Association studies among admixed populations pose many challenges including confounding of genetic effects due to population substructure and heterogeneity due to different patterns of linkage disequilibrium (LD). We use simulations to investigate controlling for confounding by indicators of global ancestry and the impact of including a covariate for local ancestry. In addition, we investigate the use of an interaction term between a single-nucleotide polymorphism (SNP) and local ancestry to capture heterogeneity in SNP effects. Although adjustment for global ancestry can control for confounding, additional adjustment for local ancestry may increase power when the induced admixture LD is in the opposite direction as the LD in the ancestral population. However, if the induced LD is in the same direction, there is the potential for reduced power because of overadjustment. Furthermore, the inclusion of a SNP by local ancestry interaction term can increase power when there is substantial differential LD between ancestry populations. We examine these approaches in genome-wide data using the University of Southern California's Children's Health Study investigating asthma risk. The analysis highlights rs10519951 ($P = 8.5 \times 10^{-7}$), a SNP lacking any evidence of association from a conventional analysis ($P = 0.5$).

confounding; genetic association studies; genome-wide association studies; heterogeneity; linkage disequilibrium; population stratification

Abbreviations: LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.
local ancestries may be required for genome-wide association scans in admixed populations (35–39).

When testing genetic markers that are proxies for a disease causal locus, the differential LD within admixed populations can result in heterogeneity of effect estimates by local ancestry. If reliable self-identified ethnicity is available, inclusion of an interaction term between the single-nucleotide polymorphism (SNP) and ethnicity may account for the heterogeneity, and subsequent stratified analyses could be performed. However, it is unclear how appropriate this approach may be for admixed populations with variation in local ancestry or when combining an admixed population with others.

In this paper, we use graphical diagrams to clarify the mechanisms by which admixture can lead to confounding and how heterogeneity in effect estimates may arise. On the basis of these mechanisms, we investigate the source and effect of confounding and test for heterogeneity via an interaction term between SNP and local ancestry through simulations. Across all models and simulation scenarios, we focus on effect estimation, type I error, and power. Finally, we apply these models to a genome-wide association study investigating the impact of genetic variation on asthma in the University of Southern California’s Children’s Health Study. We discuss the overall impact of global ancestry on this analysis and identify several empirical examples where accounting for local ancestry impacts inference.

**MATERIALS AND METHODS**

**Graphical model**

The graphical model in Figure 1A represents the relationship of several factors involved in genetic association studies among admixed populations (40, 41). Here, \( Y \) represents the outcome of interest. \( G_M \) represents the SNP at a marker being tested for association with \( Y \) (with effect \( \beta_{G_M} \)). \( G_D \) represents an unmeasured causal locus (with effect \( \beta_{G_D} \)) for which we are testing \( G_M \) as a proxy. \( X \) represents other causal factors that are associated with global ancestry (\( Q \)), including unmeasured environmental factors and/or unmeasured causal loci. The global ancestry is most often estimated from a subset set of markers (42 and/or unmeasured causal loci). The global ancestry is most often estimated from a subset set of markers (42 and/or unmeasured causal loci). The global ancestry \( (Q) \) is correlated with local ancestry \( (P) \). We assume that the local ancestry at \( G_M \) and \( G_D \) are the same, and that there are additional SNPs (\( G_L \)) that can be used to estimate local ancestry for each subject at each location.

There are paths between factors that together may lead to confounding for the relationship between the marker \( G_M \) (with effect \( \beta_{G_M} \)). \( G_D \) acts as an effect modifier of the association between \( G_M \) and \( Y \) if we are testing the disease locus \( G_D \) or even the marker locus \( G_M \).

There are 2 components that affect the magnitude of the LD between \( G_M \) and \( G_D \) in an admixed population: LD within parental populations \( (D'_L) \) and the admixture LD \( (G') \) induced by differential frequencies between ancestral populations at both the marker and the disease locus. As shown in Figure 1B, the admixture LD is indicated as the path \( G_M-L-G_D \), and the LD within parental populations is indicated as \( D'_L \). As indicated in Figure 1B, the reference alleles for \( G_M \) and \( G_D \) are determined, so that the correlation between these 2 loci is positive within the admixed population. Similarly, a reference local ancestry population for \( L \) can be defined, such that \( L \) is positively correlated with \( G_M \) in the admixed population. Thus, the reference allele is the same in both parental populations. Given these reference definitions, when \( L \) is negatively correlated with \( G_D \) (left panel), there exists an overall negative correlation between \( G_M \) and \( G_D \) through the path \( G_M-L-G_D \). In this situation, the admixture LD is in a different direction to the LD in the parental populations. This results in a corresponding reduction in the observed magnitude of the LD in the admixed population. In contrast, when \( L \) is positively correlated with \( G_D \), there is an overall positive correlation between \( G_M \) and \( G_D \) through the path \( G_M-L-G_D \) (right panel). In this situation, the admixture LD is in the same direction as the LD in the parental populations, and the observed LD between \( G_M \) and \( G_D \) in the admixed population is enhanced. In addition to the scenarios discussed above, when the LD in the 2 ancestral populations is in opposite directions, the admixture LD will always enhance the LD in one ancestry while reducing the level of LD in the other. In summary, for a marker \( G_M \), admixture LD has the potential to act as an additional confounder of the \( G_M-Y \) effect. For a disease locus \( G_D \), there is no such potential.

Finally, individual local ancestries may modify the marginal effect at the marker locus because of the differential LD existing across ancestral populations. That is, within a study population, the level of association between \( G_M \) and \( Y \) varies across individuals as a function of \( L \) (e.g., \( D'_L \neq D'_L \)). Thus, \( L \) acts as an effect modifier of the association between \( G_M \) and \( Y \).

**Models**

We use the following generalized linear models to investigate the efficiency of controlling for confounding by global ancestry and the potential impact on power by adjusting for local ancestry:

\[
g(\mu_Y) = \alpha + \beta_{G_M}G_M \\
g(\mu_Y) = \alpha + \beta_{G_M}G_M + \beta_Q Q \\
g(\mu_Y) = \alpha + \beta_{G_M}G_M + \beta_Q Q + \beta_L L
\]

Specifically, \( g(\mu_Y) \) is a logit link for a dichotomous outcome \( Y \) conditional on the covariates included in the model, although alternative outcomes can be handled in a similar manner in the generalized linear framework. \( G_M \)
represents the number of variant alleles for each individual, and $b_{GM}$ is the corresponding marginal effect. A Wald or likelihood ratio test of $b_{GM} = 0$ can be used to test association. For investigating heterogeneity, we extend model 3 to include a $G_M \times L$ interaction term:

$$g(\mu_Y) = \alpha + b_{GM}G_M + b_LL + b_{int}GML + \beta_0Q.$$ (4)

Here, we use a 2-df likelihood ratio test for a joint test of $b_{GM} = 0$ and $b_{int} = 0$. This joint test has been shown to be nearly optimal across many different scenarios for main and interacting effects (48).

**Simulations**

We conduct simulations to investigate the performance of the models defined above (refer to the Web Appendix, Simulations section, available at http://aje.oxfordjournals.org/, for more details). Simulations are based on the framework represented in the graphical model (Figure 1). We simulate data based on the confounding paths and assess the type I error and power after adjusting for individual ancestries. To test the gain in power as well as the potential overadjustment by local ancestry, we include simulation scenarios that model the admixture LD. In addition, we simulate data with and without LD differences between populations to gauge the impact of heterogeneity between ancestries. In all scenarios, we generate cases for a binary disease outcome ($Y$) using a logistic regression model incorporating the disease locus $G_D$ and the individual global ancestry $Q$, with a 50% average probability for being a case. For simplicity, we assume a direct relationship of $Q$ to $Y$.

![Figure 1. A) Potential confounding paths in genetic association studies among admixed populations. $Y$ represents the outcome of interest; $G_M$, the single-nucleotide polymorphism at a marker locus being tested for association; $L$, the individual local ancestry in the immediate neighborhood of the marker locus; $Q$, the individual global ancestry averaged through $L$ across the genome; $X$ represents other causal factors, either unmeasured environmental factors or unmeasured causal loci present across the genome, that may be associated with global ancestry; and $G_L$, the immediate neighborhood of the marker locus that is used to estimate individual local ancestry $L$. Squares represent observed variables. Solid circles represent unobserved variables that can be estimated. Dashed circles represent unobserved variables that cannot be estimated. B) Directions of admixture linkage disequilibrium (LD) and the LD ($D'$) in the parental populations. Admixture LD is represented as the path $G_M \leftrightarrow L \leftrightarrow G_D$.](http://aje.oxfordjournals.org/)

**Table 1. Simulated Scenarios A–C**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Allele Frequency Difference$^a$</th>
<th>SNP Effect, $\beta_{GM}$</th>
<th>$Q$ Effect, $\beta_0$</th>
<th>$D'$ Difference (Heterogeneity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0, 0.4</td>
<td>None</td>
<td>Log(3.0)</td>
<td>None</td>
</tr>
<tr>
<td>B</td>
<td>0, 0.4</td>
<td>Log(1.2)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>C</td>
<td>None</td>
<td>Log(1.2)</td>
<td>Log(3.0)</td>
<td>0, 1.8</td>
</tr>
</tbody>
</table>

Abbreviation: SNP, single-nucleotide polymorphism.

$^a$ Allele frequencies at both the disease and the marker loci.
Scenarios

Across all scenarios (Table 1), we vary the population-specific parameters for each ancestral population (p, q, and \( D_0 \)) and the causal model parameters \( \beta_Q \) and \( \beta_{GD} \) (refer to the Web Appendix, Scenarios section, for more details).

In scenario A, there is no genetic causal effect \( (\beta_{GD} = \log(1)) \) but a strong global ancestry effect \( (\beta_Q = \log(3.0)) \). We simulate different allele frequencies between ancestral populations to investigate the efficiency of control for confounding by individual ancestries in models 2 \( (Y \sim G + Q) \) and 3 \( (Y \sim G + Q + L) \). In scenario B, we simulate admixture LD in the same and different directions to the LD in the original ancestral populations. Finally, in scenario C, the \( D' \) difference between populations varies from 0 to 1.8 \( (D'_1 = 0.9, D'_2 \) varies from −0.9 to 0.9) to gauge the impact of heterogeneity.

For each simulated scenario, we create 1,800 individuals with an equal number of individuals \( (N_L = 600) \) within each local ancestry group \( (L = \{0, 1, 2\}) \), where \( L \) indicates the number of copies from ancestral population 1). Conditional on \( L \) and the corresponding specified parameters for allele frequency, LD, and risk, we generate \( G_D, G_M \) and \( Q \). We then probabilistically generate case status for all 1,800 individuals using a logistic regression model incorporating the disease locus \( G_D \) and the individual global ancestry \( Q \). Variables in the logistic regression model are mean centered, and there is a baseline risk of 50%, thus resulting in approximately equal numbers of cases and controls for each replicate. This simulation framework does not directly simulate potential confounding or heterogeneity by \( L \). Rather, potential confounding and heterogeneity are induced by simulating haplotypes, global ancestry, and diseases status conditional on local ancestries as reflected in our graphical framework (Figure 1). Specifically, potential confounding is induced via the path, \( G_M \rightarrow L \rightarrow Q \rightarrow Y \). The type I error and empirical power are calculated as the number of significant tests \( (\alpha = 0.05) \) over 10,000 replicates.

University of Southern California Children’s Health Study

The Children’s Health Study is an ongoing cohort study investigating environmental and genetic influences on childhood asthma (49–51). The Children’s Health Study genome-wide association study is a nested case-control study from the ongoing longitudinal Children’s Health Study cohort, with approximately equal numbers of cases and controls for non-Hispanic whites and Hispanics. All Children’s Health Study subjects and their parents gave informed consent, and the study was approved by the University of Southern California Institutional Review Board. In this study, we include a total of 2,839 samples from 2 self-reported ethnic groups: 1,396 non-Hispanic whites and 1,171 Hispanics. Among non-Hispanics samples, there are 595 cases and 801 controls; and there are 532 cases and 639 controls among Hispanics. We analyze the Children’s Health Study data stratified by ethnicity and in a combined sample, assuming that the non-Hispanic white individuals all have 2 copies of European local ancestry at each location. Genotyping of these samples was performed at the
University of Southern California Genome Center utilizing both the HumanHap550 (Illumina, Inc., San Diego, California) and the Human 610-Quad BeadChips (Illumina), and the analysis was conducted on 437,599 autosomal SNPs passing a stringent quality-control procedure. We performed several genome-wide scans on the Children’s Health Study samples with additional covariates (age, gender, community of residence, and self-reported ethnicity). We estimated individual local ancestry among Hispanic samples through HAPMIX (52) (refer to the Web Appendix, Local ancestry estimation section, for more details). The average of all local ancestry estimates across the genome for each individual was used to estimate global ancestry (\( \tilde{Q} = \frac{\sum_{m} L_{im}}{2M} \)). The estimated global ancestry is then compared with the more commonly used approach of estimating global ancestry by use of selected ancestry informative markers (42–44, 46) and the STRUCTURE program (29, 53–55) (refer to the Web Appendix, Global ancestry estimation section, for more details).

### RESULTS

#### Simulations

In scenario A (Figure 2), at the marker locus \( G_M \), when the allele frequency difference between ancestral populations is greater than 0.1 (account to 64% of the SNPs in the ENCODE regions as shown in Web Figure 1), the crude model has a substantially elevated type I error rate, while models 2 (\( Y \sim G + Q \)) and 3 (\( Y \sim G + Q + L \)) efficiently control for the confounding. The pattern is the same when testing the disease locus \( G_D \). In scenario B, there is no confounding path simulated, and all models have the correct test size (not shown). Reflecting these patterns, adjustment by global ancestry (when needed) results in an unbiased effect estimate. In contrast, there is very little impact on the effect estimate from adjustment with local ancestry. However, when the induced LD due to differential allele frequency between ancestries is in the same direction as the LD from the parental ancestries (Figure 3A), adjustment for local ancestries results in a slight loss in power. When the induced LD is in the opposite direction to the LD from the parental ancestries (Figure 3B), additional adjustment of local ancestry results in a slight increase in power. However, this decrease/increase in power is negligible for allele frequency differences of less than 0.1. When testing the disease variant directly, as opposed to a marker, we found that the pattern is the same as that shown in Figure 3A.

Figure 4 shows the performance of inclusion of the \( G_M \times L \) interaction term by comparing model 3 (\( Y \sim G + Q + L \)) and model 4 (\( Y \sim G + L + GL + Q \)). For tests at the marker locus \( G_M \), when the LDs between \( G_M \) and \( G_D \) are similar in the ancestral populations (\( D' < 0.7 \)), there is a reduction in power for the 2-df test of the joint effect of \( G_M \) and \( G_M \times L \) from model 4. As the difference in LD increases (\( D' \geq 0.7 \)), model 4 has similar or greater power than model 3. For testing a disease variant (\( D' \) difference = 0), the loss in power is 0.10 for the simulated scenario.

![Figure 4. Comparison of power in scenario C when there is heterogeneity due to differential linkage disequilibrium between ancestries.](image-url)

![Figure 5. Children’s Health Study analysis results across models \( Y \sim G + Q + L \) (top) and \( Y \sim G + L + Gl + Q \) (2 df) (bottom) for combined non-Hispanic white and Hispanic samples.](image-url)
Table 2. P Value and Effect Estimate for Selected Markers Across Ethnic Groups and Models in the Ongoing Children’s Health Study

<table>
<thead>
<tr>
<th>Marker</th>
<th>Allele</th>
<th>Local Strataa</th>
<th>Non-Hispanic Whites</th>
<th></th>
<th>Hispanics</th>
<th></th>
<th>Combinedb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Model 3c</td>
<td></td>
<td>Model 3c</td>
<td></td>
<td>Model 3c</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No. Allele Frequency</td>
<td>β0u</td>
<td>P Value</td>
<td>β0u</td>
<td>P Value</td>
</tr>
<tr>
<td>rs10119122</td>
<td>G</td>
<td>L= 0</td>
<td>12</td>
<td>0.46</td>
<td>0.03</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L= 1</td>
<td>410</td>
<td>0.60</td>
<td>-0.17</td>
<td>-0.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L= 2</td>
<td>721</td>
<td>0.64</td>
<td>-0.37</td>
<td>-0.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>1,143</td>
<td>0.62</td>
<td>-0.30</td>
<td>-0.30</td>
<td></td>
</tr>
<tr>
<td>rs10519951</td>
<td>T</td>
<td>L= 0</td>
<td>119</td>
<td>0.34</td>
<td>-1.12</td>
<td>-1.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L= 1</td>
<td>531</td>
<td>0.26</td>
<td>-0.48</td>
<td>-0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L= 2</td>
<td>510</td>
<td>0.19</td>
<td>0.16</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>1,160</td>
<td>0.24</td>
<td>-0.29</td>
<td>-0.29</td>
<td></td>
</tr>
</tbody>
</table>

a Estimated individual local ancestry. L is rounded up into 3 categorical groups: 0 (L ≤ 0.5); 1 (L > 0.5–1.5); and 2 (L > 1.5).

b Combined non-Hispanic white and Hispanic samples for analysis.

c Sample size within each local stratum.

d Effect estimate \( \beta_{0u} \) of the single-nucleotide polymorphism marginal effect followed by the corresponding P value from model 3: \( Y + G + Q \).

e The expected effect estimate \( \beta_{0u} \) of the single-nucleotide polymorphism effect within each local stratum followed by the 2-df test P value from model 4: \( Y + G + L + G_i + Q \).
only analysis within the strata of individuals carrying 0 copies of European ancestry, the estimate is \( \beta_{G_w} = -1.12 \). The contrast in estimates across strata is reflected in the more significant result from model 4 in the combined sample (\( P = 8.5 \times 10^{-7} \)).

**DISCUSSION**

When confounding arises through global ancestry via a path that links an external factor to the marker being evaluated, then global ancestry alone can control for the confounding. This assumes that the estimated global ancestry accurately captures the underlying factor. Previous studies have argued that adjustment for local ancestry is necessary for controlling for confounding (35, 36, 38); however, these papers simulated local ancestry as a strict confounder and did not allow for induced LD in admixed populations. Our simulations demonstrate that impact of adjustment for local ancestry is more nuanced within admixed populations. When the direction of the admixture LD is in a different direction from the LD in the parental ancestries, there is a reduction in the magnitude of the LD in the admixed population, and additional adjustment for local ancestry can increase the power to detect the true association at the marker locus. This potential gain in power, however, comes with risk: When the admixture LD is in the same direction as the LD within the ancestral population, adjustment for local ancestry will result in overadjustment. As shown in Web Figure 2, induced LD and the LD within the ancestral populations are in the same direction for about 57% of the loci in the ENCODE regions (36) (refer to the Web Appendix, ENCODE regions section, for more details).

When investigating heterogeneity-of-effect estimates by local ancestry, we found that there is also a potential loss in power by testing both the SNP main effect and the interaction via a 2-df test (Figure 4). In the ENCODE regions, about 30% of the estimated differences in \( D' \) between the populations are greater than 0.7 (Web Figure 3). Our simulation results demonstrate that model 4 with the SNP–local ancestry interaction has greater power than the conventional model for \( D' \) differences above 0.7. Thus, one may expect an increase in power for about 30% of the SNPs, with the remaining 70% having none or a slight reduction in power. Given this tradeoff, we believe that a genome-wide association study for discovery using only model 4 may not be the most advantageous approach. However, in practice, most investigators will first perform an analysis without an interaction term. Subsequent analyses with the interaction term included offer the potential to uncover previously unidentified regions. In such a 2-step approach, one would need to consider the impact on type I error, but for discovery and further follow-up, such impact may be negligible.

For the Children’s Health Study genome-wide association study, the most notable SNP (rs10119122) from the marginal significance levels may depend on the specific admixed population investigated, because the distribution of local ancestry for each individual across all locations in the genome will depend upon the sample. In this case, a permutation test for determining significance may be required. Whether the top SNPs are strictly significant or not, there is...
clear potential for additional information to be gained from including local ancestry in a test of heterogeneity. Overall for the Children’s Health Study and consistent with results from ENCODE, the interaction model results in smaller $P$ values for 35% of the SNPs across the genome. Notably, for those SNPs with a smaller $P$ value from model 4, this change is often substantial, suggesting that a great deal of additional information may be captured by jointly considering the main and interaction terms.

When testing the disease variant, we find that adjustment for local ancestry most often results in a loss of power from overadjustment when the allele frequency is different between ancestries. Likewise, when investigating a measured causal variant in admixed population, we find that there will be no influence of differential LD between the marker and the causal variant. Thus, the inclusion of an SNP by local ancestry interaction term will not capture any additional information, and stratified estimates across local ancestry strata should be similar. This offers a potential approach to leverage differential LD patterns in an admixed population to help identify causal variants when performing fine-scale mapping or sequencing studies (57).

In addition to capturing the heterogeneity of the SNP effect among admixed populations, it is possible that this observed effect is induced by another genetic or environmental factor that drives the observed effect modification and is correlated with self-identified ethnicity and, thus, related to local ancestry via global ancestry (i.e., $XQ-L$). In order to investigate the source (environmental or genetic) of the heterogeneity, one can perform further analyses within strata by individual global ancestry and, if available, the strata of self-reported ethnicity. For SNP rs10519951, Table 3 shows that the heterogeneity captured by local ancestry is attenuated when stratifying by global ancestry or self-identified ethnicity. These results suggest that this particular observed heterogeneity is most likely due to local genetic structure and not global genetic or environmental differences.

We have demonstrated that one needs to consider the impact of adjustment by local ancestry in addition to the common practice of adjusting for global ancestry. Although the adjustment with local ancestry reflects the induced admixture LD within admixed populations, the impact of inclusion of local ancestry depends upon the LD patterns in the ancestral populations. Furthermore, we have also demonstrated the potential for a 2-df test of SNP main effect and SNP by local ancestry interaction to increase power when there is substantial differential LD between ancestry populations. We realize that, for most genome-wide association studies utilizing admixed populations, investigators will first scan the genome with a marginal test of association. Thus, we view analyses with the interaction term as secondary follow-up to uncover previously unidentified regions with substantial heterogeneity of SNP effect by local ancestry.

ACKNOWLEDGMENTS

Author affiliations: Department of Preventive Medicine, University of Southern California, Los Angeles, California (Jinghua Liu, Juan Pablo Lewinger, Frank D. Gilliland, W. James Gauderman, David V. Conti).

This work is supported by grants R01ES016813, P30ES007048, P01ES011627, and P30ES007048 from the National Institute of Environmental Health Sciences and by grant R01HL087680 from the National Heart, Lung, and Blood Institute.

Conflict of interest: none declared.

REFERENCES


