Generalized Admixture Mapping for Complex Traits

Bin Zhu,*[,1] Allison E. Ashley-Koch,[†] and David B. Dunson[‡]
*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland 20850, [†]Center for Human Genetics, Duke University Medical Center, Duke University, Durham, North Carolina 27710, and [‡]Department of Statistical Science, Duke University, Durham, North Carolina 27708

ABSTRACT Admixture mapping is a popular tool to identify regions of the genome associated with traits in a recently admixed population. Existing methods have been developed primarily for identification of a single locus influencing a dichotomous trait within a case-control study design. We propose a generalized admixture mapping (GLEAM) approach, a flexible and powerful regression method for both quantitative and qualitative traits, which is able to test for association between the trait and local ancestries in multiple loci simultaneously and adjust for covariates. The new method is based on the generalized linear model and uses a quadratic normal moment prior to incorporate admixture prior information. Through simulation, we demonstrate that GLEAM achieves lower type I error rate and higher power than ANCESTRYMAP both for qualitative traits and more significantly for quantitative traits. We applied GLEAM to genome-wide SNP data from the Illumina African American panel derived from a cohort of black women participating in the Healthy Pregnancy, Healthy Baby study and identified a locus on chromosome 2 associated with the averaged maternal mean arterial pressure during 24 to 28 weeks of pregnancy.

KEYWORDS
generalized linear model
local ancestry
mapping by admixture linkage disequilibrium
quadratic normal moment prior
quantitative traits

Admixture mapping, also known as mapping by admixture linkage disequilibrium, has become an important tool for localizing disease genes. A number of admixture mapping studies have successfully identified candidate loci associated with common complex traits and biomarkers (Reich *et al.* 2005; Zhu *et al.* 2005; Freedman *et al.* 2006; Kao *et al.* 2008; Yang *et al.* 2011).

As a genome-wide association approach, admixture mapping aims to identify susceptibility loci, which confer risk or are linked with other loci harboring risk variants, for complex-traits that have different prevalences between ancestral populations (McKeigue 2005; Winkler *et al.* 2010). In recently admixed populations, such as African Americans or Hispanic Americans, the chromosome resembles a mosaic of ancestry blocks, with alleles inherited together from one ancestral population within each block. The ancestral populations have different risks for the trait, which is assumed to be due in part to frequency differences in risk variants. The block containing the risk variant is more likely to have originated from the high-risk ancestral population than the low-risk ancestral population. Hence, detecting the association between ancestry block and trait helps us to localize the susceptibility loci.

The ancestral status of a block at a specific genomic region, or local ancestry, is unobserved and can be estimated based on ancestry informative markers (AIMs), such as single-nucleotide polymorphisms (SNPs), which vary in frequency across ancestral populations. AIMs tag the status of an ancestry block, similar to that of tagSNPs, which are used to characterize common haplotypes in a chromosomal region. In the African-American population, the linkage disequilibrium due to admixture extends for a much wider region than the linkage disequilibrium between haplotypes (Smith *et al.* 2004; Patterson *et al.* 2004). Hence, compared with the tagSNP-based genome-wide association study, admixture mapping requires many fewer markers to tag the whole genome and therefore increases the detection power at a reduced resolution, which is still greater than linkage analysis (Patterson *et al.* 2004; Smith and O'Brien 2005). Moreover, admixture mapping is less vulnerable to allelic heterogeneity because it relies on local ancestry instead of alleles directly.

Given the local ancestries of each individual, several hypothesis testing-based approaches have been proposed to test, one locus at a time, the null hypothesis that the AIM is unlinked to the complex-trait/disease for a dichotomous trait within a case-control study design. McKeigue (1998) proposed a test for gametic disequilibrium between an AIM locus and the trait locus, conditional on the parental

doi: 10.1534/g3.113.006478
Manuscript received May 15, 2012; accepted for publication May 1, 2013

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.113.006478/-/DC1
[1]Correspondence author: 9609 Medical Center Drive, Suite 7E618, Rockville, MD 20850. E-mail: bin.zhu@nih.gov

admixture. Patterson *et al.* (2004) suggested a Bayesian likelihood ratio test, comparing the likelihood under the alternative hypothesis (a given AIM locus is associated with the trait) *vs.* the one under the null hypothesis, for cases and controls respectively. Zhu *et al.* (2004) described a Z-score statistic, similar to the one proposed by Montana and Pritchard (2004), for testing the estimated local ancestry proportion is equal to one under the null hypothesis for case-control and case-only studies.

In contrast, few methods are proposed for the quantitative traits and to consider multiple loci simultaneously while adjusting for other risk factors. To apply the aforementioned admixture methods primarily developed for a dichotomous trait, the common practice has been to dichotomize subjects with the least and greatest q% (*e.g.*, 20%) of the quantitative trait value as cases and controls; The remaining subjects with in-between quantitative trait values are then discarded (Reich *et al.* 2007; Cheng *et al.* 2010; Scherer *et al.* 2010). In addition, ADMIXMAP (Hoggart *et al.* 2003) has been proposed for quantitative traits based on generalized linear model, which is also used by Basu *et al.* (2009) and Zhu *et al.* (2011) for one locus at a time. However, complex traits are commonly caused by joint effects of the multiple genes and other risk factors, such as age, sex, and smoking status. Investigating the association between AIM loci and a trait, one locus a time, without considering other loci or risk factors may capture a rather small proportion of joint effects and will possibly lead to inconsistent conclusions. Similar considerations have been addressed in association mapping using shrinkage priors (Wu *et al.* 2009; Guan and Stephens 2011).

With these motivations, we propose regression-based generalized admixture mapping (GLEAM) for both quantitative and qualitative traits. The new approach is able to examine the association between the complex trait and single or multiple loci simultaneously while also adjusting for other risk factors. GLEAM is based on generalized linear models (GLMs) (McCullagh and Nelder 1989), with linear regression for continuous traits, logistic regression for binary (*e.g.*, case-control) traits and Poisson regression for count traits. The predictors in GLM include local ancestries at the given AIM loci and other risk factors. The local ancestry is defined as the number of alleles from the high-risk ancestral population, for example, 0, 1, or 2 alleles from African ancestry at a given AIM locus. The association examined in GLEAM can be adjusted by other risk factors. We assume for complex genetic traits that most loci have no association with the trait, a few loci may have small to modest association (*e.g.*, odds ratio <2 for binary traits), and the loci with greater proportions of disease-causing alleles from the high-risk population would possibly have stronger association with the traits. This prior knowledge is incorporated into GLEAM by using a quadratic normal moment (QNM) prior (Johnson and Rossell 2010) for the coefficients in GLM (see *Material and Methods*) with the benefit of reducing the type I error while increasing the power, as demonstrated by the simulations in *Results*.

The number of AIMs (1500~3000) (Smith *et al.* 2004) is usually larger than the number of study subjects, and keeps increasing ($>$4000) (Tandon *et al.* 2011) with advances due to the HapMap project (The International Hapmap Consortium 2005) and commercially available genome-wide SNP arrays. It is not feasible to consider loci all together simultaneously due to the "curse of dimensionality" (Bellman 1961). Rather, we propose a two-stage approach: in the first stage, we examine the association between local ancestries with the trait for one locus at a time and select a small subset of susceptibility loci; in the second stage, the associations between the various combinations of these selected loci and the trait are evaluated and the most significant ones are reported. The associations in both steps are assessed by the Bayes factor (BF), the ratio between the likelihood of observed traits under the alternative hypothesis (presence of association between single or multiple loci with traits) and that under the null hypothesis (lack of association).

Different from the association mapping based on the SNPs that are directly measured, the local ancestries are unobserved and are inferred on the basis of the AIMs via use of the Hidden Markov Model (HMM) detailed in the Appendices. At each AIM locus, the number of alleles from the high-risk ancestral population is imputed multiple times for every subject, using an Markov chain Monte Carlo (MCMC) algorithm. By using the multiple imputed datasets of local ancestries, we are able to assess the association between the traits and local ancestries directly while taking imputation uncertainty into account through Bayesian averaging. Importantly, our multiple imputation approach preserves the admixture linkage disequilibrium between the AIM loci, which is crucial for multilocus admixture mapping in GLEAM. In addition, GLEAM can also use the local ancestries sampled by other local ancestry inferring methods, such as HAPMIX (Price *et al.* 2009).

MATERIAL AND METHODS

Generalized linear model with QNM prior

GLEAM is a regression method that extends the current approaches in various ways. The most obvious extension is to accommodate both quantitative and qualitative traits y_i through a generalized linear model with the ability to adjust for covariates $\boldsymbol{E}_i = (E_{i1}, E_{i2}, \ldots, E_{iq})'$. Specifically, we use the linear model for continuous traits,

$$\mathrm{y}_i = \beta_0 + \boldsymbol{\beta}'\boldsymbol{S}_i + \boldsymbol{\alpha}'E_i + \varepsilon_i, \tag{1}$$

and the logistic model for dichotomous traits,

$$\mathrm{logit}\{\mathrm{Prob}(\mathrm{y}_i = 1)\} = \beta_0 + \boldsymbol{\beta}'S_i + \boldsymbol{\alpha}'E_i, \tag{2}$$

where p local ancestries $\boldsymbol{S}_i = (S_{i1}, S_{i2}, \ldots, S_{ip})'$ are considered and centered to have mean zero, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)'$ are the regression coefficients for $\boldsymbol{S}_i$ and $\boldsymbol{E}_i$ respectively, and $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. We use the Bayes factor to assess the admixture association between local ancestries and the trait of interest. The Bayes factor is the ratio of the marginal likelihoods under the alternative hypothesis, $\mathbf{H}_1$: $\beta_j \neq 0$ for $j = 1, \ldots, p$, and null hypothesis, $\mathbf{H}_0$: $\beta_j = 0$ for $j = 1, \ldots, p$. Marginal likelihoods remove the parameters from the likelihood by integrating over the prior distribution. The larger the Bayes factor, the stronger the evidence in favor of $\mathbf{H}_1$.

As a prior distribution for β under $\mathbf{H}_1$, we use the QNM prior having density

$$f_{\mathrm{QNM}}(\beta; \tau, \sigma^2, \boldsymbol{\Sigma}) = \frac{\beta'\boldsymbol{\Sigma}^{-1}\beta}{I\tau\sigma^2 p} f_{\mathrm{N}_p}(\beta; 0, I\tau\sigma^2\boldsymbol{\Sigma}),$$

where $f_{\mathrm{N}_p}(\cdot; \boldsymbol{m}, \boldsymbol{V})$ is the p-dimensional multivariate normal distribution with the mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{V}$, and τ is the dispersion parameter. The QNM prior is able to incorporate the case with a large number of loci of tiny effect. As shown in Figure 1A, the modes of the prior distribution will move toward zero when we reduce the value of τ. For illustration purposes, we only showed a particular value of $\tau = 0.01$, but as we decrease this value, tiny effects are accommodated. For data containing a large number of loci of tiny effect, the empirical Bayes approach should estimate a very small value, and the QNM prior will concentrate on very small effect sizes. Usual priors face major problems in distinguishing the signal from the noise, and we argue that nonlocal priors such as

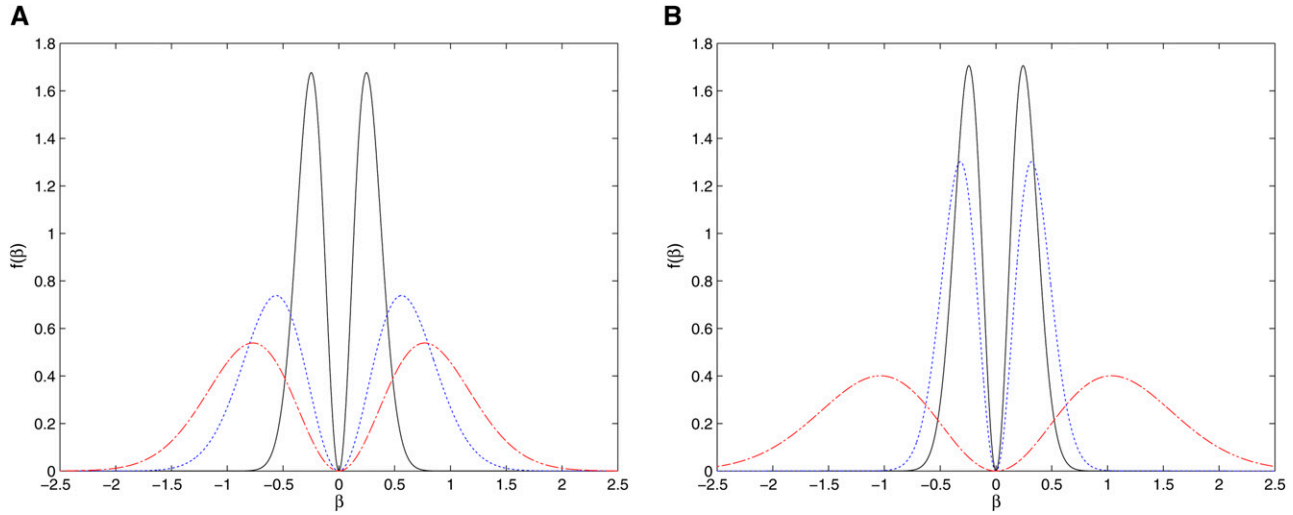

Figure 1 Univariate quadratic normal moment prior (A) for $\tau = 0.01$ (—), $\tau = 0.05$ (···), and $\tau = 0.1$ (−·−) when $p_a = 0.8$; (B) for $p_a = 0.8$ (—), $p_a = 0.9$ (. . .), and $p_a = 0.99$ (−·−) when $\tau = 0.01$. In both cases, $\sigma^2 = 1$ and $\boldsymbol{\Sigma} = (\sum_{i=1}^{1000} S_i^2)^{-1}$ with $\Pr(S_i = 0) = (1-p_a)^2$, $\Pr(S_i = 1) = 2p_a(1-p_a)$ and $\Pr(S_i = 2) = p_a^2$.

the quadratic normal provide more accurate results for genetic effects on complex traits. Hence, The QNM prior increases the evidence in favor of both the true null and true alternative hypothesis, compared to other prior distributions (*e.g.*, intrinsic and Cauchy priors) (Johnson and Rossell 2010). Moreover, we specify $\sigma^2\boldsymbol{\Sigma}$ as the covariance matrix of the (iterative weighted) least square estimation of $\boldsymbol{\beta}$ in the GLM. This choice not only leads to convenient computation but also easily incorporates the prior knowledge about the effect of local ancestry on the trait. For example, when $\boldsymbol{S}_i$ is orthogonal to $\boldsymbol{E}_i$, $\boldsymbol{\Sigma} = (\boldsymbol{S}'\boldsymbol{S})^{-1}$ with $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_I]'$ in the linear model for the continuous trait. As illustrated by the right panel of Figure 1, the QNM prior with $\boldsymbol{\Sigma} = (\boldsymbol{S}'\boldsymbol{S})^{-1}$ suggests that for each locus, the greater the proportion of alleles from the high-risk population (p_a), on average the larger the risk effect of local ancestry. Such relationships frequently are observed in admixture mapping but not in association mapping based on SNPs in general. More importantly, when we investigate multiple loci simultaneously, it is crucial to take the correlation (linkage disequilibrium, LD) between the local ancestries into consideration. Figure 2 plots several volcano-shaped bivariate QNM densities for various correlations between two local ancestries. It is clear that for two loci with admixture linkage equilibrium (as shown in Figure 2A), such as two loci on different chromosomes, their risk effects would be independent; and that for two loci with high admixture LD (as shown in Figure 2D), usually located in the same gene, they would have similar risk effects.

Under the QNM prior for β, the Bayes factor is simply

$$BF(\mathbf{y}) = \frac{p+T}{p(1+I\hat{\tau})^{p/2+1}} \exp\left(\frac{T}{2}\right), \tag{3}$$

where $T = \frac{I\tau}{\hat{\sigma}^2(1+I\hat{\tau})}\hat{\beta}'\hat{\Sigma}_{\hat{\beta}}^{-1}\hat{\beta}$, $\hat{\beta}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$, adjusted by other risk covariates when necessary, $\hat{\Sigma}_{\hat{\beta}}^{-1}$ is the corresponding covariance matrix estimate and $\hat{\tau}$ and $\hat{\sigma}^2$ are the empirical Bayes estimates. Bayes factor (3) is used to identify the loci associated with the traits, detailed as follows.

Generalized admixture mapping procedure

We propose a two-stage approach for GLEAM. In the first stage, we examine the marginal association between a single AIM locus and the trait, using the Bayes factors (3), one locus a time for J AIM loci. The loci at which $log_{10}BF(\mathbf{y}) > \delta$ are considered susceptibility loci. Although the "one locus a time" approach explores the marginal association and is widely used, marginal association only reflects part of the relationship between the AIM loci and the trait. Several loci in different regions may show associations with the trait. Thus, it is desirable to quantify the evidence for joint association of multiple loci with the trait. For this reason, in the second stage, we list all possible combinations of susceptibility loci selected in the first stage. For each set of susceptibility loci, we can again calculate the Bayes factors for the joint association at those loci simultaneously. The most significant ones are reported. The local ancestries at the AIM loci are unobserved and imputed from the HMM. The imputation uncertainty could be properly accounted for by calculating weighted average of the Bayes factors for each imputed local ancestry dataset, which is similar to the strategy used by Guan and Stephens (2008) in imputation-based association mapping for testing untyped variants.

Simulation studies

We carried out simulation studies to assess the performance of GLEAM in terms of type I error rate and power under various scenarios and compared it with the method based on Bayesian likelihood ratio (BLR) by Patterson *et al.* (2004), which is implemented by the software ANCESTRYMAP (http://genepath.med.harvard.edu/~reich/Software.htm) as well as regularized regression methods Lasso and elastic net (Tibshirani 1996; Zou and Hastie 2005; Friedman *et al.* 2010). GLEAM and ANCESTRYMAP use slightly different HMMs to impute the local ancestries and regularized regression methods require given local ancestries. Because of these differences, we assumed the true local ancestries were given and focused on evaluating the ability of localizing susceptibility loci instead of estimating local ancestries. Our simulations were based on empirical data of local ancestries for 1001 African-American subjects from the HPHB Study (Miranda *et al.* 2009), with 1296 AIM loci measured across the genome. The MATLAB codes for simulating and analyzing the data are included in a Supporting Information folder online.

We started by investigating the type I error rates for the local ancestries that were scattered around different regions of the genome and in linkage equilibrium. Under this scenario, the falsely localized

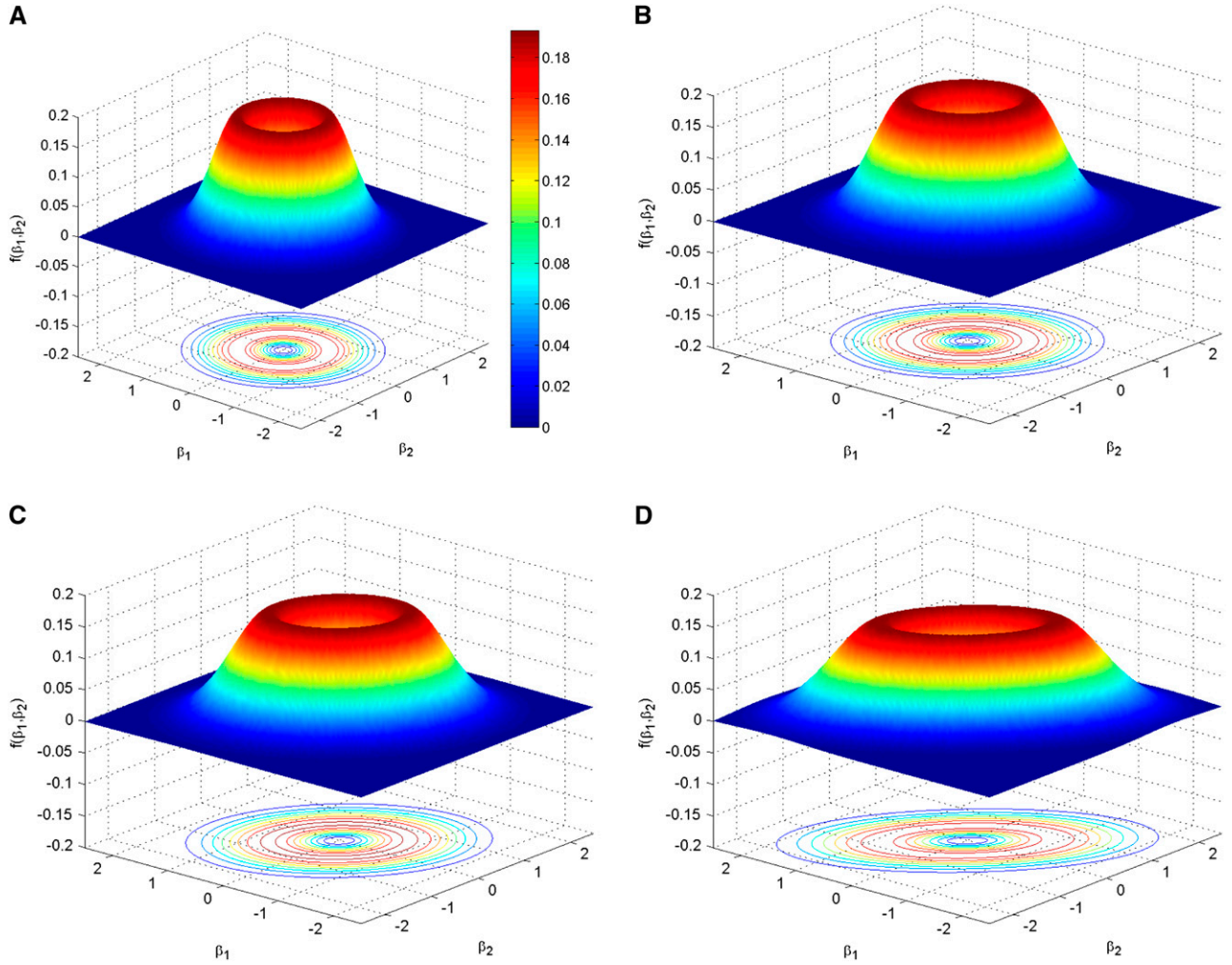

Figure 2 Bivariate quadratic normal moment prior with $\tau\sigma^2 = 0.1$ and $\boldsymbol{\Sigma} = (\boldsymbol{S}'\boldsymbol{S})^{-1}$, where $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2]'$, $\boldsymbol{S}_1 = (S_{1,1}, S_{1,2}, \ldots, S_{1000,1})'$, $\boldsymbol{S}_2 = (S_{1,2}, S_{2,2}, \ldots, S_{1000,2})'$ and $S_{i1} \in \{0, 1, 2\}$ and $S_{i2} \in \{0, 1, 2\}$. We introduce correlation between S_{i1} and S_{i2} through the latent variables (Z_{i1}, Z_{i2}), where $Z_{i1} \overset{\text{iid}}{\sim} \text{N}_1(0, 1)$, $Z_{i2} \overset{\text{iid}}{\sim} \text{N}_1(0, 1)$ and $\text{Cov}(Z_{i1}, Z_{i2}) = \rho$. let $S_{i1} = 0$ if $Z_{i1} \leq C_0$; $S_{i1} = 2$ if $Z_{i1} > C_1$; and $S_{i1} = 0$ otherwise with $C_0 = \Phi^{-1}((1-p_a)^2)$ and $C_1 = \Phi^{-1}(1 - p_a^2)$ where $\Phi^{-1}(\cdot)$ denotes normal inverse cumulative distribution function. We consider four scenarios when $p_a = 0.8$: (A) $\rho = 0$; (B) $\rho = 0.25$; (C) $\rho = 0.5$; and (D) $\rho = 0.75$ with contours drawn beneath the probability density function's surface.

AIM locus would be in the region remote from the true disease causing locus, which leads to a false positive finding. We first randomly sampled 1000 AIM loci with replacement from 1296 AIM loci for 1000 subjects. At each AIM locus, we simulated the local ancestries measured by the number of alleles from the African ancestral population from their maximum *a posteriori* (MAP) frequency estimates under the assumption of Hardy-Weinberg equilibrium. Ten sets of trait data were then generated such that we were able to assess the type I error rates under the genome-wide threshold level (*e.g.*, $\alpha = 10^{-4}$), by using the following null model for continuous traits: $y_i = \alpha E_i + \varepsilon_i$ and for binary traits, $\text{logit}\{\text{Prob}(y_i = 1)\} = \alpha E_i$; where the continuous risk covariate E_i and the measurement error ε_i followed standard normal distributions. We considered two situations whereby $\alpha = 0$ in the absence of a covariate effect and $\alpha = 1$ in the presence of a covariate effect.

We next examined power under the single locus alternative models. We simulated 100 sets of traits. Each set included 1000 subjects and one disease associated local ancestry whose location was randomly sampled from 259 AIM loci, where the proportion of African ancestral population ranged from 0.8321 to 0.8817 and was on the top 20% percentile among 1296 AIM loci. Given the local ancestry S_i, continuous covariates $\boldsymbol{E}_i$ and measurement error ε_i generated same as that for the null model, continuous traits were simulated from $y_i = \alpha E_i + \beta S_i + \varepsilon_i$ and binary traits from $\text{logit}\{\text{Prob}(y_i = 1)\} = \alpha E_i + \beta S_i$. Under both models, the β was specified as $\beta = c \times$ proportion of African ancestral population which reflected the *a priori* observation that the locus with the larger proportion of the high-risk ancestral (here African American) population usually demonstrated stronger association with the traits. For continuous traits, we chose the values of effect size multiplier c as 0.2, 0.25, 0.3, 0.35, and 0.4 respectively, with the largest possible effect size equal to 0.3527. Similarly, we picked the c values as 0.4, 0.5, 0.6, 0.7, and 0.8 for binary traits with the largest possible odds ratio equal to 1.8537.

We further considered a multilocus alternative model where two local ancestries were associated with the traits and there existed admixture linkage disequilibrium. To do so, we generated an artificial chromosome composed of two pieces from chromosome 1 and chromosome 4 with the length 139.50 Mb and 114.88 Mb, respectively, for 1000 subjects, based on empirical data on local ancestries from HPHB study. In the middle of each chromosome piece with 51 loci,

there is one locus whose proportion of African ancestry population was among the highest in all 1296 AIM loci. In the simulations, those two loci are assumed to be associated with traits. We generated 100 sets of continuous and binary traits respectively, each of which was simulated similarly to the single locus alternative model except with two local ancestries involved and both effect size multiplier c values set at 0.7 for continuous traits and 0.35 for binary traits.

The simulated datasets were analyzed by the GLEAM and the BLR method. Because the BLR method was primarily developed for binary traits, the BLR method required transformation of continuous traits into binary ones, such as defining the subjects with top 20% traits as the cases and the one with bottom 20% traits as controls.

RESULTS

Simulation studies

Figure 3 presents the empirical type I error rates for both the binary and continuous traits, with or without covariate effects. For the GLEAM and the BLR methods, we chose a threshold of 2 for $log_{10}BF(\mathbf{y})$ to control the genome-wide type I error rates. The regularization parameters of Lasso and elastic net are chosen with the minimal cross validation error. The loci with nonzero regression coefficients are regarded as the ones associated with the traits. As illustrated in Figure 3A and Figure 3B, under the null model that all the local ancestries are in linkage equilibrium, the type I error rate is controlled at a low level with the median around 5×10^{-4} for

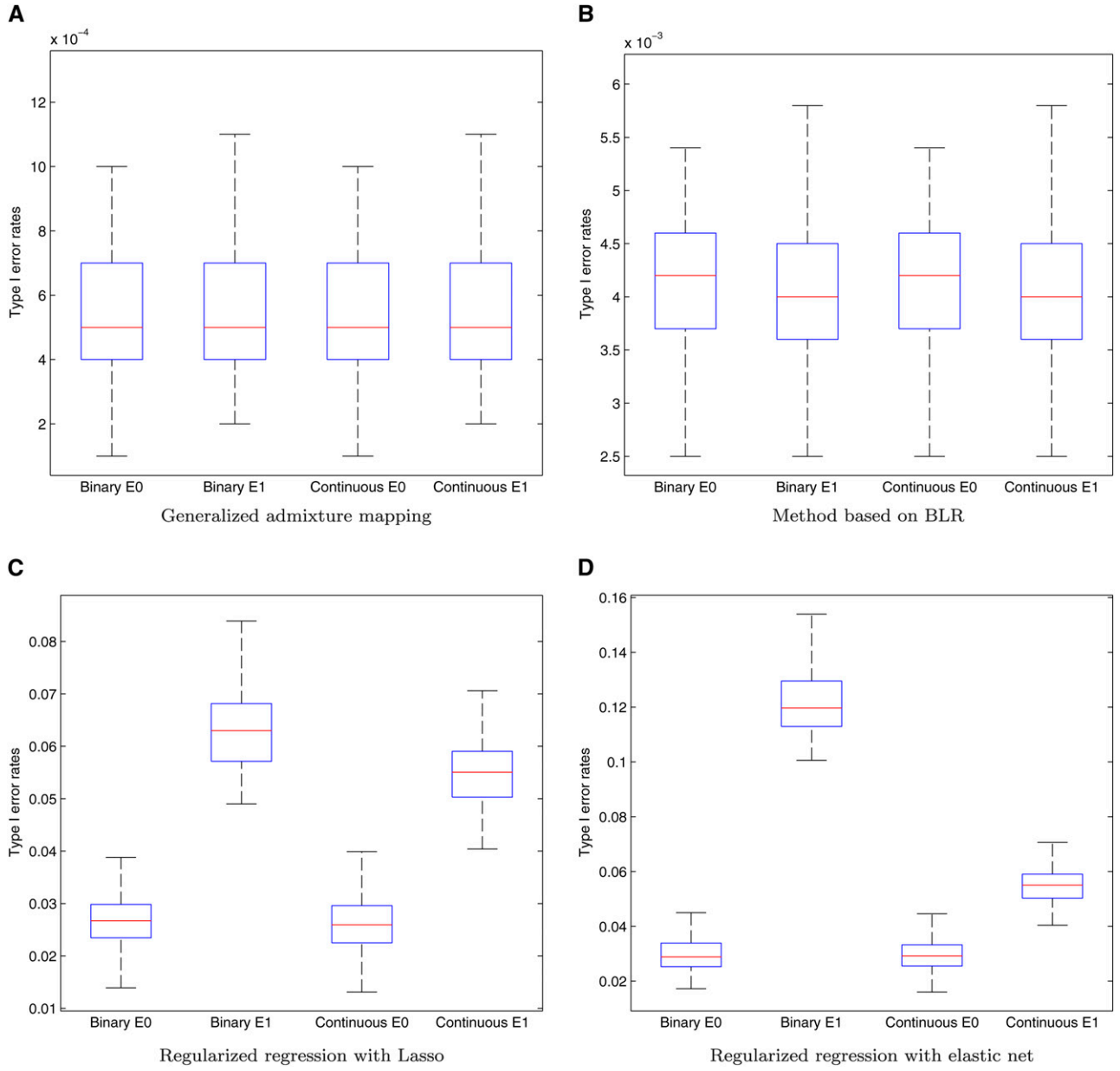

Figure 3 The type I error rates under the null model (note the different scaling of the Y-axis for panels). The type I error rates are presented for both the binary and continuous traits respectively, with or without covariate effect (denoted by E1 and E0, respectively). For each simulated dataset, we calculate one type I error rate for each method. The results for 100 replications are summarized by the box plots, where the center bar is median, bottom and top of the box are the 25th and 75th percentile and the whiskers stretch out until the extreme values. (A) Generalized admixture mapping; (B) Method based on BLR; (C) Regularized regression with Lasso; (D) Regularized regression with elastic net.

GLEAM and 4.2×10^{-3} for the BLR method. In both cases, those type I error rates seem overly conservative. However, in the application to real data, slight admixture linkage disequilibrium between the AIM loci will significantly inflate the type I error rate close to the nominal levels (*i.e.*, α = 0.05 or 0.005), which is discussed in the later paragraphs. Comparing Figure 3A and Figure 3B reveals that the type I error rates of GLEAM are consistently smaller than those of the method based on BLR and are little affected by the presence of covariate effects when properly adjusted. The covariates are not considered by the BLR method and have a mixed effect on type I error rates, where the median is slightly reduced with the maximal type I error rates increased. For the regularized regression methods Lasso and elastic net, the type I errors are significantly inflated, as shown in Figure 3C and Figure 3D. In addition, when a nonzero covariate presents, the type I errors will further increase.

Power of the methods also was evaluated for binary and continuous traits under the single locus alternative model, with or without covariate effects. We considered various effect sizes of local ancestries with the results shown in Figure 4. For the binary trait, when the effect size is small, the BLR method performs better with larger power. With the increment of the effect sizes, GLEAM gradually outperforms the BLR method. For both methods, covariates have

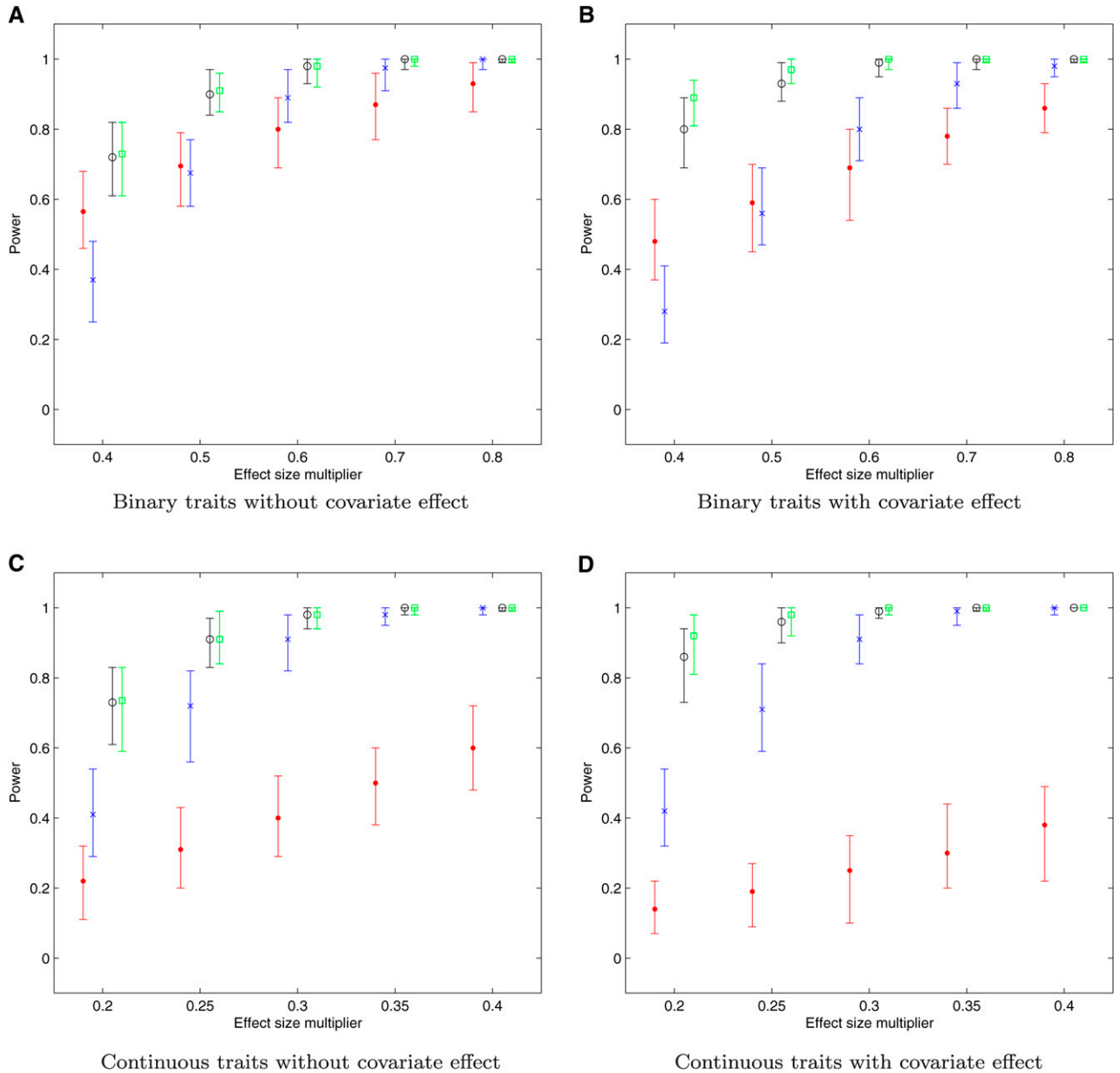

Figure 4 Powers for single locus alternative models. Power is calculated for each dataset with 100 replications total for the binary or continuous traits simulated under the single locus alternative model with or without covariate effect. The × indicates the median of powers by the GLEAM and • denotes the median of powers by the method based on Bayesian likelihood ratio; ◦ denotes the median of regularized regression with lasso; ◦ denotes the median of regularized regression with elastic net. The whiskers on each bar represent the minimal and maximal powers respectively. The effect sizes of local ancestries are equal to the multiplication of effect size multiplier *c* and the proportion of African ancestry population. (A) Binary traits without covariate effect; (B) Binary traits with covariate effect; (C) Continuous traits without covariate effect; (D) Continuous traits with covariate effect.

moderate effects on power, which is more obvious for the smaller effect sizes. For the continuous trait, the GLEAM performs significantly better at each effect size. These results were expected since the BLR method discards part of the dataset in order to transform the continuous trait into the binary one (case *vs.* control), which inevitably loses power. For all situations considered, the power of the GLEAM approach increases with the increment of the local ancestry effect size, most rapidly when the effect sizes are smaller and then levels off with larger effect sizes. In comparison, the power of the BLR method increases roughly linearly. Both GLEAM and BLR are less powerful than the regularized methods especially when the effect sizes are small. With the growth of the effect size, the power of GLEAM will quickly increase and be comparable to the ones of regularized regression.

To understand the impact of admixture linkage disequilibrium on type I error rates and to evaluate the ability of localizing multiple loci simultaneously, we generated a set of artificial chromosomes as described previously, where two loci were associated with the traits, named as Locus 1 and Locus 2. Besides Locus 1 and Locus 2, we divided the remaining loci into three regions: region 1 (REG1) with 42 loci and region 2 (REG2) with 35 loci, where the admixture linkage disequilibrium measured by the correlation coefficient between a given locus at these regions and Locus 1 or Locus 2 was larger than 0.12 respectively; and region 3 (REG3), the unassociated loci which did not belong to region 1 and region 2. Strictly speaking, the identified loci except Locus 1 and Locus 2 were all false positives. However, in contrast to the loci found in region 3, which were completely false findings, the loci identified in Region 1 and Region 2 were partially correct and could be regarded as low-resolution findings instead, because the true associated locus did exist in the nearby region. Therefore, we evaluated the false positives in three regions separately. An ideal method under the prespecified genome-wide threshold would lead to few completely false positives in region 3 and to a small number of partially false positives in regions 1 and 2, while being able to identify the true associated loci with high frequency.

Table 1 summarizes the frequencies of identified loci for each locus or locus combination at different regions by GLEAM, BLR and regularized regression methods. For the GLEAM method, we applied the two-step approach outlined in the "Generalized admixture mapping procedure" subsection. The results by applying the first step only (GLEAM1) and by applying the two-step approach (GLEAM2) were both presented. For binary traits, both the BLR method and GLEAM1 could localize both Locus 1 and Locus 2 with high power. The type I error rates in region 1 were around the nominal level (0.025 and 0.003, respectively). The type I error rates in region 1 and region 2 were higher than the ones in region 3, which would decrease the resolution of the finding. Compared with GLEAM1, further applying the second step of generalized admixture mapping procedure (GLEAM2) could significantly improve the resolution by reducing the type I errors in region 1 (from 0.013 to 0.002) and region 2 (from 0.014 to 0.003). For continuous traits, GLEAM2 also performed best with much higher power and lower type I rate than the BLR method. Similar to the simulation results under null and single locus alternative model, regularized regressions show marginally higher power at the cost of inflated type I error rate, *e.g.*, power 1 for detecting both locus 1 and 2 with type I error rates 0.023 of Lasso and 0.029 of elastic net at region 3 for the continuous trait.

Table 1 The frequency of identified loci for each locus or locus combination at different regions of the artificial chromosome

Trait	Method	REG1	REG2	REG3	Locus1	Locus2	Locus1/2[a]
Binary	BLR	0.103	0.047	0.025	0.000	0.000	1.000
	GLEAM1[b]	0.013	0.014	0.003	0.020	0.020	0.960
	GLEAM2[c]	0.002	0.003	0.001	0.030	0.030	0.940
	Lasso	0.030	0.025	0.017	0.000	0.000	1.000
	Elatic net	0.045	0.038	0.025	0.000	0.000	1.000
Continuous	BLR	0.035	0.018	0.011	0.030	0.400	0.560
	GLEAM1	0.021	0.017	0.004	0.030	0.000	0.970
	GLEAM2	0.004	0.003	0.002	0.040	0.000	0.960
	Lasso	0.039	0.031	0.023	0.000	0.000	1.000
	Elatic net	0.049	0.037	0.029	0.000	0.000	1.000

BLR, Bayesian likelihood ratio; GLEAM, generalized admixture mapping.
[a] The combination of Locus 1 and Locus 2.
[b] Applying the first step of generalized admixture mapping procedure only;
[c] Applying both steps of generalized admixture mapping procedure;

Application

We applied our approach to data from the Healthy Pregnancy, Healthy Baby (HPHB) study, which is a prospective cohort study of pregnant women aimed at identifying genetic, social, and environmental contributors to disparities in adverse birth outcomes in the Southern United States (Miranda *et al.* 2009). Consistent with previous studies, African-American women in HPHB have greater risk for maternal hypertension than white women during the pregnancy, which contributes to the poor birth outcomes (Allen *et al.* 2004). Even within the African-American subpopulation, some women have much greater blood pressure, and we hypothesize that one possible contributor may be the percentage of African ancestry. To explore this hypothesis, we applied GLEAM to investigate the association between the averaged maternal mean arterial pressure (MAP), defined as (1/3 × systolic blood pressure) + (2/3 × systolic blood pressure), during 24 to 28 weeks of pregnancy and local ancestries among these pregnant African-American women. Clinical and genetic data were available for 1004 non-Hispanic black women. A total of 1509 SNP AIMs were genotyped using the Illumina African-American admixture panel. After quality control measures described previously (A. E. Ashley-Koch, Me. E. Garrett, S. Edwards, K. S. Quinn, G. K. Swamy, and M. L. Miranda, unpublished results), the dataset consisted of 1001 non-Hispanic black women with 1296 AIMs.

The proposed GLEAM approach was applied to this dataset to identify the local ancestry associated with the averaged maternal MAP, a continuous trait, while we adjusted for mother's age. The local ancestries were multiply imputed based on the HMM. We first examined the marginal association between the trait and local ancestries, one locus a time. The results are summarized in Figure 5, where one local ancestry on the chromosome 2 was identified with its log_{10}(Bayes factor) = 2.05 exceeding the threshold 2. With only one local ancestry localized, the second step of the generalized admixture mapping procedure was unnecessary. The same data were analyzed by the BLR method, which treated the subjects with averaged maternal MAP more than 93.67 (top 20% quantile) as cases and the ones with averaged maternal MAP less than 79.33 (bottom 20% quantile) as control. No local ancestry was identified as being associated with the averaged maternal MAP with this approach, presumably due to its relatively low power compared with the GLEAM approach.

DISCUSSION

When the admixture linkage disequilibrium is used, admixture mapping is an indispensable tool to localize the alleles that are associated with the qualitative or quantitative traits and diseases that vary in prevalence across the ancestral populations. In this article, we propose a flexible and powerful generalized admixture mapping

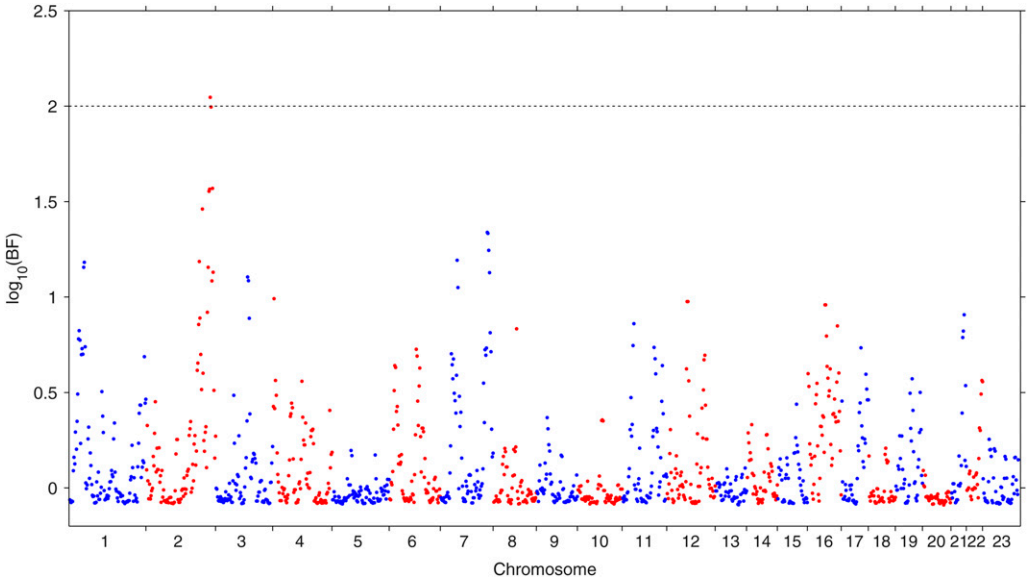

Figure 5 Manhattan plot of log_{10}(Bayes factor) for the association between the averaged maternal MAP during 24 to 28 weeks of pregnancy and genome-wide local ancestries among 1001 African-American subjects.

approach, which is based on the generalized linear model and is able to incorporate admixture prior information by using the quadratic normal moment prior and to adjust for covariates. The proposed method is applicable to both qualitative and quantitative traits with satisfactory power while controlling the type I error rates at a low level, and is able to be easily implemented as we demonstrated with our HPHB example.

In addition to the flexibility to handle different types of traits, other attractive generalizations include consideration of multiple loci simultaneously. It is known that admixture linkage disequilibrium extends much further than haplotype linkage disequilibrium. Consequently, if we only examine one locus a time, the local ancestries which are highly correlated to the true disease associated local ancestry tend to be identified as significant ones as well. As demonstrated by the simulations, those false positives can be significantly reduced by considering multiple susceptible loci simultaneously, which reduce the type I error rates and improve the mapping resolution. In addition, GLEAM specifies a hidden Markov model treating the recombination rates varying across the genome, which allows us to infer the recombination "hotspots" in admixture population. Moreover, within the generalized linear model framework, it is straightforward to extend the current method to populations with more than to two ancestral populations, such as Hispanic populations, by adding extra ancestry population covariates. It is also easy to consider the interaction between the local ancestries and covariates with the properly specification of the priors on interaction coefficients.

ACKNOWLEDGMENTS

This work was supported by Award Number R01ES017436 from the National Institute of Environmental Health Sciences, by funding from the National Institutes of Health (5P2O-RR020782-O3), and the U.S. Environmental Protection Agency (RD-83329301-0) and by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Bethesda, Maryland. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences, the National Institutes of Health, or the U.S. Environmental Protection Agency.

LITERATURE CITED

Allen, V., K. S. Joseph, K. Murphy, L. Magee, and A. Ohlsson, 2004 The effect of hypertensive disorders in pregnancy on small for gestational age and stillbirth: a population based study. BMC Pregnancy Childbirth 4: 17.

Basu, A., H. Tang, C. E. Lewis, K. North, J. D. Curb *et al.*, 2009 Admixture mapping of quantitative trait loci for blood lipids in African-Americans. Hum. Mol. Genet. 18: 2091.

Bellman, E. R., 1961 *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, New Jersey.

Cheng, C. Y., D. Reich, T. Y. Wong, R. Klein, B. E. K. Klein *et al.*, 2010 Admixture mapping scans identify a locus affecting retinal vascular caliber in hypertensive African Americans: the atherosclerosis risk in communities (ARIC) study. PLoS Genet. 6: e1000908.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567.

Freedman, M. L., C. A. Haiman, N. Patterson, G. J. McDonald, A. Tandon *et al.*, 2006 Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc. Natl. Acad. Sci. USA 103: 14068.

Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33: 1.

Guan, Y., and M. Stephens, 2008 Practical issues in imputation-based association mapping. PLoS Genet. 4: e1000279.

Guan, Y., and M. Stephens, 2011 Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. 5: 1780–1815.

Hoggart, C. J., E. J. Parra, M. D. Shriver, C. Bonilla, R. A. Kittles *et al.*, 2003 Control of confounding of genetic associations in stratified populations. Am. J. Hum. Genet. 72: 1492–1504.

Johnson, V. E., and D. Rossell, 2010 On the use of non-local prior densities in Bayesian hypothesis tests. J. R. Stat. Soc., B 72: 143–170.

Kao, W. H. L., M. J. Klag, L. A. Meoni, D. Reich, Y. Berthier-Schaad *et al.*, 2008 MYH9 is associated with nondiabetic end-stage renal disease in African Americans. Nat. Genet. 40: 1185–1192.

McCullagh, P., and J. A. Nelder, 1989 *Generalized Linear Models*, Chapman & Hall/CRC, London.

McKeigue, P. M., 1998 Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am. J. Hum. Genet. 63: 241–251.

McKeigue, P. M., 2005 Prospects for admixture mapping of complex traits. Am. J. Hum. Genet. 76: 1–7.

Miranda, M. L., P. Maxson, and S. Edwards, 2009 Environmental contributions to disparities in pregnancy outcomes. Epidemiol. Rev. 31: 67.
Montana, G., and J. K. Pritchard, 2004 Statistical tests for admixture mapping with case-control and cases-only data. Am. J. Hum. Genet. 75: 771–789.
Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321.
Patterson, N., N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler *et al.*, 2004 Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. 74: 979–1000.
Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5: e1000519.
Reich, D., N. Patterson, P. L. De Jager, G. J. McDonald, A. Waliszewska *et al.*, 2005 A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. Nat. Genet. 37: 1113–1118.
Reich, D., N. Patterson, V. Ramesh, P. L. De Jager, G. J. McDonald *et al.*, 2007 Admixture mapping of an allele affecting interleukin 6 soluble receptor and interleukin 6 levels. Am. J. Hum. Genet. 80: 716–726.
Scherer, M. L., M. A. Nalls, L. Pawlikowska, E. Ziv, G. Mitchell *et al.*, 2010 Admixture mapping of ankle–arm index: identification of a candidate locus associated with peripheral arterial disease. J. Med. Genet. 47: 1.
Scott, S. L., 2002 Bayesian methods for hidden Markov models. J. Am. Stat. Assoc. 97: 337–351.
Smith, M. W., and S. J. O'Brien, 2005 Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat. Rev. Genet. 6: 623–632.
Smith, M. W., N. Patterson, J. A. Lautenberger, A. L. Truelove, G. J. McDonald *et al.*, 2004 A high-density admixture map for disease gene discovery in African Americans. Am. J. Hum. Genet. 74: 1001–1013.
Tandon, A., N. Patterson, and D. Reich, 2011 Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. Genet. Epidemiol. 35: 80–83.
The International HapMap Consortium, 2005 A haplotype map of the human genome. Nature 437: 1299–1320.
Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc., B 73: 273–282.
Winkler, C. A., G. W. Nelson, and M. W. Smith, 2010 Admixture mapping comes of age. Annu. Rev. Genomics Hum. Genet. 11: 65–89.
Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, 2009 Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.
Yang, J. J., C. Cheng, M. Devidas, X. Cao, Y. Fan *et al.*, 2011 Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nat. Genet. 43: 237–241.
Zhu, X., R. S. Cooper, and R. C. Elston, 2004 Linkage analysis of a complex disease through use of admixed populations. Am. J. Hum. Genet. 74: 1136–1153.
Zhu, X., A. Luke, R. S. Cooper, T. Quertermous, C. Hanis *et al.*, 2005 Admixture mapping for hypertension loci with genome-scan markers. Nat. Genet. 37: 177–181.
Zhu, X., J. H. Young, E. Fox, B. J. Keating, N. Franceschini *et al.*, 2011 Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the care consortium. Hum. Mol. Genet. 20: 2285–2295.
Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. 67: 301–320.

Communicating editor: D.-J. De Koning

APPENDICES

Hidden Markov Model

For a population-based design, suppose we have I unrelated subjects, each of which has the same set of J AIMs recorded. The local ancestry is measured by $S_{ij} \in \{0, 1, 2\}$, the number of alleles from the high-risk population A (*e.g.*, African) for the ith subject and the jth AIM. S_{ij} is unknown and will be imputed using the HMM. For African Americans with African and European ancestral populations, HMM assumes that given the S_{ij}, the distribution of $X_{ij} \in \{0, 1, 2\}$, the number of variant alleles, is independent of other $S_{ij'}$ and $X_{ij'}$ with $j' \neq j$ and is specified by the observation probability mass matrix $\boldsymbol{P}_j = \{p_j(m, n)\}_{3\times3}$ with $p_j(m, n) = \text{Prob}(X_{ij} = n | S_{ij} = m)$ and

$$\boldsymbol{P}_j = \begin{array}{c} \\ S_{ij}=0 \\ S_{ij}=1 \\ S_{ij}=2 \end{array} \begin{array}{c} \begin{array}{ccc} X_{ij}=0 & X_{ij}=1 & X_{ij}=2 \end{array} \\ \begin{pmatrix} (1-p_j^B)(1-p_j^B) & 2p_j^B(1-p_j^B) & p_j^B p_j^B \\ (1-p_j^A)(1-p_j^B) & p_j^A(1-p_j^B)+p_j^B(1-p_j^A) & p_j^A p_j^B \\ (1-p_j^A)(1-p_j^A) & 2p_j^A(1-p_j^A) & p_j^A p_j^A \end{pmatrix} \end{array},$$

where p_j^A is the minor allele probability at loci j in the high-risk population A and p_j^B is the corresponding probability in the low-risk population B.

The latent states $\boldsymbol{S}_i = \{S_{ij}\}_{1\times J}$, tagging the status of the ancestry blocks, are unobserved and modeled by a Markov chain which considers the genetic recombination events. Let ρ_i denote the genome-wide proportion of alleles from the high-risk population A for subject i, $\boldsymbol{Q}_{i0} = [(1-\rho_i)^2, 2\rho_i(1-\rho_i), \rho_i^2]'$ initial state vector, $R_{ij} \in \{0, 1, 2\}$ the number of recombination events between AIM loci $j-1$ and j, $\boldsymbol{Q}_i^{(r)} = \{q_i^{(r)}(m, n)\}_{3\times3}$ the conditional state transition matrix given r recombination events between the neighboring AIM loci with $q_i^{(r)}(m, n) = \text{Prob}(S_{ij} = n | S_{i(j-1)} = m, R_{ij} = r)$. The Markov chain $\boldsymbol{S}_i$ is governed by the state transition matrix $\boldsymbol{Q}_{ij} = \{q_{ij}(m, n)\}_{3\times3}$ with $q_{ij}(m, n) = \text{Prob}(S_{ij} = n | S_{i(j-1)} = m)$. $\boldsymbol{Q}_{ij} = \sum_{r=0}^{2} \boldsymbol{Q}_i^{(r)} \text{Prob}(R_{ij} = r)$, where $\boldsymbol{Q}_i^{(0)}$, $\boldsymbol{Q}_i^{(1)}$ and $\boldsymbol{Q}_i^{(2)}$ are specified as

$$\boldsymbol{Q}_i^{(0)} = \begin{array}{c} \\ S_{i(j-1)}=0 \\ S_{i(j-1)}=1 \\ S_{i(j-1)}=2 \end{array} \begin{array}{c} \begin{array}{ccc} S_{ij}=0 & S_{ij}=1 & S_{ij}=2 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array},$$

$$\boldsymbol{Q}_i^{(1)} = \begin{array}{c} \\ S_{i(j-1)}=0 \\ S_{i(j-1)}=1 \\ S_{i(j-1)}=2 \end{array} \begin{array}{c} \begin{array}{ccc} S_{ij}=0 & S_{ij}=1 & S_{ij}=2 \end{array} \\ \begin{pmatrix} 1-\rho_i & \rho_i & 0 \\ \frac{1}{2}(1-\rho_i) & \frac{1}{2} & \frac{1}{2}\rho_i \\ 0 & 1-\rho_i & \rho_i \end{pmatrix} \end{array},$$

$$\boldsymbol{Q}_i^{(2)} = \begin{array}{c} \\ S_{i(j-1)}=0 \\ S_{i(j-1)}=1 \\ S_{i(j-1)}=2 \end{array} \begin{array}{c} \begin{array}{ccc} S_{ij}=0 & S_{ij}=1 & S_{ij}=2 \end{array} \\ \begin{pmatrix} (1-\rho_i)^2 & 2\rho_i(1-\rho_i) & \rho_i^2 \\ (1-\rho_i)^2 & 2\rho_i(1-\rho_i) & \rho_i^2 \\ (1-\rho_i)^2 & 2\rho_i(1-\rho_i) & \rho_i^2 \end{pmatrix} \end{array},$$

and $R_{ij} \sim \text{Bin}(2, \gamma_j)$ a binomial distribution with γ_j the probability that a recombination event occurs between the neighboring AIM loci in a single chromosome. Consequently, we can get,

$$\boldsymbol{Q}_{ij} = \begin{array}{c} \\ S_{i(j-1)}=0 \\ S_{i(j-1)}=1 \\ S_{i(j-1)}=2 \end{array} \begin{array}{c} \begin{array}{ccc} S_{ij}=0 & S_{ij}=1 & S_{ij}=2 \end{array} \\ \begin{pmatrix} (1-\gamma_j\rho_i)^2 & 2\gamma_j\rho_i(1-\gamma_j\rho_i) & \gamma_j^2\rho_i^2 \\ \gamma_j(1-\rho_i)(1-\gamma_j\rho_i) & \left\{1-\gamma_j(1-\rho_i)\right\}(1-\gamma_j\rho_i)+\gamma_j^2\rho_i(1-\rho_i) & \gamma_j\rho_i\left\{1-\gamma_j(1-\rho_i)\right\} \\ \gamma_j^2(1-\rho_i)^2 & 2\gamma_j(1-\rho_i)\left\{1-\gamma_j(1-\rho_i)\right\} & \left\{1-\gamma_j(1-\rho_i)\right\}^2 \end{pmatrix} \end{array}$$

We further specify informative prior distributions for the parameters p_j^A, p_j^B, γ_j and ρ_i involved in the HMM. Although the p_j^A of the high-risk population A is unknown, we have information on p_{0j}^A, the proportion of the variant allele j in a subpopulation of high-risk population A (*e.g.* YRI for African), from the HapMap or 1000 genome projects. Hence, we expect that p_j^A would be close to p_{0j}^A and specify $p_j^A \sim \text{Beta}(\tau^A p_{0j}^A, \tau^A(1-p_{0j}^A))$ with the expectation $E(p_j^A) = p_{0j}^A$ and $\tau^A \sim \cup\,[50, 1000]$ a uniform distribution to reflect the uncertainty in borrowing the subpopulation information. A similar specification is chosen for p_j^B based on the proportion of the variant allele j in a subpopulation of low-risk population B (*e.g.*, CEU for European). As for γ_j, it is well known that the recombination probability is roughly proportional to d_j the genetic distance between $(j-1)$th and jth AIM loci. A common choice is $\gamma_j = 1-\exp(-\lambda d_j)$ with $\lambda = 6$ the number of recombination events per Morgan since admixture (Falush *et al.* 2003; Patterson *et al.* 2004). However, recombination "hotspots" can occur along the chromosomes where the recombination probabilities are much greater than the other regions (Myers *et al.* 2005). For this reason, we avoid the aforementioned parametric specification of γ_j. Instead, we let $\gamma_j \sim \text{Beta}\,(\tau^\gamma\gamma_{0j}, \tau^\gamma(1-\gamma_{0j}))$ with the expectation $E(\gamma_j) = \gamma_{0j} = 1-\exp(-\lambda d_j)$. Hence, on average the probability of recombination is proportional to the genetic distance while allowing significant deviation (*e.g.* 'hotspots') from the average. The deviation is measured by τ^γ with $Var(\gamma_j) = \frac{\gamma_{0j}(1-\gamma_{0j})}{\tau^\gamma+1} = \mu_0$. In addition, for the admixed population, we often have knowledge about the proportions of ancestral populations at the population level. For example, the African American population in general consists of 80% African ancestral population and 20% European ancestral population (Smith and O'Brien, 2005; Winkler *et al.* 2010). We borrow this population level information to specify ρ_i, the subject specific proportion of high-risk population A, by letting $\rho_i \sim \text{Beta}\,(\tau^\rho\rho_{0i}, \tau^\rho(1-\rho_{0i}))$ with ρ_{0i} (*e.g.* 0.8 for African American) and $Var(\rho_i) = \frac{\rho_{0i}(1-\rho_{0i})}{\tau^\rho+1} = \nu_0$.

We use an MCMC algorithm to sample the local ancestries $\boldsymbol{S}_i$ for $i = 1.\,2, \ldots, I$, along with other parameters. The details of MCMC are given as follows.

MCMC algorithm for HMM

We propose an MCMC algorithm for posterior computation of HMM as follows.

(1) Impute the missing AIM X_{ij}^m. Given the $\boldsymbol{P}_j$ and S_{ij}, $X_{ij}^m \in \{0, 1, 2\}$ can be easily sampled with probability mass $p_j(S_{ij}, X_{ij}^m)$.
(2) Update the latent states $\boldsymbol{S}_i$ for $i = 1.\,2, \ldots, I$. Given the $\boldsymbol{Q}_{i0}$, $\boldsymbol{Q}_i^{(r)}$ and $\boldsymbol{R}_i = \{R_{ij}\}_{1\times J}$, we will use the forward filtering backward sampling (FFBS) algorithm (Scott 2002) to sample the $\boldsymbol{S}_i$ in one block. The FFBS algorithm mixes more rapidly comparing to the direct Gibbs sampler which samples one S_{ij} a time conditional on the remains of $\boldsymbol{S}_i$. Let $\boldsymbol{X}_{i1}^j = [X_{i1}, X_{i2}, \cdots, X_{ij}]'$ and $\boldsymbol{R}_i = [R_{i1}, R_{i2}, \ldots, R_{iJ}]'$. We begin the FFBS algorithm by calculating $\boldsymbol{Q}_{ij}^F = \{q_{ij}^F(m,n)\}_{3\times 3}$ with $q_{ij}^F(m,n) = \text{Prob}(S_{i(j-1)} = m, S_{ij} = n|\boldsymbol{X}_{i1}^j, \boldsymbol{R}_i)$ recursively for $j = 1.\,2, \ldots, J$ as

$$\begin{aligned} q_{ij}^F(m,n) &= \text{Prob}\left(S_{i(j-1)} = m, S_{ij} = n|\boldsymbol{X}_{i1}^j, \boldsymbol{R}_i\right) \\ &= \frac{\text{Prob}\left(S_{i(j-1)} = m, S_{ij} = n, X_{ij}|\boldsymbol{X}_{i1}^{j-1}, \boldsymbol{R}_i\right)}{\text{Prob}\left(X_{ij}|\boldsymbol{X}_{i1}^{j-1}, \boldsymbol{R}_i\right)} \\ &= \frac{q_{i(j-1)}^F(m)q_i^{(r)}(m,n)p_j\left(n, X_{ij}\right)}{\text{Prob}\left(X_{ij}|\boldsymbol{X}_{i1}^{j-1}, \boldsymbol{R}_i\right)}, \end{aligned}$$

where $q_{i0}^F(m) = \boldsymbol{Q}_{i0}$, $\text{Prob}(X_{ij}|\boldsymbol{X}_{i1}^{j-1}, \boldsymbol{R}_i) = \sum_{m=0}^{2}\sum_{n=0}^{2}\text{Prob}(S_{i(j-1)} = m, S_{ij} = n, X_{ij}|\boldsymbol{X}_{i1}^{j-1}, \boldsymbol{R}_i)$, and $q_{ij}^F(n) = \sum_{m=0}^{2} q_{ij}^F(m, n)$.

We can then sample the $\boldsymbol{S_i}$ backward from S_{iJ} to S_{i1} with

$$\text{Prob}(\boldsymbol{S}_i|\boldsymbol{X}_i, \boldsymbol{R}_i) = \text{Prob}(S_{iJ}|\boldsymbol{X}_i, \boldsymbol{R}_i)\prod_{j=1}^{J-1}\text{Prob}\left(S_{i(J-j)}|\boldsymbol{S}_{i(J-j+1)}^{J}, \boldsymbol{X}_i, \boldsymbol{R}_i\right),$$

where

$$\text{Prob}(S_{iJ}|\boldsymbol{X}_i, \boldsymbol{R}_i) = q_{iJ}^F(S_{iJ}),$$

$$\begin{aligned}\text{Prob}\left(S_{i(J-j)}|\boldsymbol{S}_{i(J-j+1)}^{J}, \boldsymbol{X}_i, \boldsymbol{R}_i\right) &= \text{Prob}\left(S_{i(J-j)}|S_{i(J-j+1)}, \boldsymbol{X}_{i1}^{J-j+1}, \boldsymbol{R}_i\right)\\ &= \frac{q_{i(J-j+1)}^F\left(S_{i(J-j)}, S_{i(J-j+1)}\right)}{q_{i(J-j+1)}^F\left(S_{i(J-j+1)}\right)}.\end{aligned}$$

The initial state S_{i0} will be sampled with $\text{Prob}(S_{i0}|\boldsymbol{S}_i, \boldsymbol{X}_i, \boldsymbol{R}_i) = \frac{q_{i1}^F(S_{i0}, S_{i1})}{q_{i1}^F(S_{i1})}$.

(3) Update the recombination count $\boldsymbol{R}_i = \{R_{ij}\}_{1\times J}$ for $i = 1, 2, \ldots, I$. R_{ij} is sampled with full conditional probability mass function

$$\text{Prob}\left(R_{ij}|S_{i(j-1)} = m, S_{ij} = n, \boldsymbol{Q}_i^{(0)}, \boldsymbol{Q}_i^{(1)}, \boldsymbol{Q}_i^{(2)}, \gamma_j\right) = \frac{q_i^{(R_{ij})}(m, n)\binom{2}{R_{ij}}\gamma_j^{R_{ij}}\left(1-\gamma_j\right)^{2-R_{ij}}}{\sum_{r=0}^{2}q_i^{(r)}(m, n)\binom{2}{r}\gamma_j^{r}\left(1-\gamma_j\right)^{2-r}}$$

(4) Update recombination probability γ_j from $\text{Beta}(\tau^\gamma\gamma_{0j} + \sum_{i=1}^{I}R_{ij}, \tau^\gamma(1-\gamma_{0j}) + 2I - \sum_{i=1}^{I}R_{ij})$ for $j = 1. 2, \ldots, J$.

(5) Update the proportion ancestry from population A ρ_i from $\text{Bin}(\tau^\rho\rho_{0i} + n_{01}^{(1)} + n_{12}^{(1)} + n_{22}^{(1)} + n_{\cdot 1}^{(2)} + 2n_{\cdot 2}^{(2)}, \tau^\rho(1-\rho_{0i}) + n_{00}^{(1)} + n_{10}^{(1)} + n_{21}^{(1)} + n_{\cdot 1}^{(2)} + 2n_{\cdot 0}^{(2)})$, where $n_{kl}^{(1)} = \sum_{j=1}^{J}I(S_{i(j-1)} = k \text{ and } S_{ij} = l \text{ and } R_{ij} = 1)$ and $n_{\cdot l}^{(2)} = \sum_{j=1}^{J}I(S_{ij} = l \text{ and } R_{ij} = 2)$.

(6) Update $\boldsymbol{Q}_i^{(0)}$, $\boldsymbol{Q}_i^{(1)}$, $\boldsymbol{Q}_i^{(2)}$ and $\boldsymbol{Q}_{i0}$ based on last ρ_i for $i = 1, 2, \ldots, I$.

(7) Update p_j^A and p_j^B for $j = 1. 2, \ldots, J$. Let $n_{kl} = \sum_{i=1}^{I}I(S_{ij} = k \text{ and } X_{ij} = l)$ and n_{11}^{VA} denotes the case that the allele from population A is variant allele when $S_{ij} = 1$ and $X_{ij} = 1$. n_{11}^{VA} is unobserved and can be imputed from $\text{Bin}\left(n_{11}, \frac{p_j^A(1-p_j^B)}{p_j^A(1-p_j^B)+p_j^B(1-p_j^A)}\right)$. p_j^A is then sampled from $\text{Beta}(\tau^A p_{0j}^A + n_{21} + 2n_{22} + n_{11}^{VA}, \tau^A(1-p_{0j}^A) + n_{21} + 2n_{20} + n_{11} - n_{11}^{VA})$; p_j^B is sampled from $\text{Beta}(\tau^B p_{0j}^B + n_{01} + 2n_{02} + n_{11} - n_{11}^{VA}, \tau^B(1-p_{0j}^B) + n_{01} + 2n_{00} + n_{11}^{VA})$

(8) Update $\boldsymbol{P}_j$ based on last p_j^A and p_j^B for $j = 1. 2, \ldots, J$.

(9) Update τ^A and τ^B using Random-Walk Metropolis-Hasting. For τ^A, we propose the new $\tau^{A*} = \tau^A + \epsilon$ where $\epsilon \sim \text{N}_1(0, \sigma_{mh}^2)$. The posterior distribution of $\tau^A, f(\tau^A|\boldsymbol{p}^A) \propto \prod_{j=1}^{J}f_{\text{Beta}}(P_j^A|\tau^A p_{0j}^A, \tau^A(1-p_{0j}^A))I(50 < \tau^A < 1000)$. Then, $\alpha(\tau^A, \tau^{A*}) = \min\left\{\frac{f(\tau^{A*}|\boldsymbol{p}^A)}{f(\tau^A|\boldsymbol{p}^A)}, 1\right\}$. We draw $\mu^A \sim \cup$ [0, 1]. If $\mu^A < \alpha(\tau^A, \tau^{A*})$, then τ^A is replaced by τ^{A*}; otherwise, τ^A is unchanged. Similar update is conducted for τ^B.