Statistics and Its Interface Volume 4 (2011) 317–326

Controlling population structure in human genetic association studies with samples of unrelated individuals

Nianjun Liu, Hongyu Zhao, Amit Patki, Nita A. Limdi and David B. Allison*

In genetic studies, associations between genotypes and phenotypes may be confounded by unrecognized population structure and/or admixture. Studies have shown that even in European populations, which are thought to be relatively homogeneous, population stratification exists and can affect the validity of association studies. A number of methods have been proposed to address this issue in recent years. Among them, the mixed-model based approach and the principal component-based approach have several advantages over other methods. However, these approaches have not been thoroughly evaluated on large human datasets. The objectives of this study are to (1) evaluate and compare the performance of the mixed-model approach and the principal component-based approach for genetic association mapping using human data consisting of unrelated individuals, and (2) understand the relationship between these two approaches. To achieve these goals, we simulate datasets based on the HapMap data under various scenarios. Our results indicate that the mixed-model approach performs well in controlling for population structure/admixture. It has a similar performance as that based on principal component analysis. However, the approach combining mixed-model and principal component analysis does not perform as well as either method itself.

AMS 2000 subject classifications: Primary 62P10, 9208; secondary 6207.

Keywords and phrases: Mixed-effects model, Principal component analysis, Population structure/admixture, Genetic association analysis.

1. INTRODUCTION

Genetic association studies have become prominent with the availability of dense single nucleotide polymorphisms (SNPs), rapid reduction in high-throughput genotyping costs, and successful identifications of thousands of variants with hundreds of human traits. However, it is well known that population structure/admixture can confound genetic association studies (Ewens and Spielman, 1995). This confounding can lead to either false positive or false negative associations. Several simulation studies have affirmed the potential confounding effect of variations in ancestry (Deng, 2001; Heiman et al., 2004a; Heiman et al., 2004b). In the work of Marchini et al. (2004), the authors simulated data with relatively simple and moderate level of population structure (i.e., departures from genotypic frequencies expected under panmixia) observed in samples with a single continental origin (e.g., all Europeans). They concluded that, "For the size of study needed to detect typical genetic effects in common diseases, even the modest levels of population structure within population groups cannot safely be ignored." In addition to the well known admixed populations, such as African-American, recent studies (Campbell et al., 2005; Dolan et al., 2005; Helgason et al., 2005) have shown that even in relatively homogenous populations, such as Irish and Icelander, there still exists population substructure. Hence for genetic studies, attention needs to be paid to control for the confounding effects because of population structure, even if such structure cannot be detected by standard methods, as suggested in Campbell et al. (2005).

Various methods have been proposed to deal with potential confounding from population structure/admixture (Price et al., 2010). Although using differently, these methods all need to use a set of unlinked genetic markers. The genomic control (GC) approach rescales the statistics, which may not follow a central distribution under the null hypothesis when population structure/admixture exists (Devlin and Roeder, 1999). Another approach is called "structured association test (SAT)", where a set of unlinked genetic markers are used to estimate the ancestry probabilities for each individual (Falush et al., 2003; Pritchard and Rosenberg, 1999; Pritchard et al., 2000; Redden et al., 2006; Satten et al., 2001; Tang et al., 2005; Wu et al., 2006). These estimates of ancestry probabilities are then used in association analysis to control for population structure/admixture (Pritchard et al., 2000; Redden et al., 2006; Satten et al., 2001; Yu et al., 2006; Zhao et al., 2007). A third approach uses genetic background derived from a set of independent genetic markers to control for population structure/admixture in association analysis (Bauchet et al., 2007; Chen et al., 2003; Paschou

*Corresponding author.

et al., 2008; Price et al., 2006; Zhang et al., 2003; Zhu et al., 2008; Zhu et al., 2002; Zou et al., 2010). Usually, principal component analysis (PCA) is used to derive genetic background, although there are some variants (Liu and Zhao, 2006; Zhu and Yu, 2009). In recent years, methods based on mixed-model have been proposed to control for population structure/admixture (Kang et al., 2010; Kang et al., 2008; Yu et al., 2006; Zhang et al., 2010). The basic idea is to use a set of genetic markers to estimate pairwise kinship coefficients between the individuals under study. This marker-based kinship is then used in the random effects of the mixed-model to control for potential confounding from population structure/admixture. The latter two approaches (i.e. the approaches based on PCA and mixed-model) have been shown to have several advantages over other approaches (e.g. GC and SAT) and are more preferred. However, the mixed-model approach has been introduced from the animal genetics literature. It has been evaluated mainly using animal and plant data, and has not been thoroughly evaluated on large human datasets (Stich and Melchinger, 2009; Stich et al., 2008; Yu et al., 2006; Zhao et al., 2007). As indicated in Zhao et al. (2007), "Comprehensive simulation studies (with known null distributions) would be required to determine how the different methods perform in terms of the false-positive rate." Similarly, in a recent paper Price et al. (2010) stated that "mixed models are relatively new and untested" in this context.

In this report, we evaluate the performance of the PCA and mixed-model approaches for controlling population structure/admixture in human genetic association studies with samples of unrelated individuals. We simulate various scenarios using human data from the HapMap Project. In addition, we aim to understand the relationship between the PCA approach and the mixed-model approach.

2. METHODS

Following the simulation procedure of Zhu et al. (2008), we simulated data under four scenarios: discrete model with two ancestral populations, admixed model with two ancestral populations, discrete model with three ancestral populations, and admixed model with three ancestral populations. The simulations with discrete models aim to illustrate the performance of the statistical methods with randomly chosen markers when samples are from two and three discrete populations, respectively. The simulations with admixed models aim to illustrate the performance of statistical methods when randomly chosen markers are used for samples from a population admixed by two and three ancestral populations, respectively.

In our simulation study, we generated samples with two or three ancestral populations using the haplotype data released by the HapMap project (Phase 2 HapMap project, HapMap release #21) (http://hapmap.ncbi.nlm.nih.gov/). There are four populations in the HapMap project: 30 trios who are European descendants living in the USA (CEU), 30 trios from Ibadan, Nigeria (YRI), 45 unrelated Han Chinese individuals from Beijing, China (CHB), and 44 unrelated Japanese from Tokyo, Japan (JPT). In these simulations, we used only the unrelated individuals from HapMap data and only haplotype data on chromosome 22 (i.e. 120 European haplotypes from the CEU, 120 African haplotypes from YRI, and 178 East Asian haplotypes from the pooled CHB and JPT). For the discrete model simulations, we used software HapGen (Marchini et al., 2007; Spencer et al., 2009) to create more independent chromosomes than in the original HapMap data. This way, we can have more data and keep the linkage disequilibrium (LD) structure across a chromosome for SNPs that are closely located. We first simulated 50,000 individuals from each of the three aforementioned HapMap populations using their haplotype data on chromosome 22. Therefore we had three large pseudo-populations, each with 50,000 individuals. Then we randomly drew individuals from the simulated larger populations to create datasets. Specifically, for the discrete model simulation with two sub-populations, we randomly drew 400 individuals from the 50,000 European individuals simulated using HapGen, and 600 individuals from the 50,000 African individuals simulated the same way. For the discrete model simulation with three sub-populations, we randomly drew 200 individuals from the simulated European population of 50,000 individuals, 300 individuals from simulated East Asian population of 50,000 individuals, and 500 individuals from the simulated African population of 50,000 individuals. The phenotype was simulated following Zhang et al. (2003). Based on the genotype at the candidate locus, the trait values were generated according to the following model:

$$y_{ij} = \mu_i + \alpha_i A_{ij} + \beta_i D_{ij} + e_{ij},$$

where $\mu_i = \mu_{00} \times R_i$, $\alpha_i = \beta_i = \mu_0 \times R_i$, e_{ij} is a normal random variable with mean 0 and variance 1, A_{ij} and D_{ij} are the additive and dominant genetic scores (additive and dominant genetic scores are the scores for the genotype each individual carries at one locus assuming additive and dominant genetic effects, respectively. Specifically, additive genetic score is the number of minor allele one individual carries at that locus, dominant genetic score is 1 if one individual carries at least one minor allele at that locus, 0 otherwise) of the j-th individual in the i-th subpopulation. The concept behind the modeling is to have a common background disease risk (controlled byμ_{00}) across all populations, specific disease risks (controlled byR_i) for each population, and genetic-related risk (controlled byμ_0) in one model. Following Zhang et al. (2003), in our simulations, we set $R_i = 1$ for individuals from Yoruba, $R_i = 1/2$ for individuals from placeEast Asia, $R_i = 1/4$ for individuals from CEU, and $\mu_{00} = 2$. Furthermore, we set $\mu_0 = 0$ and $\mu_0 = 0.2$ for type I error examination and power evaluation, respectively. A total of 10,000 randomly selected SNPs on chromosome 22 were used in the analyses for calculating the

principal components and kinship matrix. The test marker, the marker that is under study, was randomly chosen on the chromosome but was not included in the above 10,000 random markers.

For the admixed model simulations, we used a Poisson process to mimic the evolution of population admixture, following Zhu et al. (2008). Specifically, using a Poisson process, we first generated haplotype exchange points on the chromosome among the populations. Same as in Zhu et al. (2008), we used an average of 6 crossovers per Morgan to simulate a population that has been admixed for an average of 6 generations. Between two exchange points generated by the Poisson process, we drew haplotype from one ancestral population chosen from a distribution of admixture proportions of Africans, Europeans, and East Asians, which we set to be (0.7, 0.2, 0.1) (for admixed model with two ancestral populations, we used African and European HapMap samples with admixture proportions (0.7, 0.3)), following Zhu et al. (2008). We then applied the same method as in the discrete model to generate an individual's genotypes based on the selected ancestral population. For the admixed model simulation with two ancestral populations, we simulated 1,000 individuals from the European population and the African population. For the admixed model simulation with three ancestral populations, we simulated 1,000 individuals from the European population, the East Asian populations, and the African population. For data with three ancestral populations, the trait values were generated according to the following model:

$$y_i = \mu_i + \alpha_i A_i + \beta_i D_i + e_i,$$

where $\mu_i = \mu_{00} \times (R_1\lambda_1 + R_2\lambda_2 + R_3\lambda_3)$, $\alpha_i = \beta_i = \mu_0 \times (R_1\lambda_1 + R_2\lambda_2 + R_3\lambda_3)$, e_i is a normal random variable with mean 0 and variance 1, A_i and $D_i$are the additive and dominant genetic scores of the i-th individual. In our simulations, we set $R_1 = 1$ for alleles from Yoruba, $R_2 = 1/2$ for alleles from East Asia, $R_3 = 1/4$ for alleles from CEU, and $\mu_{00} = 2$. $(\lambda_1, \lambda_2, \lambda_3)$ are the admixture proportions of an individual from the three sub-populations. The same as in the discrete model with two sub-populations, we used CEU and Yoruba HapMap data for the admixed model simulation with two ancestral populations. Furthermore, we set $\mu_0 = 0$ and $\mu_0 = 0.2$ for type I error examination and power evaluation, respectively. A total of 10,000 randomly selected SNPs on chromosome 22 were used in the analyses for calculating principal components and kinship matrix. The test marker was randomly chosen on the chromosome but was not included in the above 10,000 random markers. Again, this way the LD structure across a chromosome is preserved for adjacent SNPs.

We have compared four statistical approaches under each of the four scenarios described above: linear model without considering population structure/admixture (denoted as "None" below), the approach using PCA to control for population structure/admixture (denoted as "P" as in literature), mixed-model using kinship matrix (denoted as "K" as in literature), and mixed-model with both kinship matrix and principal components (denoted as "K+P" following literature). The four statistical models can be expressed in the following form:

$$y = X\beta + P\alpha + Zu + e,$$

where **y** is a vector of observed phenotypes, **X** is a matrix of predictor variables, here is the genotypes of genetic markers. β is a vector of coefficients corresponding to fixed effects. **P** contains the top principal components, with corresponding effects contained in α. **Z** is an incidence matrix. **u** is the random effect of the mixed model with variance $Var(u) = \sigma_g^2 K$, where **K** is the matrix of kinship coefficients (a kinship coefficient is a measure of degree of genetic correlation between two individuals), σ_g^2 is the genetic variance that may capture the extent of similarity in phenotype for individuals who are similar in genotype. **e** is the random error vector with variance-covariance matrix $Var(e) = \sigma_e^2 I$, where σ_e^2 is the variance of a single random error and **I** is identity matrix. Therefore the phenotypic variance-covariance matrix is $Var(y) = \sigma_g^2 K + \sigma_e^2 I$. Model "None" corresponds to equation (1) without terms $P\alpha$ and Zu. Model "P" corresponds to equation (1) without term Zu. Model "K" corresponds to equation (1) without term $P\alpha$. Model "P+K" is the full model in equation (1). Mixed model with kinship matrix is usually used in analysis of pedigree data where the kinship coefficients can be estimated from the consanguineous relationship of the individuals in the pedigree. Here the kinship coefficients are inferred from the individuals' genotypes. Different methods have been proposed to estimate the kinship coefficients from genotypes and studies have shown that these marker-based kinship estimates are useful in genetic studies (Balding and Nichols, 1995; Kang et al., 2010; Lynch and Ritland, 1999; Ritland, 2005, 2009; Yu et al., 2006).

All analyses were performed using EMMA (Kang et al., 2008) which is an R package for association mapping correcting for population structure/admixture using mixed-model. SmartPCA (Patterson et al., 2006) was used to generate principal components. Following Zhu et al. (2008), we used only the first 10 principal components in our study.

3. RESULTS

To compare the performance of these four statistical approaches, we mainly focused on type I error rate and power. Table 1 presents the type I error rates and power values from the four statistical approaches under different simulated scenarios. Because the "None" model can not control type I error rate, we don't include the power values for it in table 1. In general, ignoring population structure/admixture may induce substantially inflated type I error rate, and some loss in power (i.e. mild inflated type II error rate). This is

Table 1. Type I error rates and power under different scenarios at significance levels 5% and 1%

Statistical model	Discrete genetic model		Admixed genetic model	
	Two ancestral populations	Three ancestral populations	Two ancestral populations	Three ancestral populations
No kinship or PCA	0.879	0.824	0.237	0.213
	0.840	0.764	0.117	0.092
Kinship only	0.067 (0.903)	0.049 (0.919)	0.039 (0.953)	0.053 (0.972)
	0.014 (0.813)	0.007 (0.816)	0.006 (0.873)	0.011 (0.913)
PCA only	0.061 (0.895)	0.046 (0.922)	0.046 (0.952)	0.052 (0.973)
	0.014 (0.820)	0.010 (0.833)	0.005 (0.893)	0.011 (0.909)
Both kinship and PCA	0.054 (0.803)	0.045 (0.842)	0.042 (0.885)	0.047 (0.906)
	0.005 (0.532)	0.003 (0.549)	0.000 (0.649)	0.001 (0.666)

Note: In each cell, the upper numbers are for $\alpha = 0.05$; the lower numbers for are $\alpha = 0.01$. The numbers are Type I error rates and power (in parentheses). Each entry is based on 1,000 data sets with sample size 1,000.

not surprising and is consistent with previous studies. The problem is more severe for data under discrete genetic models (i.e. population structure) than under admixed genetic models (i.e. population admixture). The performance of "K" and "P" models are comparable and exceed all other models under all scenarios considered. The "K+P" model has a slightly conserved type I error rate, and lower power. Figure 1 shows the QQ-plots for the four statistical models under null hypothesis with discrete model with two ancestral populations. It is clear that the "None" model has severe type I error rate. The other three statistical methods can control the type I error well. The figures are similar for other simulated scenarios (data not shown). Figure 2 shows the ROC curves of the four statistical methods under four simulated scenarios. Clearly, "K" and "P" perform consistently and significantly better. The "K+P" model performs better than the "None" model. The difference in performance between these two models is larger under discrete model than admixed model, and is the smallest under admixed model with three ancestral populations.

In addition, all statistical approaches seem to perform better under admixed genetic models (i.e. population admixture) than under discrete genetic models (i.e. population structure), with three ancestral populations than with two ancestral populations. This is expected.

4. DISCUSSION

Genetic association studies are very promising in disease fine mapping but can suffer from the potential confounding from population structure/admixture. With the availability of genome-wide genetic markers, this issue may be handled more efficiently. Many statistical methods have been proposed for this purpose. Among them, mixed-model approach seems to be very appealing. However, this approach is mainly used and evaluated in animal and plant genetics. Although it has been applied to human studies, its performance is still not comprehensively evaluated. In this study, we compared the performance of some statistical approaches, including the mixed-model approach, in controlling the confounding from population structure/admixture using human data with unrelated individuals.

In our simulation study, the performance of the "P" model and the "K" model is comparable and better than other models under all scenarios simulated. Not surprisingly, our study showed that ignoring population structure/admixture may induce substantially inflated type I error rate and mild type II error rate, in genetic association studies with samples of unrelated individuals. This is consistent with previous finding. Surprisingly, the approach with both random effects (i.e. with kinship matrix) and principal components does not perform as well as either the mixed-model approach (i.e. with kinship matrix) or the principal component approach. From previous studies, this "P+K" model performed better, at least as good as, the other two approaches (Stich and Melchinger, 2009; Stich et al., 2008; Zhao et al., 2007). The major difference is that we used human data consisting of unrelated samples, whereas the aforementioned previous studies used plant data where inbreds are usually included. The "P+K" approach can be thought of as a two stage model: The first stage is the "P" model. After fitting the "P" model, we can calculate the residual which is then used as response in the second stage. In the second stage, the "K" model is fitted with residual from the first stage as response. This may explain the potential problem of the "P+K" approach. In the first stage, the population structure/admixture is taken into account by the principal components. If the "P" model works well (this is supported by our study and previous studies), the potential confounding effect from population structure/admixture should be eliminated. This means that population structure/admixture should not confound the residual from the "P" model. In the second stage, however, the "K" model takes into account the population structure/admixture, which should no longer exist with the residual. We conjecture that it is this "double control" of population structure/admixture that affects the performance of

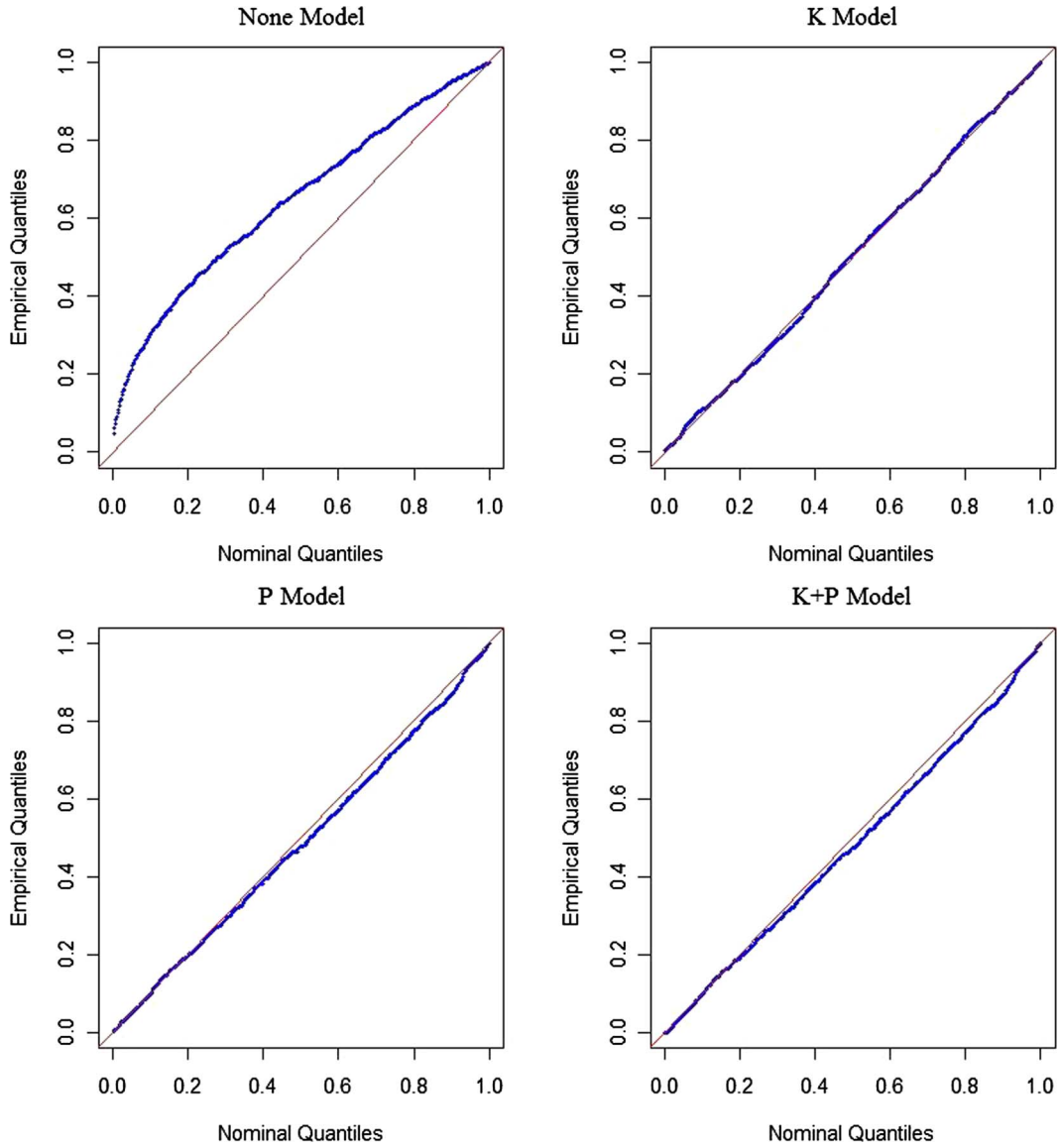

Figure 1. QQ plot (nominal p-values vs. observed p-values) under null hypothesis with discrete model with two ancestral populations using different statistical approaches. Each plot is based on 1,000 data sets with sample size 1,000.

the "P+K" approach. To test our conjecture, we conducted two simulations. Both simulations were performed under the alternative model with discrete and admixed models with two ancestral populations. In the first simulation, we used only the first two principal components in the "P" model and "P+K" model. In the second simulation, we conducted a two-stage analysis as described above: in the first stage, we used only the top 10 principal components as covariates to fit the model; in the second stage, we used residual from the first stage as response, and used two "K" models. In the first "K" model, the kinship matrix was estimated from the genotypes, the same as in the "P+K" model. In the second "K" model, an identity matrix was used as the kinship matrix. Therefore, this model is actually the "P" model with two-stage, but with the same number of parameters as the two-stage "P+K" model. In our first simulation, the power values are a little bit higher (less than 2% in absolute value) than that using ten principal components (data not shown), indicating that adding more principal components (i.e. with extra parameters in the model) may lower the power slightly, if any. But this amount of power decrease may not explain the power decrease of the "P+K" model. In our second simulation study, for the scenario of two discrete populations, the power values are 0.818 and 0.381 at $\alpha = 0.01$ and 0.859 and 0.717 at $\alpha = 0.05$, for approaches using identity kinship matrix and marker-based kinship matrix, respectively; for the scenario of two admixed populations, the power values are 0.837 and 0.588 at $\alpha = 0.01$ and 0.912

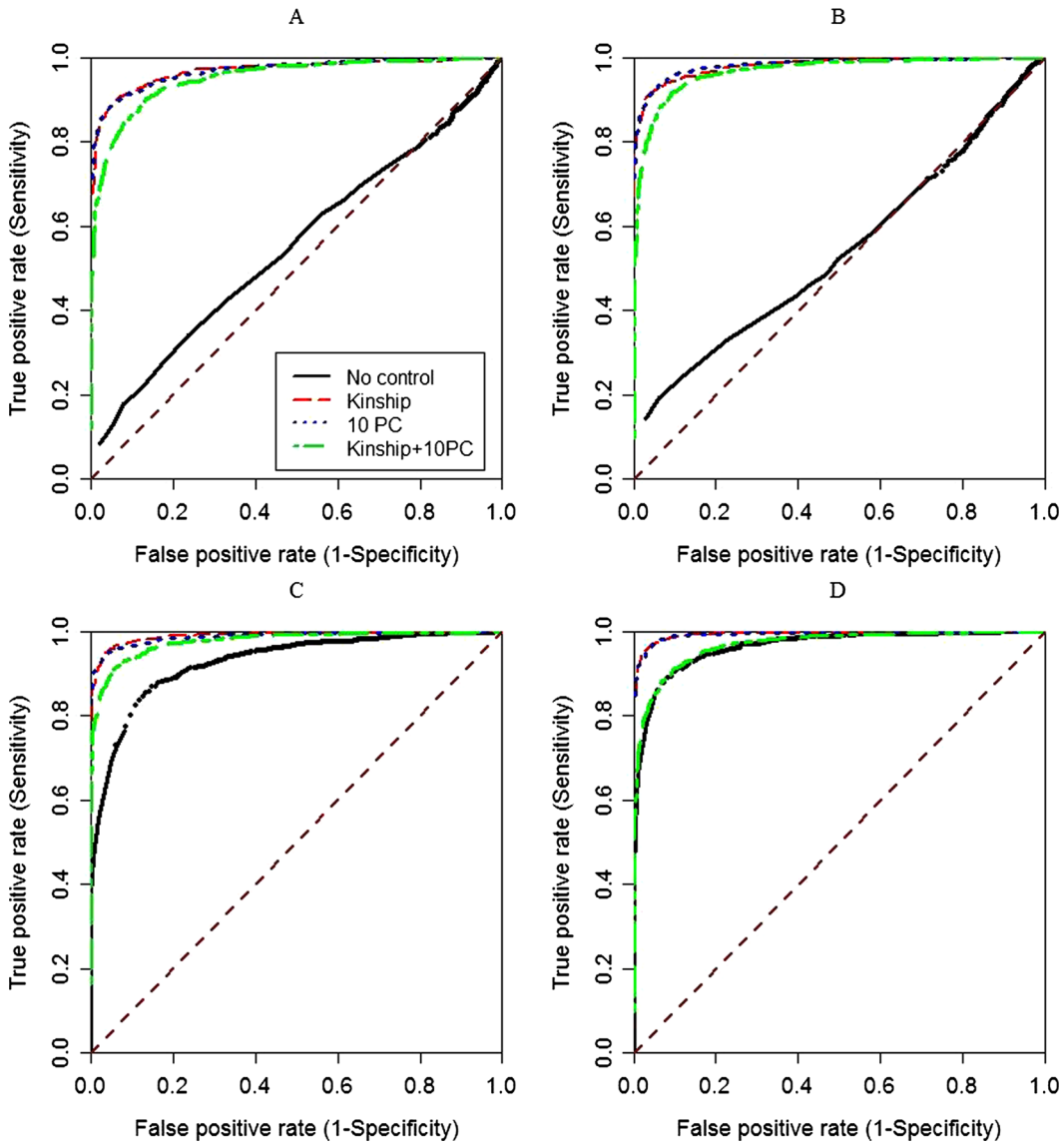

Figure 2. ROC plots comparing true and false-positive rates of four statistical models. Panels A to D are for simulated scenarios of discrete model with two ancestral populations, discrete model with three ancestral populations, admixed model with two ancestral populations, and admixed model with three ancestral populations, respectively. Each plot is based on 1,000 data sets with sample size 1,000.

and 0.838 at $\alpha = 0.05$, for approaches using identity kinship matrix and marker-based kinship matrix, respectively. Although the power values for the two-stage approaches are slightly lower than their one-stage counterparts, which is expected, the general pattern is the same. This confirms our conjecture that the lower power of "K+P" model is mainly due to the "double control", instead of the extra parameters that are estimated in the "K+P" model. It is also interesting that the "P+K" approach has reasonable type I error rate (may be a little bit conserved), which is consistent with the finding in Price et al. (2010), but inflated type II error rate (i.e. loss in power).

Price et al. (2010) reviewed the approaches to population stratification in genome-wide association studies. They also conducted simulations to evaluate the performance of these methods. There are some differences between that paper and our work. First, they considered population structure and family structure or cryptic relatedness together. That is, they mainly focused on pedigree data, although they had one simulation with unrelated individuals. Our study focused on unrelated individuals only. Second, they focused on type I error, as they said "Power to detect causal variants may vary between methods, but our focus here was on correcting false-positive associations." However, we evaluated

both type I error rate and power. In their simulation study, they had one scenario with unusually differentiated markers (allele frequency difference equal to 0.6). Our simulations were based on real data and may be more realistic. The allele frequency differences in our simulation range from 0 to 0.47. There is no marker in our study having such unusual allele frequency difference among the ancestry populations, therefore our findings do not apply to this unusual case.

Although models "P" and "K" performed similarly, the computation time differed greatly. Running EMMA on cluster with 3.0GHz quad-core Intel Xeon E5450 processors, it took about 1550s to calculate the likelihood ratio test (the "K" model) (it took about 1535s to compute kinship matrix, and about 15s to run the mixed-model). It took about 30s to run the "P" model (it took about 15s to calculate the principal components using smartPCA). It took about 10s to run the "None" model. Calculating the kinship matrix is extremely time-consuming.

The choice of number of principal components in the "P" (and "P+K") model is a practical concern. In our analysis, we used top 10 principal components to control for population structure/admixture. This number was suggested by Price et al. (2006) and used by Zhu et al. (2008). Because we had only two and three ancestral populations in our simulation studies, we think that this number should be sufficient. To further investigate the effect of number of principal components, we performed more simulation studies (data not shown). In addition to the use of top ten principal components in the "P" model, we have tried using one and two principal components in all our simulations. Our results showed that using an insufficient number of principal components may not fully control the confounding effect of population structure, resulting in both inflated false positives and false negatives. On the other hand, using more principal components may have slightly increased type I error rate and decreased power, but the effect is subtle. In our simulation, for the scenario of discrete model with three ancestry populations, two principal components may be enough. For all other scenarios, one principal component may be sufficient.

When using principal components to control for population structure, a practical approach is to include all the covariates, including the candidate marker and some top principal components in the model and see if these principal components are significant or not. If the principal components are significant, then keep them in the analysis; if not, exclude them from the analysis, i.e. do not control population structure. It has been discussed in the literature that this is not the correct way to handle confounding, as stated in (Kleinbaum et al., 1998) (Page 195) "One approach sometimes used (incorrectly) to assess confounding is, for example, to conduct a statistical test of $H_0 : \beta_2 = 0 \ldots$. Such a test does not address confounding, but rather precision." Redden et al. (2006) mentioned this as well. We will illustrate in a very simple scenario where there is only one candidate marker and no other covariates. We denote Y as the trait, S stands for population structure, and M stands for candidate marker. We assume that there is correlation between trait and population structure, between the candidate marker and population structure, but no correlation between trait and the candidate marker after controlling for population structure. That is, $r(Y,S) \neq 0$, $r(M,S) \neq 0$ and $r[Y,(M|S)] = 0$, where the first two are correlations and the last one is a semipartial correlation between the candidate marker and trait after controlling the candidate marker for population structure. We know that (Kleinbaum et al., 1998; Muller and Fetterman, 2002)

$$r[Y,(M|S)] = \frac{r(Y,M) - r(Y,S)\cdot r(M,S)}{\sqrt{1-r(M,S)^2}}$$
$$r[Y,(S|M)] = \frac{r(Y,S) - r(Y,M)\cdot r(M,S)}{\sqrt{1-r(M,S)^2}}.$$

Therefore we have $r(Y,M) = r(Y,S)\cdot r(M,S) \neq 0$ which means that the candidate marker is correlated with the trait if not controlling for population structure. When $|r(Y,S)| = |r(Y,M)|$, or equivalently $|r(M,S)| = 1$, we can have $r[Y,(S|M)] = 0$, which means that semipartial correlation between trait and population structure controlling for the candidate marker could be zero if the candidate marker and population structure are highly correlated. We know that testing covariates (variable added-last test) equal to zero is equivalent to testing semipartial correlations equal to zero (Muller and Fetterman, 2002). The above means that even if population structure is a confounder, the test of it could be zero with all the covariates in the linear model. Therefore if it is removed from the model (i.e., no control for population structure), there will be false positives. Because now the model contains only the candidate marker and we know that $r(Y,S) \neq 0$, which means that now the test of the candidate marker is not zero, even though we know that when controlling for population structure, candidate marker and trait are not correlated. Because of the relationship between semipartial correlation and full partial correlation, the above observation holds even if we control both trait and candidate marker for population structure, as in Price et al. (2006). In short, excluding principal components from the "P" model based on the test of them in the model may not be appropriate.

Another practical consideration in using principal components to control for population structure is how to control, on trait, on candidate marker, or on both. In order for population structure to be a confounder, it should be correlated with both the trait and the candidate marker. Therefore in terms of controlling for confounding, we can control either trait, or candidate marker, or both for population structure, as long as such controlling can eliminate the correlation with population structure. This can also be seen from the relationship between semipartial and full partial correlations, and the relationship between partial correlations and regression.

$$r[(Y|S),M] = \frac{r(Y,M) - r(Y,S)\cdot r(M,S)}{\sqrt{1-r(Y,S)^2}}$$
$$r[(Y,M)|S] = \frac{r(Y,M) - r(Y,S)\cdot r(M,S)}{\sqrt{(1-r(M,S)^2)(1-r(Y,S)^2)}},$$

with the same notation as above. From these equations we can see that the full partial and semipartial correlations have the same numerators, and we know that testing full partial and semipartial correlations equal to zero is equivalent to variable-added last test of regression coefficients to be zero (Muller and Fetterman, 2002), therefore controlling trait, candidate marker, or both for population structure should have the same type I error rate. The difference may be in power. We conducted simulations under both the null and alternative models with discrete and admixed models with two ancestral populations. We controlled both trait and candidate marker for population structure, as did in Price et al. (2006) (Price et al., 2006). However, the type I error rates and power values were the same as those from the "P" model as shown in table 1. Further studies may be needed to evaluate the benefits of this approach.

EMMA provides two tests: a likelihood ratio test and a t-test. Our simulation study showed that the likelihood ratio test performed as good as, or better (depends on the simulation scenarios) than, the t-test (data not shown). The results in this report are from the likelihood ratio test.

Recently we have data from a genome-wide association study on warfarin dose response. Warfarin is the most widely used oral anticoagulant and has become the case-study for pharmacogenetics. The evolution of our understanding of warfarin pharmacodynamics and pharmacokinetics and the recognition of genetic regulation of warfarin response has stimulated efforts aimed at quantifying this influence (Cooper et al., 2008; Klein et al., 2009; Limdi et al., 2008; Limdi et al., 2010; Rieder et al., 2005; Takeuchi et al., 2009). Our data consist of 290 unrelated African-American patients who were at least 20 years of age and were followed monthly for up to two years from initiation of therapy. Genotyping was performed using the Illumina 1M array with an overall 99.5% genotyping call rate and no gender discrepancies. After quality control, 991,457 SNPs on autosomes were left for subsequent analysis. Because our samples are all African-Americans, a well known admixed population, we may have the potential confounding effect from population admixture. Based on our results in this work, currently we are using the top two principal components to control for potential confounding effect from population admixture. We think that this should control the potential confounding but without incurring much computational burden. Actually, this dataset serves as our practical motivation for our current work.

In summary, we evaluated and compared the performance of some statistical methods for controlling the confounding effect from population structure/admixture in human genetic association studies with samples of unrelated individuals. Our results confirmed the conclusion that ignoring population structure/admixture in human genetic studies may incur severe false positive rates and mild false negative rates. We also showed that the mixed-model with principal components (i.e. the "K+P" model) did not perform very well, with inflated type II error rate (loss in power). In addition, we showed empirically that the performance of the mixed-model (the "K" model) and the principal component approach (the "P" model) were comparable. We also showed through simulation that inclusion of additional principal components may have little effect on type I error rate and power. We showed theoretically that removing principal components from "P" based on testing may not be appropriate. Based on our results, we suggest that it is always prudent to control for population structure if there is no strong evidence for its existence. It should be noted that our simulations are based on human data with unrelated individuals. Therefore these results may only apply to this kind of data. In addition, theoretical work is certainly warranted to access the relationship between the "P"model and "K" model. Further studies are needed for other data such as pedigree data.

ACKNOWLEDGEMENTS

We thank Xiaofeng Zhu and Qiuying Sha for sharing their programs for simulation with us. This work was supported in part by NIH grants GM077490 (DBA), GM081488 (NL), GM59507 (HZ), GM57672 (HZ), HL092173 (NAL), and K23NS45598 (NAL) from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the Associate Editor and the anonymous reviewer for their valuable comments and suggestions that improved the article.

Received 2 June 2010

REFERENCES

Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96** 3–12.

Bauchet, M., McEvoy, B., Pearson, L. N., Quillen, E. E., Sarkisian, T., Hovhannesyan, K., Deka, R., Bradley, D. G. and Shriver, M. D. (2007). Measuring European population stratification with microarray genotype data. *Am. J. Hum. Genet.* **80** 948–956.

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G. and Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nat. Genet.* **37** 868–872.

Chen, H. S., Zhu, X., Zhao, H. and Zhang, S. (2003). Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann. Hum. Genet.* **67** 250–264.

Cooper, G. M., Johnson, J. A., Langaee, T. Y., Feng, H., Stanaway, I. B., Schwarz, U. I., Ritchie, M. D., Stein, C. M., Roden, D. M., Smith, J. D., Veenstra, D. L., Rettie, A. E. and Rieder, M. J. (2008). A genome-wide scan for common genetic variants with

a large influence on warfarin maintenance dose. *Blood* **112** 1022–1027.

Deng, H. W. (2001). Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. *Genetics* **159** 1319–1323.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55** 997–1004.

Dolan, C., O'Halloran, A., Bradley, D. G., Croke, D. T., Evans, A., O'Brien, J. K., Dicker, P. and Shields, D. C. (2005). Genetic stratification of pathogen-response-related and other variants within a homogeneous Caucasian Irish population. *Eur. J. Hum. Genet.* **13** 798–806.

Ewens, W. J. and Spielman, R. S. (1995). The transmission disequilibrium test - history, subdivision and admixture. *Am. J. Hum. Genet.* **57** 455–464.

Falush, D., Stephens, M. and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164** 1567–1587.

Heiman, G. A., Hodge, S. E., Gorroochurn, P., Zhang, J. and Greenberg, D. A. (2004a). Effects of population stratification on false positive rates in association analysis: A simulation study. *Am. J. Epidemiol.* **159** S25–S25.

Heiman, G. A., Hodge, S. E., Gorroochurn, P., Zhang, J. Y. and Greenberg, D. A. (2004b). Effect of population stratification on case-control association studies - I. Elevation in false positive rates and comparison to confounding risk ratios (a simulation study). *Hum. Hered.* **58** 30–39.

Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J. and Stefansson, K. (2005). An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* **37** 90–95.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C. and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42** 348–354.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178** 1709–1723.

Klein, T. E., Altman, R. B., Eriksson, N., Gage, B. F., Kimmel, S. E., Lee, M. T., Limdi, N. A., Page, D., Roden, D. M., Wagner, M. J., Caldwell, M. D. and Johnson, J. A. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Medc.* **360** 753–764.

Kleinbaum, G. D., Kupper, L. L., Muller, E. K. and Nizam, A. (1998). *Applied Regression Analysis and Multivariable Methods*, 3rd ed. Duxbury Press.

Limdi, N. A., Beasley, T. M., Crowley, M. R., Goldstein, J. A., Rieder, M. J., Flockhart, D. A., Arnett, D. K., Acton, R. T and Liu, N. (2008). VKORC1 polymorphisms, haplotypes and haplotype groups on warfarin dose among African-Americans and European-Americans. *Pharmacogenomics* **9** 1445–1458.

Limdi, N. A., Wadelius, M., Cavallari, L., Eriksson, N., Crawford, D. C., Lee, M. T., Chen, C. H., Motsinger-Reif, A., Sagreiya, H., Liu, N., Wu, A. H., Gage, B. F., Jorgensen, A., Pirmohamed, M., Shin, J. G., Suarez-Kurtz, G., Kimmel, S. E., Johnson, J. A., Klein, T. E. and Wagner, M. J. (2010). Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups. *Blood* **115** 3827–3834.

Liu, N. and Zhao, H. (2006). A non-parametric approach to population structure inference using multilocus genotypes. *Hum. Genomics* **2** 353–364.

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152** 1753–1766.

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39** 906–913.

Muller, E. K. and Fetterman, A. B. (2002). *Regression and ANOVA: An Integrated Approach Using SAS Software*. SAS Publishing.

Paschou, P., Drineas, P., Lewis, J., Nievergelt, C. M., Nickerson, D. A., Smith, J. D., Ridker, P. M., Chasman, D.I., Krauss, R. M. and Ziv, E. (2008). Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet.* **4** e1000114.

Patterson, N., Price, A. L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* **2** e190.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

Price, A. L., Zaitlen, N. A., Reich, D. and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11** 459–463.

Pritchard, J. K. and Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65** 220–228.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* **67** 170–181.

Redden, D. T., Divers, J., Vaughan, L. K., Tiwari, H. K., Beasley, T. M., Fernandez, J. R., Kimberly, R. P., Feng, R., Padilla, M. A., Liu, N., Miller, M. B. and Allison, D. B. (2006). Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. *PLoS Genet.* **2**.

Rieder, M. J., Reiner, A. P., Gage, B. F., Nickerson, D. A., Eby, C. S., McLeod, H. L., Blough, D. K., Thummel, K. E., Veenstra, D. L. and Rettie, A. E. (2005). Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.* **352** 2285–2293.

Ritland, K. (2005). Multilocus estimation of pairwise relatedness with dominant markers. *Mol. Ecol.* **14** 3157–3165.

Ritland, K. (2009). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res.* **67** 175–185.

Satten, G. A, Flanders, W. D. and Yang, Q. H. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* **68** 466–477.

Spencer, C. C., Su, Z., Donnelly, P. and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5** e1000477.

Stich, B. and Melchinger, A. E. (2009). Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. *BMC Genomics* **10** 94.

Stich, B., Mohring, J., Piepho, H. P., Heckenberger, M., Buckler, E. S. and Melchinger, A. E. (2008). Comparison of mixed-model approaches for association mapping. *Genetics* **178** 1745–1754.

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N., Soranzo, N., Whittaker, P., Ranganath, V., Kumanduri, V., McLaren, W., Holm, L., Lindh, J., Rane, A., Wadelius, M. and Deloukas, P. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet.* **5** e1000433.

Tang, H., Peng, J., Wang, P. and Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genet. Epidemiol.* **28** 289–301.

Wu, B., Liu, N. and Zhao, H. (2006). PSMIX: an R package for population structure inference via maximum likelihood method. *BMC Bioinformatics* **7** 317.

Yu, J., Pressoir, G., Briggs, W. H., Vroh, Bi. I, Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38** 203–208.

Zhang, S. L., Zhu, X. F. and Zhao, H. Y. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet. Epidemiol.* **24** 44–56.

Zhang, Z., Ersoz, E., Lai, C. Q, Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J, Yu, J., Arnett, D. K., Ordovas, J. M. and Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42** 355–360.

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P. and Nordborg, M. (2007). An arabidopsis example of association mapping in structured samples. *PLoS Genet.* **3** e4.

Zhu, C. and Yu, J. (2009). Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* **182** 875–888.

Zhu, X., Li, S., Cooper, R. S. and Elston, R. C. (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. *Am. J. Hum. Genet.* **82** 352–365.

Zhu, X. F., Zhang, S. L., Zhao, H. Y. and Cooper, R. S. (2002). Association mapping, using a mixture model for complex traits. *Genet. Epidemiol.* **23** 181–196.

Zou, F., Lee, S., Knowles, M. R. and Wright, F. A. (2010). Quantification of population structure using correlated SNPs by shrinkage principal components. *Hum. Hered.* **70** 9–22.

Nianjun Liu
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, AL 35294
USA

Hongyu Zhao
Department of Epidemiology and Public Health
Department of Genetics
Yale University School of Medicine
New Haven, CT 06520
USA

Amit Patki
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, AL 35294
USA

Nita A. Limdi
Department of Neurology
University of Alabama at Birmingham
1719 6th Avenue South, CIRC-312
Birmingham, AL 35294
USA

David B. Allison
Section on Statistical Genetics
Department of Biostatistics
University of Alabama at Birmingham
USA
E-mail address: DAllison@ms.soph.uab.edu