Adjustment for local ancestry in genetic association analysis of admixed populations

Xuexia Wang1, Xiaofeng Zhu2, Huaizhen Qin2, Richard S. Cooper3, Warren J. Ewens4, Chun Li5,6,* and Mingyao Li1,∗

1Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, 2Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine, Cleveland, OH 44106, 3Department of Preventive Medicine and Epidemiology, Loyola University Chicago School of Medicine, Maywood, IL 6015, 4Department of Biology, University of Pennsylvania, PA 19104, 5Department of Biostatistics and 6Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

ABSTRACT

Motivation: Admixed populations offer a unique opportunity for mapping diseases that have large disease allele frequency differences between ancestral populations. However, association analysis in such populations is challenging because population stratification may lead to association with loci unlinked to the disease locus.

Methods and results: We show that local ancestry at a test single nucleotide polymorphism (SNP) may confound with the association signal and ignoring it can lead to spurious association. We demonstrate theoretically that adjustment for local ancestry at the test SNP is sufficient to remove the spurious association regardless of the mechanism of population stratification, whether due to local or global ancestry differences among study subjects; however, global ancestry adjustment procedures may not be effective. We further develop two novel association tests that adjust for local ancestry. Our first test is based on a conditional likelihood framework which models the distribution of the test SNP given disease status and estimated discrete subpopulations and then aggregates evidence of association within each subpopulation. The current state-of-the-art contrast, structured association methods assign the study subjects to three categories: (i) genomic control (Devlin et al., 1999), (ii) structured association (Pritchard et al., 2000; Satten et al., 2001) and (iii) principal components-based methods (Epstein et al., 2007; Li et al., 2010; Price et al., 2006; Zhang et al., 2003; Zhu et al., 2002). The genomic control approach attempts to correct for population stratification by adjusting association statistics with a single overall inflation factor obtained from a set of random markers that are not associated with the phenotype of interest. In contrast, structured association methods assign the study subjects to estimated discrete subpopulations and then aggregates evidence of association within each subpopulation. The current state-of-the-art approach is the principal component-based approach implemented in EIGENSTRAT (Price et al., 2006), which computes coefficients of principal components using SNPs across the genome to control for population structure.

With the increasing availability of large volumes of high-density SNP genotyping data generated in genome-wide association studies (GWAS), the analysis of admixed populations is now moving toward SNP association. In admixed populations, the proportion of admixture may vary across individuals. This variation can lead to associations of the disease with loci unlinked to the disease locus, a phenomenon well known as ‘population stratification’, which can produce both false-positive and false-negative association signals if not appropriately controlled.

Over the past decade, various methods have been developed to deal with this population stratification effect (Price et al., 2010; Sillanpaa, 2010). In general, these methods can be classified into three categories: (i) genomic control (Devlin et al., 1999), (ii) structured association (Pritchard et al., 2000; Satten et al., 2001) and (iii) principal components-based methods (Epstein et al., 2007; Li et al., 2010; Price et al., 2006; Zhang et al., 2003; Zhu et al., 2002). The genomic control approach attempts to correct for population stratification by adjusting association statistics with a single overall inflation factor obtained from a set of random markers that are not associated with the phenotype of interest. In contrast, structured association methods assign the study subjects to estimated discrete subpopulations and then aggregates evidence of association within each subpopulation. The current state-of-the-art approach is the principal component-based approach implemented in EIGENSTRAT (Price et al., 2006), which computes coefficients of principal components using SNPs across the genome to control for population structure.

INTRODUCTION

African Americans and Hispanic Americans represent the two largest racial minority groups in the USA, comprising ~28% of the US population. Both groups are recently admixed and have inherited ancestry from more than one ancestral population. Such admixed populations offer a unique opportunity for mapping disease genes that have large allele frequency differences between ancestral populations. Recently, admixture mapping has become one of the main approaches for gene mapping studies in admixed populations (Hoggart et al., 2004; McKeigue, 1998; Montana and Pritchard, 2004; Patterson et al., 2004; Zhang et al., 2004; Zhu et al., 2004). However, traditional admixture mapping has a substantially lower resolution than association analysis (Smith et al., 2005; Zhu et al., 2008). Moreover, admixture mapping cannot identify disease loci that have similar allele frequencies or disease prevalences in the ancestral populations.

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
The central idea of all the above-mentioned methods is to use markers across the genome to capture the global population structure within the study subjects. Variation in global population structure is driven mainly by demographic histories such as migration and genetic random drift. However, for admixed populations, a person’s genome is a mosaic of ancestral chromosomes. This introduces considerable amount of variation in local ancestry at certain genomic regions, and the local ancestry may have little correlation with global ancestry. Local ancestry has been useful in localizing disease susceptibility genes in admixture mapping studies (Hoggart et al., 2004; McKeigue, 1998; Montgomery and Pritchard, 2004; Patterson et al., 2004; Zhang et al., 2004; Zhu et al., 2004). In this article, we show that local ancestry may confound with the association signal at the test SNP, and if ignored, can lead to spurious association in genetic association analysis. Methods that adjust for global ancestry may fail to remove this spurious association because the global ancestry information obtained from all markers across the genome may not accurately reflect the amount of ancestry variation in local regions.

There are two types of mechanisms that may lead to population stratification: (i) stratification due to local ancestry difference driven by natural selection at certain genomic regions and (ii) stratification due to global ancestry difference driven by the demographic history of a population or genetic random drift due to finite population size. The existing methods for population stratification correction are appropriate when the stratification is due to global ancestry difference; however, they may be ineffective in removing the effect of population stratification when the stratification is induced by natural selection that occurs only in certain genomic regions. Population stratification due to natural selection is not uncommon. A well-known example is the lactase gene, which has been shown to be under recent positive selection (Bersaglieri et al., 2004). Since both the lactase gene ancestry and height track with northwest versus southeast European ancestry, naïve association analysis between the lactase gene and height will lead to a spurious association (Campbell et al., 2005). Another example of recent natural selection was reported in admixed Puerto Ricans (Tang et al., 2007). In this study, three chromosomal regions, including the human leukocyte antigen region on 6p, 8q and 11q, were reported to exhibit deficiencies at the test SNP. A key advantage of this test lies in its ability to incorporate different directions of association in the ancestral populations, a ‘flip-flop’ phenomenon due to either population differences in allele frequencies, or to multi-locus effects and variation in inter-locus correlation (Hoggart et al., 2007). Our second test, denoted as Logistic-Local, is based on a simple logistic regression in which the test SNP is a predictor and the estimated local ancestry proportion at the test SNP is included as a covariate in the analysis. Compared with the first test, this test imposes additional assumptions on the disease risk model but is computationally simpler. Our procedures are developed as general tests of genetic association for admixed populations and can be applied in the GWAS setting. In contrast to a requirement in admixture mapping, neither of our tests requires the allele frequencies or the disease prevalences to be different in the ancestral populations, and thus they can identify disease loci that may be missed by admixture mapping.

We conducted extensive simulations to evaluate the performance of the proposed tests. Our results indicate that regardless of the mechanism of population stratification—whether due to local or global ancestry differences—the Type I error rates of our procedures were always under control; however, global ancestry adjustment procedures such as EIGENSTRAT (Price et al., 2006) may fail to control Type I error rates when population stratification is due to local ancestry differences.

2 METHODS

2.1 Notation

We assume that admixture has occurred between two ancestral populations, denoted by X and Y, in a recently admixed population. Assume a set of markers with known genetic locations is available for estimating ancestry. Consider a test SNP with alleles A and a (with frequencies \( p_X \), \( q_X \) and \( p_Y \), \( q_Y \) in populations X and Y, respectively). Let \( \pi \) be the probability that a randomly selected allele at the test SNP comes from population X. For each individual, let Z be the disease status (1 = affected; 0 = unaffected), \( I_{Z=0} \) (0, 1, 2) denote the number of alleles at the test SNP that come from population X, \( G_{SNP} \) (0, 1, 2) denote the number of allele A at the test SNP, and \( G_{SRC} \) denote the flanking marker genotypes that are used to infer the local ancestry at the test SNP.

2.2 Confounding due to local ancestry

We now show that local ancestry at the test SNP can confound with the association signal in genetic association analysis. The local ancestry at the test SNP can be considered as a ‘dialectic marker’ with alleles ‘X’ and ‘Y’. There are four marker-ancestry haplotypes, AX, aX, AY and aY, with respective frequencies \( P(AX) = P(A|X)P(X) = p_X\pi \), \( P(aX) = P(a|X)P(X) = q_X\pi \), \( P(AY) = P(A|Y)P(Y) = p_Y(1 - \pi) \), \( P(aY) = P(a|Y)P(Y) = q_Y(1 - \pi) \).

Based on these haplotype frequencies, we can calculate the square of the correlation coefficient between the test SNP and the local ancestry as

\[
r^2(\text{SNP, ancestry}) = \frac{\pi(1 - \pi)\left[p_X^2 - p_Y^2\right]}{[p_X + p_Y(1 - \pi)][q_X + q_Y(1 - \pi)]^2}
\]

Therefore, \( r^2 = 0 \) when \( \pi = 0 \) or 1, or \( p_X = p_Y \) and \( r^2 > 0 \) when \( 0 < \pi < 1 \) and \( p_X \neq p_Y \). In an admixed population, \( \pi \) is always between 0 and 1. Thus, there is linkage disequilibrium (LD) between the test SNP and the local ancestry at the test SNP as long as the allele frequencies are different in the two ancestral populations. In this situation, if the local ancestry is correlated with disease.
risk, the test SNP will appear to be disease associated, even when the SNP is not close to the disease locus.

To help understand this phenomenon, consider a typical admixture mapping scan with unrelated cases and controls. In the traditional admixture mapping analysis, one may identify a region that shows association evidence between a local ancestry estimate and disease status; however, due to the shared ancestry among many markers, this region may extend over several megabases. For a marker that falls in the region, the local ancestry at the marker will correlate with both the marker (due to the marker-ancestry LD described above) and the disease status (due to the admixture mapping scan). However, most markers in the region will be far from the disease-cause variant and will not be in LD with the variant in any of the ancestral populations. Naïve testing for association between a marker and disease status while ignoring the confounding effect due to the local ancestry may lead to a spurious association. Below we describe statistical procedures that can remove this confounding effect. In the Appendix (Supplementary Material), we show that such local ancestry adjustment is sufficient and that adjustment for global ancestry may not be enough. Our tests do not require the disease allele frequencies or the disease prevalences to be different in the ancestral populations.

### 2.3 Conditional likelihood for unrelated cases and controls

The conditional likelihood of $G_{SNP}$, the genotype at the test SNP, given disease status $Z$ and flanking marker genotypes $G_{ANC}$, is

$$P(G_{SNP}|G_{ANC}) = \frac{P(G_{SNP}, Z|G_{ANC})}{P(Z|G_{ANC})}$$

$$= \frac{1}{P(Z|G_{ANC})} \sum_{I_{ANC}=0}^{1} P(G_{SNP}, X_{I_{ANC}} G_{ANC}|G_{ANC})$$

$$= \frac{1}{P(Z|G_{ANC})} \sum_{I_{ANC}=0}^{1} \sum_{I_{SNP}=0}^{2} P(G_{SNP}, X_{I_{ANC}}=a_{I_{SNP}} G_{ANC}|G_{ANC})$$

$$= \frac{1}{P(Z|G_{ANC})} \sum_{I_{ANC}=0}^{1} \sum_{I_{SNP}=0}^{2} P(Z|G_{ANC}, I_{SNP}|G_{ANC}) P(G_{SNP}|I_{SNP}, G_{ANC}).$$

Given a set of unrelated cases and controls, the overall likelihood of the observed data is $L = \prod P(G_{SNP}|G_{ANC})$, where the product is taken over all individuals. To estimate the expression in (1) and (2), we need to estimate: (i) $P(G_{SNP}|Z, G_{ANC})$, the probability that an individual is affected or unaffected given his genotype and ancestry state at the test SNP; (ii) $P(G_{SNP})$, the genotype frequency given the ancestry state at the test SNP and (iii) $P(I_{SNP}|G_{ANC})$, the probability of ancestry state at the test SNP given flanking marker genotypes. Below we describe how to estimate each of the three probabilities.

### 2.4 Model for disease risk

To calculate $P(Z|G_{ANC}, I_{SNP})$, we need to have a model that relates disease risk to the SNP genotype and the underlying latent ancestry state. It is generally believed that disease risks are the same in the ancestral populations at the true disease locus (Kaplan et al., 1998). However, in genetic association studies, it is more likely that the test SNP is not the true disease locus but instead is in LD with it. Since the degrees of LD may differ in the ancestral populations, the penetrances manifested at the test SNP may be different. Therefore, two sets of penetrances are needed to reflect the fact that the probability of being affected depends not only on the individual’s genotype at the test SNP but also on the underlying ancestry state. Taking this into consideration, we propose to model the penetrances at the test SNP as shown in Table 1.

Here $(\theta_{X}, f_{1X}, f_{2X})$ and $(\theta_{Y}, f_{1Y}, f_{2Y})$ denote the penetrances of genotypes $aa, aA$ and $AA$ in ancestral populations $X$ and $Y$, respectively. For individuals carrying one allele from population $X$ and one allele from population $Y$, we allow the ancestry risk (defined as the risk of having disease due to ancestry state at the test SNP) to be additive, recessive or dominant (Table 1). We note that imposing different penetrances in different ancestral populations has been previously considered by Pritchard et al. (2000). The key advantage of using two sets of penetrances is that this allows us to incorporate different directions of association in the ancestral populations, a ‘flip-flop’ phenomenon due either to population differences or multi-locus effects and variation in inter-locus correlations (Lin et al., 2007).

### 2.5 Genotype frequencies given local ancestry state

Assuming Hardy–Weinberg equilibrium (HWE) in both ancestral populations, the genotype frequencies at the test SNP can be easily calculated. For example, when $I_{SNP} = 2$, i.e. both alleles come from population $X$, the frequency for genotype $aa$ is $q_{X}^{2}$. Table 2 shows the genotype frequencies for all three scenarios of $I_{SNP}$ (Table 2).

| $I_{SNP}$ | $I_{ANC}$ | $P(Z)$ | $P(G_{SNP}|I_{SNP})$ |
|-----------|-----------|--------|----------------------|
| 0         | $f_{0X}$  | $f_{0X}$ | $\frac{f_{0X} + f_{0Y}}{2}$ |
| 1         | $f_{1X}$  | $f_{1X}$ | $f_{1X}$ |
| 2         | $f_{2X}$  | $f_{2X}$ | $f_{2X}$ |

### 2.6 Estimation of ancestry state at the test SNP

The conditional probability distribution of the ancestry state at the test SNP given flanking marker genotypes can be obtained from external programs. Several software packages are available for estimating local ancestry, including ANCESTRYMAP (Patterson et al., 2004), MALDISOFT (Montana and Pritchard, 2004), ADMIXPROGRAM (Zhu et al., 2006), SABER (Tang et al., 2006), LAMP (Sankararaman et al., 2008), HAPPA (Sandquist et al., 2008) and HAPMIX (Price et al., 2009). The choice of which program to use will depend on the nature of the data. For studies where the flanking markers are not densely spaced, programs such as ANCESTRYMAP and ADMIXPROGRAM would be sufficient. For markers from GWAS, programs such as SABER, HAPPA, LAMP and HAPMIX would be more appropriate. The current state-of-the-art method is HAPMIX, which for GWAS data can yield an estimated ancestry that has as high as 98% correlation with the true ancestry (Price et al., 2009).

### 2.7 Tests of association

With the previously developed likelihood framework, we can evaluate whether the test SNP is associated with the disease of interest. Under the null hypothesis that the test SNP is not associated with the disease, the
An admixed population was generated by forming inter-population marriages to evaluate the performance of the proposed tests, we conducted extensive continuous gene flow model. In brief, in Generation 0, the marker genotypes adjacent AIMs in each population. Samples were simulated according to the λ of children in each marriage was assumed to follow a Poisson distribution in subsequent generations. Specifically, in each generation, we took a population stratification: (i) stratification due to local ancestry difference at 0.1; for individuals with one copy of population X allele and one copy of population Y allele, the probability of being affected was either 0.2 (additive ancestry risk), 0.1 (recessive ancestry risk) or 0.3 (dominant ancestry risk). When stratification was due to local ancestry difference, the probability of being affected was equal to the global ancestry proportion of the individual, calculated as the average of the ancestry proportions across all markers in the genome. For power evaluation, we simulated case-control data following disease models specified in Table 7. As noted earlier, an advantage of our tests is that they do not require the disease allele frequencies or disease prevalences to be different in the ancestral populations. To evaluate the performance of our tests in this setting, we simulated data assuming the disease allele frequency was 0.4, disease prevalence was 0.3, and genotype relative risk (GRR) was 1.2 in both ancestral populations. Type I error rates were estimated based on 10000 replicate datasets, and power was estimated based on 1000 replicates. Each replicate dataset consisted of 1000 unrelated cases and 1000 unrelated controls. We analyzed each simulated dataset using the following tests: (i) LRT, (ii) Logistic-Local, (iii) Logistic-Global, a logistic regression procedure with the global ancestry proportion (calculated as the average of local ancestry proportions across all markers in the genome) as a covariate and (iv) EIGENSTRAT. We assumed additive model for the test SNP genotype in all tests. For the LRT approach, we conducted the test assuming the ancestry risk model is additive, dominant or recessive. For EIGENSTRAT, we included the first 10 principal components in the analysis.

2.9 Simulation of a synthetic GWAS dataset

To test the performance of our tests in GWAS settings, we simulated a synthetic GWAS dataset based on ancestry characteristics in the Maywood study (Kang et al., 2010). In this study, 701 unrelated African Americans were collected from Maywood, IL, USA. All study subjects were genotyped using Affymetrix 6.0 SNP array. Due to its small sample size and the lack of disease phenotypes, the Maywood dataset is not appropriate for testing the performance of different tests. However, we can use it to simulate GWAS dataset with realistic admixture patterns. Specifically, we used ADMIXTEST (Zhu et al., 2006) to infer each individual’s ancestry using 2660 selected ancestry informative SNPs and obtained the distribution of ancestry proportion. Based on this distribution, we then simulated an admixed population with average 20% Caucasian and 80% African ancestries. For each individual with average ancestry proportion \( \nu_i \) (sampled from the ancestry distribution estimated from Maywood), we simulated SNP genotypes for 22 autosomal chromosomes using data generated by HapMap, which includes 1969739 SNPs with complete haplotype information for CEU (Utah residents with ancestry from northern and western Europe) and YRI (Yoruba in Ibadan, Nigeria) samples. For each chromosome, we simulated the number of crossover points \( s \) from a Poisson distribution with mean \( \mu = l \times g \times 10^{-7} \), where \( l \) is the length of the chromosome, \( g \) is the generation since the individual began admixture and was randomly sampled from 1 to 10. We then randomly sampled haplotype segments from CEU or YRI haplotypes between two crossovers independently according to the average ancestry proportion. Next, we introduced selection to a 5 Mb region centered at rs6576848 (27053359 bp) on Chromosome 1 by simulating haplotype segments with 40% Caucasian ancestry and 60% African ancestry. The size of the selection region is similar to that reported by Tang et al. (2007). Similar simulation strategies have been employed by Qin et al. (2010). We assigned disease status according to local ancestry at rs6576848: the probabilities of being affected are 0.1, 0.2 and 0.3, respectively, if the
Table 3. Comparison of Type I error rates (%) when population stratification is due to local ancestry difference between cases and controls

<table>
<thead>
<tr>
<th>Allele</th>
<th>Ancestry Risk</th>
<th>LRT-A</th>
<th>LRT-R</th>
<th>LRT-D</th>
<th>LL</th>
<th>LG</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>pX = 0.4</td>
<td>Add</td>
<td>3.31</td>
<td>3.50</td>
<td>3.35</td>
<td>3.20</td>
<td>12.44</td>
<td>7.62</td>
</tr>
<tr>
<td>pX = 0.2</td>
<td>Rec</td>
<td>3.67</td>
<td>3.25</td>
<td>3.81</td>
<td>3.17</td>
<td>32.92</td>
<td>16.47</td>
</tr>
<tr>
<td>pX = 0.6</td>
<td>Add</td>
<td>2.83</td>
<td>3.06</td>
<td>3.59</td>
<td>2.65</td>
<td>40.89</td>
<td>34.67</td>
</tr>
<tr>
<td>pX = 0.2</td>
<td>Rec</td>
<td>3.17</td>
<td>3.25</td>
<td>3.50</td>
<td>2.95</td>
<td>84.70</td>
<td>53.36</td>
</tr>
<tr>
<td>pY = 0.8</td>
<td>Add</td>
<td>2.88</td>
<td>2.86</td>
<td>3.48</td>
<td>2.87</td>
<td>79.47</td>
<td>74.28</td>
</tr>
<tr>
<td>pY = 0.2</td>
<td>Add</td>
<td>3.45</td>
<td>3.25</td>
<td>3.58</td>
<td>3.09</td>
<td>99.40</td>
<td>91.08</td>
</tr>
<tr>
<td>pY = 0.8</td>
<td>Add</td>
<td>3.01</td>
<td>2.87</td>
<td>3.16</td>
<td>2.77</td>
<td>20.32</td>
<td>12.07</td>
</tr>
</tbody>
</table>

Significance was assessed at the 5% level. LRT-A, LRT assuming ancestry risk is additive; LRT-R, LRT assuming ancestry risk is recessive; LRT-D, LRT assuming ancestry risk is dominant; LL, Logistic-Local; LG, Logistic-Global; ES, EIGENSTRAT.

Table 4. Comparison of Type I error rates (%) when population stratification is due to global ancestry difference between cases and controls

<table>
<thead>
<tr>
<th>pX</th>
<th>pY</th>
<th>LRT-A</th>
<th>LRT-R</th>
<th>LRT-D</th>
<th>LL</th>
<th>LG</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>3.25</td>
<td>3.15</td>
<td>3.88</td>
<td>3.14</td>
<td>2.98</td>
<td>3.08</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2</td>
<td>3.20</td>
<td>3.11</td>
<td>3.59</td>
<td>2.82</td>
<td>3.48</td>
<td>3.36</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>3.18</td>
<td>3.18</td>
<td>3.18</td>
<td>3.28</td>
<td>3.69</td>
<td>4.04</td>
</tr>
</tbody>
</table>

Significance was assessed at the 5% level.

3 RESULTS

3.1 Comparison of Type I error rates

Table 3 shows the estimated Type I error rates of different tests when population stratification is due to local ancestry difference at the test SNP between cases and controls. Not surprisingly, tests that adjust for global ancestry cannot effectively remove the effect of the test SNP between cases and controls. Therefore, we anticipate that the amount of uncertainty in practical GWAS may not be a concern.

In the above simulations, we assumed the ancestry state at the test SNP was known. We also assessed the performance of our tests when there is uncertainty in ancestry probability estimation. Ideally, one should use programs such as HAPMIX to estimate ancestry probabilities. However, these programs are computationally intensive and it is not feasible to run these programs on all simulated datasets. To circumvent this difficulty, we added uncertainties to the ancestry states according to the error model specified in Table 5, which creates similar patterns and magnitude of uncertainty as those in the estimated ancestry from HAPMIX and LAMP based on our analysis of testing datasets. In our error model, we set the uncertainty parameter for population X to be less than that for population Y. The rationale is that for the data we simulated (mimicking African Americans), 80% of the genome came from population X and only 20% came from population Y. Thus, the effective sample size for population X was larger, and this led to more accurate ancestry probability estimates than for population Y. Table 6 shows the Type I error rates of the LRT and Logistic-Local approaches when there are uncertainties in ancestry estimation. Its Type I error rates were below the nominal level under all the settings we considered. For the LRT approach, its Type I error rates were under control when assuming additive or dominant model for the ancestry risk, and slightly inflated when the assumed ancestry risk model was recessive and the degree of uncertainty was high. We note that when GWAS data are available, with the current state-of-the-art program such as HAPMIX, ancestry probabilities can be reliably estimated for an admixed population such as African American population. Therefore, we anticipate that the amount of uncertainty in practical GWAS may not be a concern.

Table 5. Uncertainty model for ancestry probability estimation

<table>
<thead>
<tr>
<th>Ancestry probability</th>
<th>True probability</th>
<th>Probability with uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(\text{Es}=0</td>
<td>\text{Anc}=\text{X})</td>
<td>0</td>
</tr>
<tr>
<td>P(\text{Es}=1</td>
<td>\text{Anc}=\text{X})</td>
<td>0</td>
</tr>
<tr>
<td>P(\text{Es}=2</td>
<td>\text{Anc}=\text{X})</td>
<td>1</td>
</tr>
<tr>
<td>P(\text{Es}=0</td>
<td>\text{Anc}=\text{Y})</td>
<td>0</td>
</tr>
<tr>
<td>P(\text{Es}=1</td>
<td>\text{Anc}=\text{Y})</td>
<td>1</td>
</tr>
<tr>
<td>P(\text{Es}=2</td>
<td>\text{Anc}=\text{Y})</td>
<td>0</td>
</tr>
</tbody>
</table>

\epsilon_x is randomly generated from uniform (0, \text{E}_X), and \epsilon_y is generated from uniform (0, \text{E}_Y), where \text{E}_X and \text{E}_Y are uncertainty parameters.
Table 6. Type I error rates (%) of LRT and Logistic-Local when there are uncertainties in ancestry probabilities

<table>
<thead>
<tr>
<th>Ancestry Risk</th>
<th>$E_{x}$</th>
<th>$E_{y}$</th>
<th>LRT-A</th>
<th>LRT-R</th>
<th>LRT-D</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add</td>
<td>0.01</td>
<td>0.05</td>
<td>2.81</td>
<td>3.86</td>
<td>2.99</td>
<td>2.63</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.10</td>
<td>3.18</td>
<td>5.37</td>
<td>2.81</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.05</td>
<td>3.04</td>
<td>4.39</td>
<td>2.92</td>
<td>2.70</td>
</tr>
<tr>
<td>Rec</td>
<td>0.01</td>
<td>0.05</td>
<td>3.20</td>
<td>3.86</td>
<td>4.19</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.10</td>
<td>3.45</td>
<td>4.86</td>
<td>3.03</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.05</td>
<td>3.51</td>
<td>4.37</td>
<td>4.22</td>
<td>2.95</td>
</tr>
<tr>
<td>Dom</td>
<td>0.01</td>
<td>0.05</td>
<td>2.90</td>
<td>3.44</td>
<td>2.72</td>
<td>2.52</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.10</td>
<td>3.22</td>
<td>5.45</td>
<td>2.48</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.05</td>
<td>3.12</td>
<td>4.09</td>
<td>2.63</td>
<td>2.53</td>
</tr>
</tbody>
</table>

The reference allele frequency for population $X$ at the test marker is 0.6 and 0.2 for population $Y$. $E_{x}$ and $E_{y}$ are uncertainty parameters.

Table 7. Comparison of power (%)

<table>
<thead>
<tr>
<th>GRR</th>
<th>Ancestry Risk</th>
<th>LRT-A</th>
<th>LRT-R</th>
<th>LRT-D</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRR = 1.2</td>
<td>Add</td>
<td>67.3</td>
<td>68.1</td>
<td>67.6</td>
<td>77.0</td>
</tr>
<tr>
<td>GRR = 1.1</td>
<td>Rec</td>
<td>61.2</td>
<td>62.0</td>
<td>59.4</td>
<td>68.2</td>
</tr>
<tr>
<td></td>
<td>Dom</td>
<td>75.9</td>
<td>74.9</td>
<td>75.7</td>
<td>84.8</td>
</tr>
<tr>
<td>GRR = 1.2</td>
<td>Add</td>
<td>63.5</td>
<td>62.0</td>
<td>61.5</td>
<td>72.8</td>
</tr>
<tr>
<td>GRR = 1.0</td>
<td>Rec</td>
<td>56.3</td>
<td>60.5</td>
<td>47.1</td>
<td>57.9</td>
</tr>
<tr>
<td></td>
<td>Dom</td>
<td>75.0</td>
<td>74.8</td>
<td>75.9</td>
<td>84.9</td>
</tr>
<tr>
<td>GRR = 1.2</td>
<td>Add</td>
<td>66.5</td>
<td>58.3</td>
<td>57.3</td>
<td>53.4</td>
</tr>
<tr>
<td>GRR = 0.7</td>
<td>Rec</td>
<td>81.1</td>
<td>96.8</td>
<td>20.9</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>Dom</td>
<td>75.2</td>
<td>69.6</td>
<td>80.0</td>
<td>80.6</td>
</tr>
</tbody>
</table>

The prevalence of population $X$ is 0.5 and the prevalence of population $Y$ is 0.1. $p_{x}$ is the genotype relative risk for population $X$ at the test marker. GRR is the genotype relative risk for population $Y$ at the test marker.

3.2 Comparison of power

Next, we compared the power of the LRT and Logistic-Local approaches using data simulated under disease models specified in Table 7. We did not compare with the Logistic-Global and EIGENSTRAT approaches as they have inflated Type I error rates when local ancestry is the main factor for stratification. We considered three scenarios for power evaluation: (i) the test SNP was associated with the disease in the same direction in populations $X$ and $Y$ (GRR$_x$ = 1.2, GRR$_y$ = 1.1), (ii) the test SNP was associated with the disease only in population $X$ (GRR$_x$ = 1.2, GRR$_y$ = 1.0) and (iii) the test SNP was associated with the disease in opposite directions in populations $X$ and $Y$ (GRR$_x$ = 1.2, GRR$_y$ = 0.7). Table 7 shows the power comparison results. For the first two scenarios, the power of the Logistic-Local approach was slightly higher than LRT, especially when the true ancestry risk model was additive or dominant. However, when the test SNP was associated with the disease in opposite directions in the ancestral populations, the power of the Logistic-Local approach can be substantially lower than LRT. For example, in the third scenario, the power of the LRT approach assuming recessive ancestry risk was 96.8% when the true ancestry risk model was recessive, but the power of the Logistic-Local approach was only 14.8%. The power of the LRT approach can be higher than Logistic-Local even when the ancestry risk model is misspecified. Such a significant drop in power for the Logistic-Local approach is due to: (i) its inherent assumption on the additivity of the ancestry risk on the logit scale and (ii) its inability to model opposite directions of association in the ancestral populations. The power of this test can be greatly attenuated when these assumptions are not satisfied. In contrast, the LRT approach uses two sets of penetrances to model disease risk for the two ancestral populations, and thus is more flexible.

For the situation where the disease allele frequency and disease prevalence are equal in the two ancestral populations, we compared the power of the LRT and Logistic-Local approaches with the traditional admixture mapping method, which tests for the correlation between disease status and local ancestry proportion at the test SNP. When the disease allele frequency was 0.4, disease prevalence was 0.3 and GRR was 1.2 in both ancestral populations, the power of the traditional admixture mapping method was close to the 5% nominal level, whereas the power of the LRT and Logistic-Local approaches were 82% and 89%, respectively. Therefore, our tests can identify signals that are missed by the traditional admixture method and be used as a general tool for genetic association analysis.

3.3 Analysis of the synthetic GWAS dataset

We analyzed the synthetic GWAS dataset simulated based on the Maywood study (Kang et al., 2010) using five approaches, including LRT, Logistic-Local, Logistic-Global, EIGENSTRAT and unadjusted trend test. For the LRT, we assumed the disease prevalences in Africans and Caucasians to be 0.3 and 0.1, respectively, and analyzed the data assuming additive, dominant and recessive models for ancestry risk. Figure 1 shows the results from different tests for the region on Chromosome 1 that undergoes selection. Using local adjustment tests, we did not observe any evidence of association; however, all the other tests yield highly significant results, with many SNPs reaching genome-wide significance ($P$-values < 5 × 10$^{-8}$). The global ancestry adjustment procedures such as Logistic-Global and EIGENSTRAT perform similarly as the unadjusted trend test, suggesting that global ancestry cannot sufficiently capture the local ancestry variation for regions with natural selection. We note that the significant results from Logistic-Global, EIGENSTRAT and trend tests are not due to allelic association between the SNPs and disease status, but are rather driven by the association between the disease status and local ancestry uncertainty parameters.
ancestry in the region. Our results demonstrate that global ancestry adjustment procedures may identify a wrong region if local ancestry in the region happens to be associated with disease status. As an example, it has been previously shown that EIGENSTRAT cannot completely remove the effect of population stratification at the lactase gene (Epstein et al., 2007; Li et al., 2010; Price et al., 2006). We suspect that this failure is probably because global ancestry cannot fully reflect the ancestry variation at the lactase gene.

We also note that admixture mapping searches for correlation between local ancestry and disease status. Our results indicate that even if there is a causal SNP in the region, adjusting for global ancestry will result in a wide region with association signals (≈5 Mb or even larger depending on the size of the ancestry block) due to inflated signals induced by association with local ancestry. Therefore, the global ancestry adjustment methods may point to the wrong region that harbors causal SNPs, but this region might be too wide and may lead to following up of wrong SNPs/genes. In contrast, the local ancestry adjustment procedures will appropriately control the background signals induced by association with local ancestry, and thus can pinpoint the correct SNPs/genes for follow-up study. This makes the local adjustments procedures a useful tool for fine mapping of regions identified from admixture mapping studies.

4 DISCUSSION

We showed that in genetic association analysis of admixed populations, local ancestry difference between study subjects can confound with association signal and lead to spurious associations if not appropriately controlled. We also showed theoretically that to remove the confounding effect of local ancestry, it is sufficient to condition on local ancestry at the test SNP, whereas conditioning on global ancestry may be ineffective. To remove the spurious associations due to the confounding effect of local ancestry, we further proposed two novel association tests that adjust for local ancestry at the test SNP. We conducted extensive simulations and evaluated the performance of different tests under various settings. Our simulation results indicate that regardless of the mechanism of population stratification, whether due to local or global ancestry difference, it is sufficient to control population stratification by conditioning on local ancestry at the test marker. In contrast, global ancestry adjustment procedures such as EIGENSTRAT and Logistic-Global cannot completely remove the effect of population stratification induced by local ancestry difference. The reason is that global ancestry information as represented by a global ancestry proportion or a few principal components obtained from all markers across the genome cannot accurately reflect the amount of ancestry variation in local regions.

We proposed two tests for association analysis in admixed populations. Our first test, LRT, is based on a conditional likelihood framework which models the distribution of a test SNP given disease status and flanking marker genotypes. This conditional likelihood allows us to explicitly model local ancestry differences among study subjects and thus it eliminates the effect of population stratification at the test SNP. Our second test, Logistic-Local, is based on a logistic regression model that adjusts for local ancestry proportion at the test SNP. Although the LRT approach is computationally more involved than Logistic-Local, it is more flexible and is particularly useful when the directions of association are different in the ancestral populations. Another advantage of the LRT approach is that, in addition to testing the null hypothesis as described in Section 2, we can test two other hypotheses H0: no disease–SNP association in population X, and H0: no disease–SNP association in population Y, separately. This can be achieved by testing H0: f0 = f1 = f2, X = f0, Y = f1, Y = f2, respectively. These tests allow a cross-ethnicity replication since evidence of disease association in the two ancestral populations can be directly compared (Risch et al., 2006).

Our tests are different from admixture mapping in that we directly assess the correlation between a phenotype and SNP genotypes. In contrast, admixture mapping examines the association between a phenotype and local ancestry without fully using the actual genotypes at each SNP. Therefore, SNPs falling within the same ancestry block will share similar admixture mapping signal, which explains why admixture mapping has substantially lower resolution than direct SNP association analysis. Since the association tests we proposed directly compare the allele frequencies between cases and controls, they can serve as a fine-mapping tool for regions identified from admixture mapping studies. As shown in our simulations, another advantage of our tests is that, unlike admixture mapping, which may miss disease variants with similar allele frequencies in the ancestral populations, our tests are able to identify such variants since the allele frequencies between cases and controls are directly compared. Therefore, our tests can be used as a general tool for genetic association analysis.

Our tests rely on the estimates of local ancestry probabilities. It is computationally intensive to estimate these probabilities for GWAS datasets. In addition, our LRT test requires the estimation of several parameters, which is computationally more involved than logistic regression; for example, with 1000 cases and 1000 controls, it took the LRT test about 3 days to finish the analysis of 500 000 SNPs using a single CPU. However, the computations can be parallelized across chromosomes, and thus the computation is tractable even for very large datasets if a computing cluster is available.

Our methods treat estimated ancestry as plug-in estimates in the likelihood calculation. Although simple, such a two-step procedure is not as efficient as approaches that estimate ancestry and disease model parameters simultaneously. Another problem is that the disease phenotypes are not used when estimating ancestry. However, since the allele frequencies between cases and controls are directly expected by chance, and thus may lead to ancestry estimates that are biased towards the null. How to jointly model local ancestry and disease status would merit further research.

Our methods share similarity with a recent paper published by Qin et al. (2010), who proposed to correct for population stratification using local principal components (PCs). In principle, this method can be applied to admixed populations as well. However, local PCs can only approximate the local ancestry and require a predefined window size, whereas for admixed populations such as African Americans, locus-specific ancestry can be inferred accurately. As noted by Qin et al. (2010), the local PC-based approach will be more appropriate for a population whose substructure is subtle, due to either the lack of information on the ancestral population or when admixture has occurred within similar populations, for example, European Americans.

In summary, we have proposed two novel association tests for admixed populations. We showed that dependence between local ancestry and disease phenotype can lead to spurious associations.
Our results indicate that it is better to adjust for local ancestry than for global ancestry to appropriately control for population stratification. The method in this article is implemented in a C program and can be obtained by contacting the last author.

ACKNOWLEDGEMENTS
We thank Dr Hongzhe Li for helpful discussions.

**Funding:** National Institutes of Health (grants R01HG004517 to M.L. and C.L., R01HG005854 to M.L. and R01HL074166 and R01 HG003654 to X.Z.).

**Conflict of Interest:** none declared.

REFERENCES