases; thus $(\varphi - \varphi_l) \to 0$ at the extremes of recombination and the bias term disappears. However, the bias may be non-negligible for modest amounts of recombination (as with moderately recent populations) if $c$ is large. Notably, as recombination decreases ($\varphi$ and $\varphi_l$ approach 1), we can show that $\beta_L^{'} \to \beta_L^*$. This makes sense because with little recombination, local ancestry and global ancestry measures should be similar for each individual. As recombination increases ($\varphi$ and $\varphi_l$ approach 1), we find that $\beta_L^{'} \to \frac{a(D^{**}-\delta\Delta\Delta_L)}{\omega\Delta_L^2+\gamma_L^{'}}$, which is the marker-genotype coefficient for the model adjusted for subpopulation membership (Table III).

For the subpopulation-membership adjusted model, we see that $\beta_L^{''}$ is 0 if $a = 0$ (Table III), showing that it is a valid test of genetic effect regardless of the value of $c$. It is also 0 if there is no LD in both of the admixed subpopulations, $D_0^*=D_1^*=0$. The coefficient is not 0, however, if $D^{**} = 0$, unless $\Delta = 0$ or $\Delta_L = 0$, nor is it 0 if $W^* = 0$, unless $\Delta = 0$ or $\Delta_L = 0$ or $r_{gl} = 1$. This means adjusting for subpopulation membership does not generally provide a valid test of LD in the stratified population or LD in the ancestral populations, but does provide a valid test of no genetic effect or of no LD in both admixed subpopulations.

### 3.2 Simulation results

We conducted simulations of single admixed and stratified admixed populations to validate our theoretical findings (Supplemental Methods and Results). Consistent with the theoretical conclusions above, our simulations of a single admixed population show that all models result in valid tests of the null hypothesis $a = 0$ (at the QTL and marker loci). For $a \neq 0$, the rejection rates depend on LD as expected. The local-ancestry model, but not necessarily the unadjusted or global adjusted models, provides valid tests of $W = 0$. The unadjusted model provides valid tests of the null hypothesis of $D^* = 0$. In general we found that power depends on the relative size of $W$ and $D^*$. When $D^* > W$ the unadjusted tends to be most powerful and when $W > D^*$ the local-ancestry adjusted model tends to be most powerful. In general, we found that the global-ancestry adjusted model performs similarly to the unadjusted model or has power between the unadjusted and local-ancestry adjusted model. This agrees with our theory that the regression parameter for the global-adjusted model should be close to the unadjusted model if there has been sufficient recombination, which is the case with our simulations involving 20 generations of random mating following admixture.

Results from simulations of stratified populations are also consistent with theory. When $c \neq 0$, such that there is a difference in trait mean between strata, only the local-ancestry and membership adjusted models are valid for the null hypothesis of $a = 0$. Interestingly, the global-ancestry adjusted model also shows correct rejection rates, suggesting that the specification bias discussed in the methods section is small for these examples. When $a > 0$, we show that the local-ancestry adjusted model is valid as a test of the null hypothesis of $W^* = 0$ (or $r_{gl} = 0$) and the membership-adjusted model provides a valid test of no LD in both subpopulations. Rejection rates for the unadjusted model are greater than the nominal rate, even for distant markers. The exceptions are markers in low LD with the QTL in the stratified population ($D^{**}$) and with small differences between marker allele frequencies in the ancestral populations. Power again depends on the relative level of LD in the ancestral populations and in the stratified population.

### 3.3 Estimates of LD in admixed datasets

When testing at a marker, we have seen that the regression coefficient for the genetic effect depends on the genetic effect at the QTL and LD between the QTL and marker genotypes. To put these results into the context of actual admixed populations, we examined the relationship between estimates of LD (disequilibrium coefficient: $D^*$ or $D_i$) and intermarker distance for six admixed Hispanic datasets (Colombian, Honduran, Cuban, Puerto Rican, Dominican and Mexican), an admixed non-Hispanic dataset (African-American), and non-admixed “ancestral” datasets (European, African and Native American). As expected, the estimates of LD tend to be larger (in absolute value) and more variable in the admixed datasets relative to the non-admixed datasets (Supplemental Figure 6). Colombian, Honduran and Dominican datasets show the largest variation in LD estimates. Cuban, Mexican, and African-American datasets show the least variation and are similar to estimates in non-admixed datasets. For all datasets, LD values get closer to 0 as intermarker distance increases. Within each dataset, pairs beyond ~100–200kb demonstrate LD values with a similar distribution to unlinked pairs, but the range remains relatively wide in Colombians, Hondurans and Dominicans, even for unlinked pairs (e.g., 17–20% of unlinked pairs in these admixed datasets have estimates of $|D^*|<0.05$ compared to <4% of unlinked pairs in non-admixed datasets).

We have shown that the key quantities in the formulas for the regression coefficients of the unadjusted and local-ancestry adjusted models (as well as global-adjusted but in a more complex way) are $D^*$, computed directly in an admixed population, and $W$, computed as an average of LD in the ancestral populations weighted by mixing proportions. Simulations also demonstrated that these have strong influences on relative power. Figure 1 compares estimates of $W$ and $D^*$ across varying intermarker distances in Colombians, Cubans, Dominicans and African Americans (results in Hondurans, Puerto Ricans and Mexicans are shown in Supplemental Figure 7). We see as expected a smaller median and tighter distribution (smaller IQR) for $W$ compared to $D^*$ for all admixed populations; however there is variability among populations, with the Dominicans being most strikingly different (e.g., for unlinked markers in Dominicans the median $|D^*|$ is about 3 times the median of $|W|$ 0.023 vs 0.008). Interestingly, the differences in distribution are most apparent for distant markers; for nearby markers (e.g., <~25kb) the distributions of $W$ and $D^*$ are fairly similar for all admixed populations.

## 4. DISCUSSION

Our aim was to better understand the properties of global- and local-ancestry adjustments in linear regression models used for genetic association studies. Specifically, we examined statistical confounding, the forms of regression parameters, and relative power in terms of genetic parameters. Our results regarding confounding are not surprising. When testing a QTL itself in a single admixed population, (global or local) ancestry is not a confounder because, in our model, the QTL explains all of the non-random phenotypic variation. However, in all other scenarios considered (testing at marker loci or in stratified admixed populations), ancestry can be a confounder, and our formulas shed light on when this is the case. Since typically we assume we are not testing the QTL directly, we must consider ancestry a potential confounder, and should adjust for it at some level.

Studying the forms of the regression parameters for unadjusted and adjusted regression models was particularly informative, as it shed light on our implicit assumptions in hypothesis testing. We saw that tests of the genotype coefficient can correspond to tests of different null hypotheses in terms of genetic parameters. In a single admixed population, we have shown that the unadjusted model and the models adjusted for local and global ancestry all provide valid tests for the hypothesis of no genetic effect; however, in the presence of a genetic effect, tests at marker loci depend on LD between the marker and QTL in different ways. The unadjusted model provides a valid test of the null hypothesis of no LD in the admixed population, while the local-ancestry adjusted model provides a valid test of no LD in the ancestral populations. The global-adjusted model will perform similarly to the unadjusted model as recombination increases and similarly to the local-ancestry adjusted model if there has been little recombination.

In a stratified population, only adjusting for local ancestry or subpopulation membership guarantees valid tests of no genetic effect. As in the single admixed population, in the presence of a genetic effect, the different adjustments provide tests of different null hypotheses with respect to LD. The model adjusted for subpopulation membership provides a valid test for no LD in both of the admixed subpopulations, while the local-ancestry adjusted model provides a valid test of no LD in the ancestral populations. The coefficient for the global-ancestry adjusted model ranges between that for the local-ancestry adjusted and population-membership adjusted models, depending on the amount of recombination across the genome. The unadjusted model is generally invalid, even at the QTL itself, since it is sensitive to phenotypic differences between the subpopulations not explained by the QTL.

Our estimates of disequilibrium coefficients in data from actual admixed and “ancestral” populations provide insight into how these different hypothesis tests may behave in real data. Though we cannot go directly from these estimates to statements about power, we can make some useful observations: (1) Values of $W$ (average LD from the ancestral populations) get closer to 0 quite rapidly with increasing distance, suggesting that local-ancestry adjusted models will likely only have power at markers close to the QTL (<~1Mb for our examples); (2) The admixed populations show more extreme values of LD ($D^*$) than the ancestral populations, suggesting that unadjusted (and sometimes global-adjusted models) will have more power in admixed populations than non-admixed populations; (3) The more extreme values of LD seen in the admixed populations persist even between distant and unlinked loci, suggesting there will be a lot of noise at distant markers when using unadjusted tests.

Our results are in agreement with previous studies that considered global vs. local ancestry adjustment in association tests. Like others (Liu et al., 2013; Zhang & Stram, 2014), we find that using global ancestry as a covariate works well to weed out spurious associations at distant loci and control type I error. Also in agreement, we show that the local-ancestry adjusted models will result in tests that generally have lower rejection rates than the global-adjusted model. Our derivation of explicit formulas for the regression coefficients helps us to understand these findings, and in particular understand that these models actually provide tests of different null hypotheses. For example, when testing at a marker in a single-admixed population, we have shown that the unadjusted model provides a test of LD in the admixed population while the local-ancestry adjusted model provides a test of LD in the ancestral populations. Since LD in admixed populations tends to be larger and span larger distances than the LD in the ancestral populations, it makes sense that the unadjusted model has a higher rejection rate. As we show, the coefficient for the global model is intermediate to the unadjusted and local adjusted model, depending on the admixture dynamics (in particular the amount of recombination). We did not explore the scenario posed by Wang *et al* (2011) in which forces such as selection lead to local-ancestry effects at the test marker, but this would appear to be a case for which additional adjustment for local-ancestry may be required. We also did not consider inclusion of a genotype × ancestry interaction term, which Liu *et al* (2013) show can improve power when LD varies in ancestral populations. Examining the forms of this interaction and main effect coefficients under these models will be the subject of future work.

We have made several assumptions to simplify the theoretical calculations. The model of admixture and population evolution assumes only two mixing populations as well as specific pulses of admixture followed by random mating. These assumptions may not be valid for Hispanic populations, which often show three-way admixture. Other assumptions include a known (not estimated) haplotype ancestry and equal trait variance. Perhaps the most unrealistic simplification is that of a single common QTL. We expect variation of quantitative traits to be influenced by multiple genetic and environmental factors as well as their potential interactions. The relationship of these factors to ancestry may influence our conclusions about ancestry adjustment. For example, we have assumed that the trait mean depends only on genotype at the QTL, but if the trait depends on ancestry -specific factors (genetic or environmental), our conclusions may not hold. Such considerations are particularly important as we think about the contributions of multiple rare variants within a gene to a trait. Rare variants are likely to be ancestry-specific, and thus trait means with respect to a particular variant genotype will depend on both the genotype and ancestry of the gene region. This relationship complicates theoretical calculations but is an important topic for future exploration.

We have focused here on analysis of continuous traits, but many studies use a case-control design that focuses on a binary outcome and use logistic regression. There has been interesting discussion on the properties of covariate adjustment in linear regression models versus randomized and retrospective case-control logistic regression models (Mefford & Witte, 2012; Pirinen, Donnelly, & Spencer, 2012; Robinson & Jewell, 1991). In particular, the conclusions with respect to relative power of adjusted and unadjusted models when considering a non-confounding, predictive covariate are similar for randomized studies with continuous or binary traits; but for a case-control design, where selection is based on phenotype, the relative power depends on disease prevalence (Pirinen et al., 2012; Robinson & Jewell, 1991). It is not clear how our findings will generalize to commonly used case-control designs. Extension of our theory to such designs, particularly approximations of relative power, are more difficult since the selection of subjects can induce a bias that must be accounted for in addition to relative variance (Robinson & Jewell, 1991). Nevertheless, such extensions would be interesting and informative.

As practical recommendations based on our results, we suggest that global-adjusted models should be used for initial association analyses. Though we did show that in stratified admixed populations, the parameter can be biased, we demonstrate that this bias is typically small. If one can be relatively certain that the population under study represents a single admixed population, e.g. African American, an unadjusted model may be used for this initial scan as it will be sensitive to LD in the admixed population. However, it is likely that more distant markers with LD driven by ancestral allele frequency differences will also be detected. For fine-mapping and to exclude results due entirely to LD induced by admixture; local-ancestry adjusted models should be used, given that in the presence of a genetic effect, they provide a valid test of LD in the ancestral populations. Our results also suggest that local-ancestry models should also be used to characterize effect size when we believe we have found the underlying QTL. This will more appropriately estimate the true genetic effect if there is population stratification than the global-adjusted or unadjusted model and more closely reflect the effect in non-admixed ancestral populations.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research was supported by grants from the National Institute of General Medical Sciences (NIGMS, R01GM090087 Eden Martin/Carlos Bustamante MPIs) and the National Heart Lung and Blood Institute (NHLBI, R01HL102487 Gary Beecham/David Seo MPIs) of the National Institutes of Health (NIH). Several samples from Puerto Rico and the Dominican Republic were provided by Dr. Richard Mayeux and supported by the Genetic Studies of Alzheimer's disease in Caribbean Hispanics (EFIGA) funded by the National Institute on Aging (NIA) (5R37AG015473, RF1AG015473, R56AG051876). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We would like to thank Dr. Lily Wang for providing useful comments. We also wish to thank the participants in the GOAL and EFIGA studies as well as research and support staff for their contributions to this study.

## References

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Abecasis GR. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. [doi]. DOI: 10.1038/nature15393 [PubMed: 26432245]

Adeyemo AA, Tekola-Ayele F, Doumatey AP, Bentley AR, Chen G, Huang H, Rotimi CN. Evaluation of genome wide association study associated type 2 diabetes susceptibility loci in sub saharan africans. Frontiers in Genetics. 2015; 6:335. [doi]. doi: 10.3389/fgene.2015.00335 [PubMed: 26635871]

Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011; 12 246-2105-12-246. [doi]. doi: 10.1186/1471-2105-12-246

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009; 19(9):1655–1664. [doi]. DOI: 10.1101/gr.094052.109 [PubMed: 19648217]

Armstrong DL, Zidovetzki R, Alarcon-Riquelme ME, Tsao BP, Criswell LA, Kimberly RP, Jacob CO. GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. Genes and Immunity. 2014; 15(6):347–354. [doi]. DOI: 10.1038/gene.2014.23 [PubMed: 24871463]

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Halperin E. Fast and accurate inference of local ancestry in latino populations. Bioinformatics (Oxford, England). 2012; 28(10):1359–1367. [doi]. DOI: 10.1093/bioinformatics/bts144

Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, Montine TJ. Genome-wide association meta-analysis of neuropathologic features of alzheimer's disease and related dementias. PLoS Genetics. 2014; 10(9):e1004606. [doi]. doi: 10.1371/journal.pgen.1004606 [PubMed: 25188341]

Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of african americans, latinos, and european americans across the united states. American Journal of Human Genetics. 2015; 96(1):37–53. [doi]. DOI: 10.1016/j.ajhg.2014.11.010 [PubMed: 25529636]

Burnett MS, Strain KJ, Lesnick TG, de Andrade M, Rocca WA, Maraganore DM. Reliability of self-reported ancestry among siblings: Implications for genetic association studies. American Journal of Epidemiology. 2006; 163(5):486–492. doi:kwj057 [pii]. [PubMed: 16421243]

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Cavalli-Sforza LL. A human genome diversity cell line panel. Science (New York, N.Y.). 2002; 296(5566):261–262.

Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proceedings of the National Academy of Sciences of the United States of America. 1988; 85(23):9119–9123. [PubMed: 3194414]

Cox, TF., Cox, MAA., editors. Multidimensional scaling. Chapman and Hall/CRC; Boca Raton, FL: 2001.

Cruchaga C, Kauwe JS, Harari O, Jin SC, Cai Y, Karch CM, Goate AM. GWAS of cerebrospinal fluid tau levels identifies risk variants for alzheimer's disease. Neuron. 2013; 78(2):256–268. [doi]. DOI: 10.1016/j.neuron.2013.02.026 [PubMed: 23562540]

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics. 2003; 164(4):1567–1587. [PubMed: 12930761]

Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources. 2009; 9(5):1322–1332. [doi]. DOI: 10.1111/j.1755-0998.2009.02591.x [PubMed: 21564903]

Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, Tang H. Ancestral components of admixed genomes in a mexican cohort. PLoS Genetics. 2011; 7(12):e1002410. [doi]. doi: 10.1371/journal.pgen.1002410 [PubMed: 22194699]

Liu J, Lewinger JP, Gilliland FD, Gauderman WJ, Conti DV. Confounding and heterogeneity in genetic association studies with admixed populations. American Journal of Epidemiology. 2013; 177(4): 351–360. [doi]. DOI: 10.1093/aje/kws234 [PubMed: 23334005]

Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. American Journal of Human Genetics. 2013; 93(2):278–288. [doi]. DOI: 10.1016/j.ajhg.2013.06.020 [PubMed: 23910464]

Mefford J, Witte JS. The covariate's dilemma. PLoS Genetics. 2012; 8(11):e1003096. [doi]. doi: 10.1371/journal.pgen.1003096 [PubMed: 23162385]

Melton PE, Carless MA, Curran JE, Dyer TD, Goring HH, Kent JW Jr, Almasy L. Genetic architecture of carotid artery intima-media thickness in mexican americans. Circulation. Cardiovascular Genetics. 2013; 6(2):211–221. [doi]. DOI: 10.1161/CIRCGENETICS.113.000079 [PubMed: 23487405]

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Bustamante CD. Reconstructing the population genetic history of the caribbean. PLoS Genetics. 2013; 9(11):e1003925. [doi]. doi: 10.1371/journal.pgen.1003925 [PubMed: 24244192]

Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, Schellenberg GD. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset alzheimer's disease. Nature Genetics. 2011; 43(5):436–441. DOI: 10.1038/ng.801 [PubMed: 21460841]

Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, Singleton AB. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for parkinson's disease. Nature Genetics. 2014; [doi]. doi: 10.1038/ng.3043

Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WH, Price AL. Enhanced statistical tests for GWAS in admixed populations: Assessment using african americans from CARe and a breast cancer consortium. PLoS Genetics. 2011; 7(4):e1001371. [doi]. doi: 10.1371/journal.pgen. 1001371 [PubMed: 21541012]

Peng B, Kimmel M. simuPOP: A forward-time population genetics simulation environment. Bioinformatics (Oxford, England). 2005; 21(18):3686–3687. doi:bti584 [pii].

Pino-Yanes M, Gignoux CR, Galanter JM, Levin AM, Campbell CD, Eng C, Burchard EG. Genome-wide association study and admixture mapping reveal new loci associated with total IgE levels in latinos. The Journal of Allergy and Clinical Immunology. 2015; 135(6):1502–1510. [doi]. DOI: 10.1016/j.jaci.2014.10.033 [PubMed: 25488688]

Pirinen M, Donnelly P, Spencer CC. Including known covariates can reduce power to detect genetic effects in case-control studies. Nature Genetics. 2012; 44(8):848–851. [doi]. DOI: 10.1038/ng. 2346 [PubMed: 22820511]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38(8):904–909. DOI: 10.1038/ng1847 [PubMed: 16862161]

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945–959. [PubMed: 10835412]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Sham PC. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007; 81(3):559–575. DOI: 10.1086/519795 [PubMed: 17701901]

Robinson, LD., Jewell, NP. International statistical review / revue internationale de statistique. Vol. 59. International Statistical Institute (ISI); 1991. Some surprising results about covariate adjustment in logistic regression models; p. 227-240.

Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics. 2003; 73(6):1402–1422. doi:S0002-9297(07)63990-1 [pii]. [PubMed: 14631557]

Serfling, RJ. Approximation theorems of mathematical statistics. Wiley; 2009.

Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: Analytical and study design considerations. Genetic Epidemiology. 2005; 28(4):289–301. [doi]. DOI: 10.1002/gepi.20064 [PubMed: 15712363]

Wang X, Zhu X, Qin H, Cooper RS, Ewens WJ, Li C, Li M. Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics (Oxford, England). 2011; 27(5):670–677. DOI: 10.1093/bioinformatics/btq709

Zhang J, Stram DO. The role of local ancestry adjustment in association studies using admixed populations. Genetic Epidemiology. 2014; 38(6):502–515. [doi]. DOI: 10.1002/gepi.21835 [PubMed: 25043967]

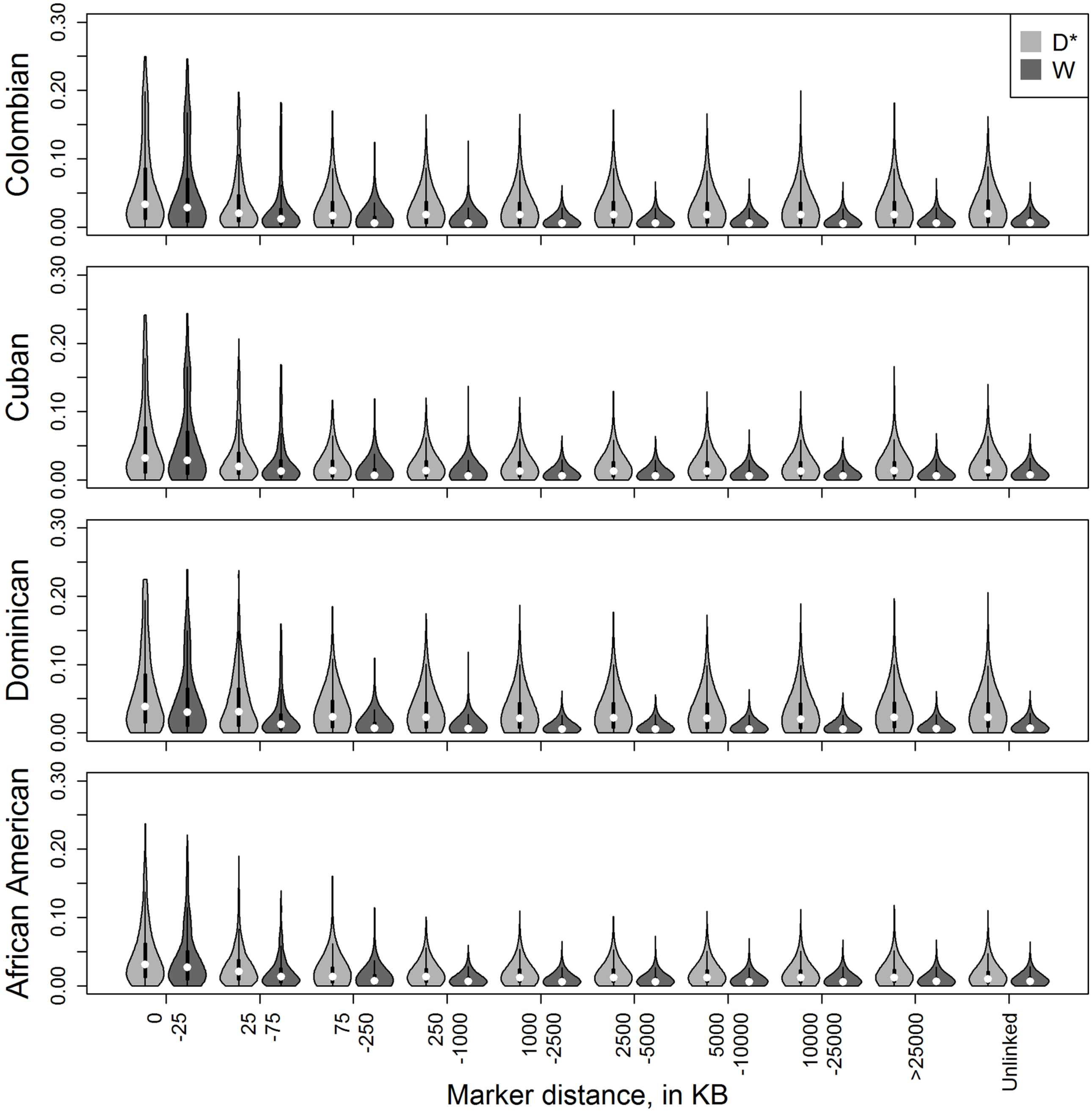

**Figure 1.**

Violin plots for absolute value of estimates of disequilibrium coefficients three admixed datasets. *D** is the estimate in the admixed dataset and *W* is the average of disequilibrium coefficients in the ancestral populations, weighted by estimates of mixing proportions: Colombian (0.68 European, 0.07 African, 0.25 Native American), Cuban (0.81 European, 0.15 African, 0.04 Native American), Dominican (0.70 European, 0.27 African, 0.03 Native American), African American (0.24 European, 0.75 African, 0.01 Native American). Results are binned by intermarker distance, with the final bin being pairs of unlinked markers.

## Table I

Table of Notation

| | | |
|---|---|---|
| <u>**Single Admixed Population**</u> | | |
| **Random Variables** | | |
| $G$ | | QTL Genotype |
| $L$ | | Marker Genotype |
| $Y$ | | Quantitative Trait |
| $A$ | | Local Ancestry |
| $\bar{A}$ | | Global Ancestry |
| **Constants** | | |
| $K$ | | number of chromosomes |
| $m$ | | number of markers per chromosome |
| $M = mK$ | | total number of markers |
| **Parameters** | | |
| $a$ | | genetic effect of QTL |
| $\sigma^2$ | | Within-genotype variance of quantitative trait |
| $q$ | | probability any locus on a random haplotype is from ancestral population 1 |
| $r_{jk}$ | | recombination probability between locus $j$ and $k$ |
| $r_u$ | | recombination probability between loci on different chromosomes |
| $\varphi=1-\left(\frac{K-1}{K}\right)r_U-\frac{1}{\mathrm{Mm}}\sum_{j=1}^{m}\sum_{k\neq j}^{m}r_{\mathrm{jk}};$<br>$\varphi_g=1-\left(\frac{K-1}{K}\right)r_U-\frac{1}{M}\sum_{j\neq g}^{m}r_{\mathrm{gj}};$<br>$\varphi_l=1-\left(\frac{K-1}{K}\right)r_U-\frac{1}{M}\sum_{j\neq l}^{m}r_{\mathrm{lj}}$ | | Functions of recombination probabilities, where $g$ is QTL and $l$ is marker locus |
| <u>*QTL*</u> | <u>*Marker*</u> | |
| $p_i$ | $p_{Li}$ | allele frequency in ancestral population $i$, for $i = 0,1$ |

| | | |
|---|---|---|
| $\Delta = p_1 - p_0$ | $\Delta_L = p_{L1} - p_{L0}$ | ancestral allele frequency difference |
| $p^* = qp_1 + (1 - q)p_0$ | $p_L^* = \mathrm{qp}_{L1} + (1 - q)p_{L0}$ | allele frequency in the admixed population |
| $\gamma = qp_1(1 - p_1) + (1 - q)p_0(1 - p_0)$ | $\gamma_L = qp_{L1}(1 - p_{L1}) + (1 - q)p_{L0}(1 - p_{L0})$ | average variance of allele frequency in the admixed population |
| $D_i = P_{TLi} - p_{Li}p_i$ | | disequilibrium coefficient between marker and QTL alleles in ancestral population *i* |
| $D^* = (1 - r_{gl})(W + q(1 - q)\Delta\Delta_L)$ | | disequilibrium coefficient in admixed population |
| $W = qD_1 + (1 - q)D_0$ | | average disequilibrium coefficient |
| **<u>Stratified Admixed Populations</u>** | | |
| **Random Variables** | | |
| $S$ | | admixed subpopulation membership |
| **Parameters** | | |
| $c$ | | shift in mean quantitative trait value in subpopulation 1 compared to subpopulation 0 |
| $Q$ | | probability of being in subpopulation 1 |
| $q_s$ | | probability any locus on a random haplotype in subpopulation *s* is from ancestral population 1 |
| $\alpha = Qq_1 + (1 - Q)q_0$ | | average ancestry in stratified population |
| $\delta = Q(1 - Q)(q_1 - q_0)^2$ | | covariance of ancestry of different haplotypes at the same locus within an individual |
| $\omega = \alpha(1 - \alpha) - \delta = Qq_1(1 - q_1) + (1 - Q)q_0(1 - q_0)$ | | convenient function of population structure parameters |
| *<u>QTL</u>* | *<u>Marker</u>* | |
| $p_s^* = q_s p_1 + (1 - q_s)p_0$ | $p_{\mathrm{Ls}}^* = q_s p_{L1} + (1 - q_s)p_{L0}$ | allele frequency in subpopulation *s*, for *s* = 0,1 |
| $p' = \mathrm{Qp}_1^* + (1 - Q)p_0^* = \alpha p_1 + (1 - \alpha)p_0$ | $p_L^{'} = \mathrm{Qp}_{L1}^* + (1 - Q)p_{L0}^* = \alpha p_{L1} + (1 - \alpha)p_{L0}$ | allele frequency in stratified population |
| $\gamma' = \alpha p_1(1 - p_1) + (1 - \alpha)p_0(1 - p_0)$ | $\gamma_L{}' = \alpha p_{L1}(1 - p_{L1}) + (1 - \alpha)p_{L0}(1 - p_{L0})$ | average variance of allele frequency in stratified population |
| $D_s^* = (1 - r_{\mathrm{gl}})(W_s + q_s(1 - q_s)\Delta\Delta_L)$ | | disequilibrium coefficient in *s*th subpopulation |

| | |
|---|---|
| $W_s = q_s D_1 + (1 - q_s) D_0$ | average disequilibrium coefficient in subpopulation *s* |
| $D^{**} = \mathrm{QD}_1^* + (1 - Q) D_0^* + Q(1 - Q)(q_1 - q_0)^2 \Delta\Delta_L = (1 - r_{\mathrm{gl}})(W^* + \omega \Delta\Delta_L) + \delta \Delta\Delta_L$ | disequilibrium coefficient in stratified population |
| $W^* = Q W_1 + (1 - Q) W_0$ | average disequilibrium coefficient over subpopulations |

## Table II

Regression models for tests at the QTL and marker adjusted by ancestry or subpopulation membership covariates.

| | Testing QTL | Testing Marker |
|---|---|---|
| **Unadjusted (UN)** | $E(Y\|G) = \beta_0 + \beta_G G$ | $E(Y\|L) = \beta_{L0} + \beta_L L$ |
| **Local-ancestry adjusted (LA)** | $E(Y\|G, A_{\cdot g}) = \beta_0^* + \beta_G^* G + \beta_A^* A_{\cdot g}$ | $E(Y\|L, A_{\cdot l}) = \beta_{L0}^* + \beta_L^* L + \beta_{\mathrm{LA}}^* A_{\cdot l}$ |
| **Global-ancestry adjusted (GA)** | $E(Y\|G, \overline{A}) = \beta_0' + \beta_G' G + \beta_{\overline{A}}' \overline{A}$ | $E(Y\|L, \overline{A}) = \beta_{L0}' + \beta_L' L + \beta_{\mathrm{L}\overline{\mathrm{A}}}' \overline{A}$ |
| **Subpopulation-membership adjusted (MA)** | $E(Y\|G, S) = \beta_0'' + \beta_G'' G + \beta_S'' S$ | $E(Y\|L, S) = \beta_{L0}'' + \beta_L'' L + \beta_{\mathrm{LS}}'' S$ |

### Table III

Regression coefficients for the genotype term in single admixed and stratified admixed populations. Regression models are the unadjusted model (UN) and models adjusted for local ancestry (LA), global ancestry (GA), and subpopulation membership (MA) (only for stratified admixed populations).

| **Model** | **Regression Coefficient** |
|---|---|
| | *Single Admixed Population* |
| ***At QTL*** | |
| **UN** | $\beta_G = a$ |
| **LA** | $\beta^*_G = a$ |
| **GA** | $\beta'_G = a$ |
| ***At Marker*** | |
| **UN** | $\beta_L = \frac{\mathrm{aD}^*}{q(1-q)\Delta^2_L + \gamma_L}$ |
| **LA** | $\beta^*_L = \frac{a(1-r_{\mathrm{gl}})W}{\gamma_L}$ |
| **GA** | $\beta'_L = \frac{a\left[D^* - q(1-q)\Delta\Delta_L \frac{\varphi_g \varphi_l}{\varphi}\right]}{q(1-q)\Delta^2_L \left(1 - \frac{\varphi^2_l}{\varphi}\right) + \gamma_L}$ |
| | *Stratified Admixed Population* |
| ***At QTL*** | |
| **UN** | $\beta_G = a + \frac{cQ(1-Q)(q_1 - q_0)\Delta}{(\omega + 2\delta)\Delta^2 + \gamma'}$ |
| **LA** | $\beta^*_G = a$ |
| **GA** | $\beta'_G = a + c\frac{Q(1-Q)(q_1 - q_0)\omega\Delta \frac{(\varphi - \varphi_g)}{(\omega\varphi + 2\delta)}}{\left[(\omega + 2\delta) - \frac{(\omega\varphi_g + 2\delta)^2}{(\omega\varphi + 2\delta)}\right]\Delta^2 + \gamma'}$ |
| **MA** | $\beta''_G = a$ |
| ***At Marker*** | |
| **UN** | $\beta_L = \frac{a(D^{**} + \delta\Delta\Delta_L) + \mathrm{cQ}(1 - Q)(q_1 - q_0)\Delta_L}{(\omega + 2\delta)\Delta^2_L + \gamma'_L}$ |
| **LA** | $\beta^*_L = \frac{a(1-r_{\mathrm{gl}})W^*}{\gamma'_L}$ |

| **Model** | **Regression Coefficient** |
|---|---|
| **GA** | $\beta'_L=\frac{a\left[(D^{**}+\delta\Delta\Delta_L)-\Delta\Delta_L\frac{(\omega\varphi_g+2\delta)(\omega\varphi_l+2\delta)}{(\omega\varphi+2\delta)}\right]+cQ(1-Q)(q_1-q_0)\Delta_L\omega\frac{(\varphi-\varphi_l)}{(\omega\varphi+2\delta)}}{\Delta_L^2\left[(\omega+2\delta)-\frac{(\omega\varphi_l+2\delta)^2}{(\omega\varphi+2\delta)}\right]+\gamma'_L}$ |
| **MA** | $\beta''_L=\frac{a(D^{**}-\delta\Delta\Delta_L)}{\omega\Delta_L^2+\gamma'_L}$ |