Published in final edited form as:

Ann Hum Genet. 2011 May ; 75(3): 418–427. doi:10.1111/j.1469-1809.2010.00639.x.

A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies

Chengqing Wu, Andrew DeWan, Josephine Hoh, and Zuoheng Wang*

Department of Epidemiology and Public Health, Yale University, New Haven CT 06510, USA

Summary

Population stratification is an important issue in case–control studies of disease–marker association. Failure to properly account for population structure can lead to spurious association or reduced power. In this article, we compare the performance of six methods correcting for population stratification in case–control association studies. These methods include genomic control (GC), EIGENSTRAT, principal component-based logistic regression (PCA-L), LAPSTRUCT, ROADTRIPS, and EMMAX. We also include the uncorrected Armitage test for comparison. In the simulation studies, we consider a wide range of population structure models for unrelated samples, including admixture. Our simulation results suggest that PCA-L and LAPSTRUCT perform well over all the scenarios studied, whereas GC, ROADTRIPS, and EMMAX fail to correct for population structure at single nucleotide polymorphisms (SNPs) that show strong differentiation across ancestral populations. The Armitage test does not adjust for confounding due to stratification thus has inflated type I error. Among all correction methods, EMMAX has the greatest power, based on the population structure settings considered for samples with unrelated individuals. The three methods, EIGENSTRAT, PCA-L, and LAPSTRUCT, are comparable, and outperform both GC and ROADTRIPS in almost all situations.

Keywords

Population structure; association testing; type I error; power

Introduction

Genome-wide association studies (GWAS) have emerged as popular tools for identifying genetic variants that underlie human complex disease (McCarthy et al., 2008). In genetic case–control studies, typical approaches compare the frequency distribution of marker alleles or genotypes between cases and controls. A difference in the frequency of an allele or genotype of a polymorphism between the two groups suggests the possibility that the genetic marker may have a causal influence on the disease or trait, or may be in linkage disequilibrium (LD) with a polymorphism that does. One major concern in case–control studies is population stratification. When cases and controls are sampled from a population that contains two or more subpopulations, a false positive result can arise if both rates of disease and allele frequencies differ by subpopulation. Spurious association can also occur if samples arise from an admixed population and the ancestry distributions differ between cases and controls (Lander and Schork, 1994; Marchini et al., 2004).

To deal with unknown population structure in case-control association studies, methods using a set of ancestry-informative markers (AIMs) or genome-wide data have been developed to correct single-SNP association tests for population stratification (Devlin and Roeder, 1999; Pritchard et al., 2000; Satten et al., 2001; Zhu et al., 2002; Zhang et al., 2003; Chen et al., 2003; Hoggart et al., 2003; Price et al., 2006; Kimmel et al., 2007; Epstein et al., 2007; Luca et al., 2008; Rakovski and Stram, 2009; Zhang et al., 2009; Lee et al., 2010; Thornton and McPeck, 2010; Kang et al., 2010; Zhang et al., 2010). In principle, we can divide them into two different general approaches to correcting association tests. In the first, we can think of population membership as an unmeasured covariate, or more generally, proportion of ancestry from different populations as unmeasured covariates. Several studies propose efficient methods to reconstruct values of unmeasured covariates and use them to correct the test statistic (Pritchard et al., 2000; Satten et al., 2001; Zhu et al., 2002; Zhang et al., 2003; Chen et al., 2003; Price et al., 2006; Zhang et al., 2009; Lee et al., 2010). Among these methods, structured association (SA) (Pritchard et al., 2000), and principal component analysis (PCA) (Zhu et al., 2002; Zhang et al., 2003; Chen et al., 2003; Price et al., 2006) are most widely used. SA is a Bayesian Markov Chain Monte Carlo (MCMC) method that estimates population ancestry proportions for each sample. This information is then used in association testing. An advantage of the SA method is that it provides direct information about population structure for easy interpretation. However, it can lose power when too many subpopulations are specified. In addition, the current implementation is computationally intensive in large genome-wide datasets (Price et al., 2006). A simple and appealing alternative is PCA. PCA is a linear dimensionality reduction technique used to infer continuous axes of genetic variation. Price et al. (2006) developed the program EIGENSTRAT to correct for population structure in association tests. It uses the top eigenvectors of the sample covariance matrix as covariates in a regression setting. More recently, Zhang et al. (2009) and Lee et al. (2010) propose an alternative to PCA using spectral graph theory, a nonlinear dimensionality reduction technique, for inferring genetic ancestry. Their algorithms are implemented in LAPSTRUCT (Zhang et al., 2009) and Spectral-GEM (Lee et al., 2010), respectively. In the second category of methods, we can think of unobserved genealogy as creating dependence among individuals; thus, the regular test statistic would be inflated. One can directly estimate a correction factor for the variance such as genomic control (GC) (Devlin and Roeder, 1999), or estimate dependence structure and use it to correct the test statistic (Rakovski and Stram, 2009; Thornton and McPeck, 2010; Kang et al., 2010; Zhang et al., 2010). GC is a routinely used, simple approach to detecting stratification, where an inflation factor λ is estimated from the distribution of single-marker test statistics to measure the extent of inflation. It is also a computationally fast approach to correcting for stratification, where the same correction factor is applied to all SNPs. Methods that incorporate the estimated covariance structure across individuals have also been proposed, including ROADTRIPS (Thornton and McPeck, 2010), EMMAX (Kang et al., 2010), and TASSEL (Zhang et al., 2010). These three methods are applicable even when samples contain related individuals.

In this article, we compare six methods for association testing with population structure, including GC (Devlin and Roeder, 1999), EIGENSTRAT (Price et al., 2006), principal component-based logistic regression (PCA-L) (Zeggini et al., 2008; Need et al., 2009), LAPSTRUCT (Zhang et al., 2009), ROADTRIPS (Thornton and McPeck, 2010), and EMMAX (Kang et al., 2010). We first review these methods and discuss their differences in terms of modeling and computation. We then assess their performance based on simulations with various types of population structure, including admixture. Methods, such as EIGENSTRAT and LAPSTRUCT, were developed for unrelated samples, and they cannot be properly calibrated with family structure (Astle and Balding, 2009; Price et al., 2010). In the current study, we limit our comparisons to unrelated case-control samples. This

comparison allows us to explore the limitations and the relative power of these methods for association testing with population structure.

Methods

Genomic Control

GC is a nonparametric method that controls for population stratification in case-control studies (Devlin and Roeder, 1999). In the presence of population stratification, the association test statistic, for instance, the Armitage trend test (Armitage, 1955), may not follow a central χ^2 distribution under the null hypothesis of no association, and is generally inflated. The degree of inflation is dependent on the nature of the structure and is also proportional to sample size (Price et al., 2010). The key idea behind GC is that a single parameter λ , which measures the overall inflation in test statistic due to population stratification, can be estimated as the median of association test statistic across a set of unlinked markers divided by the theoretical median of a central χ^2_1 distribution. The GC approach simply applies the correction factor λ to the test statistic, and the resulting statistic is assumed to follow a χ^2 distribution with 1 degree of freedom. This method is computationally simple and fast. It can be used with a moderate number of markers. In addition, it can be used to correct for family structure and cryptic relatedness (Thornton and McPeck, 2010). However, the uniform inflation factor λ applied to all SNPs by GC could result in both an increase in type I error, due to undercorrection of highly differentiated noncausal SNPs, as well as a loss of power, due to overcorrection of causal SNPs (Thornton and McPeck, 2010; Price et al., 2010).

EIGENSTRAT

EIGENSTRAT is a widely used program in GWAS. It uses PCA to detect and correct for population stratification. PCA is a classical statistical tool to achieve dimension reduction through consideration of linear combinations of the original data. The top few principal components generally explain the greatest amount of variation in the data. PCA has long been used to study population structure in genetic data (Menozzi et al., 1978; Cavalli-Sforza et al., 1994; Reich et al., 2008; Novembre and Stephens, 2008). In GWAS, PCA has been used to explicitly model ancestry differences between cases and controls along continuous axes of variation. The algorithm is as follows. Suppose there are N individuals genotyped at M SNPs. For each individual i , let X_{ij} be the genotype data at SNP j , taking a value 0, 1, or 2, and let Y_i be the phenotype information. PCA considers an empirical covariance matrix Ψ with (i, j) th element being

$$\Psi_{ij} = \frac{1}{M} \sum_{s=1}^M \frac{(X_{is} - 2\hat{p}_s)(X_{js} - 2\hat{p}_s)}{2\hat{p}_s(1 - \hat{p}_s)}, \quad (1)$$

where $\hat{p}_s$ is an allele frequency estimate for the type 1 allele at marker s . EIGENSTRAT uses the top principal components obtained from Ψ as covariates in a multi-linear regression

$$Y_i = \alpha + \beta_1 v_{i1} + \beta_2 v_{i2} + \cdots + \beta_K v_{iK} + \gamma X_{is} + \varepsilon_i.$$

Here, v_{ik} is the ancestry of individual i along the k th axis of variation, which equals the i th coordinate of the k th eigenvector, and K is the number of axes of variation used to adjust for ancestry. The Armitage trend test is used to test $H_0: \gamma = 0$ vs. $H_A: \gamma \neq 0$. In the computation of EIGENSTRAT, the test statistic is defined as

$$EG = (N - K - 1) \text{Corr}^2(X_s^{\text{adj}}, Y^{\text{adj}}),$$

where X_s^{adj} is the adjusted genotype at marker s , defined as the residuals after regressing genotypes on the top K principal components. The adjusted phenotype Y^{adj} is similarly defined. The test statistic EG approximately follows a χ_1^2 distribution under the null hypothesis of no association. An important feature of EIGENSTRAT is that it provides information about population structure. The top principal components represent broad differences across individuals, thus effectively capture a few major axes of population structure. It is also computationally fast. It has been shown that the EIGENSTRAT method possesses power advantage over GC because the correction in EIGENSTRAT is specific to a candidate marker's variation in frequency across ancestral populations, which will minimize spurious associations as well as maximize power to detect true associations (Price et al., 2006).

Principal Component-Based Logistic Regression

An alternative way to incorporate PCA in case-control studies to correct for population stratification is by applying a generalized linear model. Recently, there have been several case-control association studies that combine PCA with logistic regression (Zeggini et al., 2008; Need et al., 2009). They use the top principal components as covariates in a logistic regression model (PCA-L). As the case-control phenotypes do not follow a normal distribution, applying a generalized linear model using logit or probit link function is preferable. However, the computational cost of a generalized linear model is much higher than that of a linear model such as EIGENSTRAT.

LAPSTRUCT

As an alternative to PCA, LAPSTRUCT is a software that uses spectral graph theory to infer population structure (Zhang et al., 2009). The spectral graph approach is more flexible than PCA and allows for different ways of modeling similarities and structure in the sample (Lee et al., 2010). In LAPSTRUCT, each subject is viewed as a vertex in a graph. An appropriate weight matrix is assigned to the graph to reflect the local dependence structure of the study subjects, with higher weight given when individuals are genetically closer. Then Laplacian eigenmaps are employed to obtain a new presentation of the original data. As opposed to PCA, this algorithm is nonlinear and robust to outliers, because it tends to emphasize substructure that affects many data points rather than just a few extreme points. The top eigenvectors of the associated graph Laplacian matrix contain important information on population structure. Comparing with PCA, the spectral graph approach tends to separate the data into more meaningful clusters, especially when outliers are present. In GWAS, LAPSTRUCT uses the top eigenvectors of the graph Laplacian as covariates in a logistic regression to correct for population stratification.

ROADTRIPS

ROADTRIPS is a recently proposed method that simultaneously corrects for both pedigree and population structure (Thornton and McPeck, 2010). In this method, the genotype data are treated as random and the phenotype information as fixed; thus, no assumption is made on the distribution of the phenotypes. The marginal distribution of genotypes is modeled as a fixed function of phenotypes, and the genotypic correlation structure is estimated from genome-wide data. An advantage of ROADTRIPS is that it can explicitly incorporate the pedigree information, if available, to improve power of the association test; while the four methods, we describe above cannot take advantage of this type of information.

ROADTRIPS uses the same sample covariance matrix Ψ (Equation (1)), estimated from genome-wide data, to correct the single-marker association test for unknown population structure and relatedness among the sampled individuals. Another feature of ROADTRIPS is that it takes into account the missing genotype pattern at each SNP to properly construct a valid test.

EMMAX

EMMAX is a variance component model to account for sample structure in GWAS. In this method, the effect of an allele is modeled as a main effect, while the population structure and relationships among samples are taken into account by means of variance components of random polygenic effects. It considers a model

$$Y_i = \alpha + \gamma X_{is} + \eta_i + \varepsilon_i, \quad \text{Var}(\eta) = \sigma_a^2 \Gamma,$$

where Γ is the normalized pairwise relatedness matrix, estimated by identity by state (IBS) or sample covariance matrix Ψ (Equation (1)). As oppose to EIGENSTRAT, which uses principal components as a set of fixed effects in a regression model, this method models population structure as random effects. Not only does the relatedness matrix Γ capture population structure, it also encodes a wider range of structures, including cryptic relatedness and family structure. Thus, EMMAX is applicable in related samples. In contrast to ROADTRIPS, this method can easily incorporate additional covariates, such as age and environmental factors, in the association analysis.

Simulation Study

We conduct simulation studies to compare the performance of the six methods described in the previous section, including GC, EIGENSTRAT, PCA-L, LARSTRUCT, ROADTRIPS, and EMMAX. We also include the standard uncorrected Armitage trend test for comparison. In EIGENSTRAT and PCAL, we use the default setting of 10 principal components. For LAPSTRUCT, the top 10 Laplacian eigenvectors are included in the analysis. To achieve a more comprehensive evaluation, we design two sets of simulations. In Simulation I, we consider population structure from discrete subpopulations. In Simulation II, we consider an admixed population.

Simulation I—We consider four different settings of population structure. The first two settings, S1 and S2, involve individuals sampled from two subpopulations. S3 and S4 involve samples from three and four subpopulations, respectively. In S1, there are 500 individuals sampled from subpopulation 1, and 500 individuals sampled from subpopulation 2. The number of cases from each subpopulation are specified by the fraction parameters a and b , and the remaining study subjects are controls. We vary the value of a (or b) from 0.1 to 0.5 with 0.1 increment, and set b (or a) at 0.5 in the simulations. For S2–S4, we generate a total of 500 cases and 500 controls, with varying proportions of individuals from each subpopulation. Following the design of Li and Yu (2008), in each setting, we choose the case–control sampling fractions at two levels, which correspond to moderate and extreme ancestral differences between cases and controls, respectively. Table 1 summarizes the fraction parameters and the number of cases and controls in each subpopulation for S1–S4.

In all simulations, we use genotype data at 100,000 random SNPs that are neither linked nor associated with the trait to correct for stratification. To generate genotypes for the 100,000 SNPs, we use the Balding-Nichols model (Balding and Nichols, 1995; Price et al., 2006). Hardy–Weinberg equilibrium (HWE) is assumed to hold at each SNP within subpopulations. For each SNP, the ancestral allele frequency p is drawn from a uniform(0.1,

0.9) distribution, and the allele frequency in each subpopulation is generated independently from the beta distribution with two parameters $p(1 - F_{ST})/F_{ST}$ and $(1 - p)(1 - F_{ST})/F_{ST}$, where F_{ST} is Wright's fixation index, a measure of genetic divergence among subpopulations (Wright, 1951). In our simulations, F_{ST} is set at 0.01, a typical value for the European population.

To evaluate the performance of different methods correcting for population stratification, we simulate three categories of testing SNPs: random SNPs with no association with the trait; differentiated SNPs with no association with the trait; and causal SNPs. Random SNPs were generated the same way as those SNPs chosen for detecting and correcting for population structure. For differentiated SNPs, we consider two sets of allele frequencies at each setting of population structure. In S1 and S2, the two sets of parameters of allele frequencies in subpopulations 1 and 2 are set to be (0.2, 0.8) and (0.1, 0.5). For S3, the allele frequencies in subpopulations 1, 2, and 3 are (0.2, 0.8, 0.8) and (0.1, 0.5, 0.5). For S4, we choose the allele frequencies to be (0.2, 0.8, 0.2, 0.8) and (0.1, 0.5, 0.1, 0.5) for subpopulations 1, 2, 3, and 4. In the simulation of causal SNPs, we assume a multiplicative trait model with a relative risk of 1.5 for the causal allele. We follow the algorithm in Price et al. (2006) to generate genotypes for causal SNPs conditioned on the disease status. Testing random and differentiated SNPs allows us to evaluate the type I error rates of different methods, whereas testing causal SNPs allows us to evaluate the power at different settings of population structure.

Simulation II—We sample cases and controls from an admixed population, formed by two ancestral populations. Individuals in the admixed population are assumed to have independent and identically distributed (i.i.d.) admixture vectors of the form $(a, 1 - a)$, where a is a Uniform (0, 1) random variable. The case-control status is simulated according to the model used in Price et al. (2006) with the ancestral risk r setting at two levels: $r = 2$ and $r = 3$.

To evaluate the type I error and power in Simulations I and II, for every population structure, we generate 10 datasets with each consisting of 100,000 random SNPs that are used to infer substructure and 100,000 testing SNPs in each of the three categories (random, differential, and causal).

Results

We set the significance level at 0.0001 to mimic the stringent cutoff for association at the genome-wide level. For each of the combinations of population structure and parameter setting, we test a total of 1,000,000 SNPs, across 10 datasets, for association, using various test statistics. The empirical type I error and power are calculated as the proportion of SNPs in which the test statistic exceed the χ^2_1 quantile corresponding to the nominal type I error of 0.0001. The statistics compared are the uncorrected Armitage trend test, GC, EIGENSTRAT, PCA-L, LAPSTRUCT, ROADTRIPS, and EMMAX. For type I error comparison, using an exact binomial calculation, we determine that empirical type I error rates falling in the range of 0.00008–0.00012 are not significantly different from the nominal 0.0001 level.

Simulation I Results

Tables 2–5 report the empirical type I error and power for each statistic under various discrete subpopulation structures. Under S1 (Table 2), when the ancestral distributions are the same in cases and controls, which corresponds to the case that $a = b = 0.5$, empirical type I error rates of all tests are not significantly different from the nominal level. This is to

be expected, because all tests have an approximate χ_1^2 distribution when there is no population stratification. However, when population structure exists, that is, $a \neq b$, type I error of the uncorrected Armitage trend test is inflated in all cases, because this test does not correct for population structure. In the case of testing for random SNPs, EIGENSTRAT, PCA-L, LAPSTRUCT, ROADTRIPS, and EMMAX are effective at controlling the type I error (Table 2, top). In contrast, GC tends to be conservative, especially when the ancestral difference is large between cases and controls. For differentiated SNPs, empirical type I error of PCA-L and LAPSTRUCT is in general not significantly different from the nominal 0.0001 level, whereas GC, ROADTRIPS, and EMMAX have inflated type I error (Table 2, centre). This is probably because GC, ROADTRIPS, and EMMAX correct all the Armitage test statistics across the genome either by a common factor or by a common correlation matrix that estimates an average effect due to population structure. This type of correction works well for random SNPs but is insufficient for SNPs that demonstrate strong differentiation across ancestral populations. Interestingly, EIGENSTRAT is conservative when there are more cases from subpopulation 1 and is anticonservative when there are more cases from subpopulation 2. At causal SNPs, when there is no population stratification ($a = b = 0.5$), power of all seven tests is extremely close. With population structure, EMMAX has the highest power among all methods that adjust for stratification (excluding the uncorrected Armitage test), and the EIGENSTRAT, PCA-L, and LAPSTRUCT methods have similar power, whereas GC and ROADTRIPS have relatively reduced power (Table 2, bottom). The power difference is even greater when the case-control fractions differ substantially between subpopulations. This suggests that the correction used in GC and ROADTRIPS is superfluous, leading to a loss in power.

The results for S2–S4 are presented in Tables 3–5, respectively. In all three settings, PCA-L and LAPSTRUCT can effectively adjust for the effect of population structure and their type I error rates in general are well controlled for both random and differentiated SNPs. However, GC, ROADTRIPS, and EMMAX fail to reduce the false positive rate to the nominal level when testing differentiated SNPs. EMMAX also appears to be anticonservative for random SNPs when there is extreme ancestral difference between cases and controls. This suggests that using a common correction factor or modeling population structure as a random effect could not provide a sufficient correction for stratification in data, especially for SNPs that are highly differentiated across ancestral populations. As before, the uncorrected Armitage test has inflated type I error in all settings. Although both EIGENSTRAT and PCA-L use principal components as covariates to correct for population structure, we find that PCA-L has better control of type I error than EIGENSTRAT. For moderately differentiated SNPs, for example, the ancestral allele frequencies are 0.1 and 0.5 in subpopulations 1 and 2, the empirical type I error of EIGENSTRAT is 0.00015 and 0.00038 under the moderate and more extreme ancestral difference between cases and controls, respectively, whereas the type I error of PCA-L is 0.00012 under both cases, not significantly different from the nominal 0.0001 level. For power comparisons, the three methods that model population structure as a fixed effect, EIGENSTRAT, PCA-L, and LAPSTRUCT, have similar power results, and outperform the GC and ROADTRIPS methods. Among the six correction methods compared, the power of EMMAX is the highest.

Simulation II Results

Table 6 displays the type I error and power of various statistics in an admixed population. Once again, the uncorrected Armitage test has inflated type I error, whereas the other methods, GC, EIGENSTRAT, PCA-L, LAPSTRUCT, ROADTRIPS, and EMMAX can adequately control the type I error when testing random SNPs. For differentiated SNPs, the pattern is similar to that in discrete populations. PCA-L and LAPSTRUCT yield the type I

error that is close to the nominal level, whereas GC, ROADTRIPS, and EMMAX have type I error that is significantly larger than the expected. The EIGENSTRAT method has a slightly inflated type I error when the ancestral risk $r = 3$. In terms of power, EMMAX is the greatest, and the power results are similar for EIGENSTRAT, PCA-L, and LAPSTRUCT. In contrast, the GC and ROADTRIPS methods have reduced power.

Discussion

Population stratification is a major concern in GWAS. Association analysis correcting for unknown population structure has drawn a great deal of attention in recent years. Several novel methods and programs have been developed in this context (Devlin and Roeder, 1999; Pritchard et al., 2000; Zhu et al., 2002; Price et al., 2006; Zhang et al., 2009; Lee et al., 2010; Thornton and McPeck, 2010; Kang et al., 2010; Zhang et al., 2010). However, the performance of these methods has not been carefully evaluated and compared. In this article, we examine six methods for correcting for population structure in unrelated case-control studies. Our simulations consider population structure that consists of two, three, or four discrete subpopulations. We also consider an admixed population. Our results suggest that the uncorrected Armitage test fails to control the false positive rate in all cases. PCA-L and LAPSTRUCT have comparable performance and they both outperform GC and ROADTRIPS in terms of type I error and power in all simulation settings. Interestingly, although EMMAX has inflated type I error when testing differentiated SNPs, it has the greatest power for causal SNPs, based on the population structure settings considered for samples with unrelated individuals. In our simulations, the program EIGENSTRAT can be either too liberal or too conservative compared to PCA-L, though both of them infer case-control ancestry through PCA.

When random SNPs are tested, we have shown that all six methods have type I error that is close to or lower than the nominal level of 0.0001. However, when unusually differentiated SNPs are tested, PCA-L and LAPSTRUCT are still valid, whereas GC, ROADTRIPS, and EMMAX are not properly calibrated for this setting. The GC method corrects for population structure by estimating a common correction factor that summarizes the average inflation of test statistic due to stratification. The uniform inflation factor applied by GC tends to underestimate the inflation at SNPs that show strong differentiation across ancestral populations. Similarly, ROADTRIPS and EMMAX estimate a common dependence structure from genome-wide data. When all SNPs have the same pattern of missing genotypes, the same correlation matrix used to correct for stratification at all SNPs is inadequate for highly differentiated SNPs. Although both EIGENSTRAT and PCA-L use PCA to estimate continuous axes of genetic variation, they differ in how they incorporate the inferred ancestry information in the downstream analysis. EIGENSTRAT uses the top principal components as covariates in a linear regression model, whereas PCA-L uses a logistic regression model. In the simulation studies, we find that the logistic regression model used in PCA-L gives better control of type I error. EIGENSTRAT can result in inflated type I error when testing differentiated SNPs. For LAPSTRUCT, the top eigenvectors of graph Laplacian are used in a logistic regression model, and it has similar performance to that of PCA-L. This suggests that modeling population structure as a fixed effect in a regression setting in general provides better control of type I error than methods that include structure as random effects. In terms of power, EMMAX performs best in all simulation settings. The three methods, EIGENSTRAT, PCA-L, and LAPSTRUCT, have similar power and are more powerful than ROADTRIPS and GC. This suggests that proper modeling of population structure as random effects can gain power compared to the fixed effect model. In addition, when family structure and cryptic relatedness are present, PCA may fail to infer genetic ancestry because of artificial principal components (Patterson et al., 2006; Thornton and McPeck, 2010; Price et al., 2010). In this case, both ROADTRIPS and

EMMAX have been shown to be practical approaches to simultaneously addressing confounding due to population structure, family structure, and cryptic relatedness (Thornton and McPeck, 2010; Price et al., 2010).

A recent article conducted a simulation study to compare the type I error of EIGENSTAT, EMMAX, ROADTRIPS, and the combined approach of PCA and EMMAX for random SNPs and differentiated SNPs with or without family structure (Price et al., 2010). In their simulation settings, all four methods can effectively correct for population stratification for random SNPs in unrelated case–control samples, while EMMAX and ROADTRIPS tend to have inflated type I error at highly differentiated SNPs. Their conclusions are similar to our simulation results, except that we also identified that EIGENSTRAT can result in inflated type I error when testing differentiated SNPs. For all simulated datasets, we apply the combined approach of PCA and EMMAX, that is, we include leading principle components as covariates in linear mixed models of EMMAX. The empirical type I error of the combined method is similar to that of EIGENSTRAT, and the power is slightly higher (results not shown).

Current GWAS typically have much larger sample sizes than those that we have simulated. We expect that our conclusions can be extended to larger samples. The type I error and power results are expected to have similar patterns with larger samples. The advantage of large sample size is that we would be able to maintain high power even with much smaller effects.

In summary, we have compared six association methods correcting for population stratification in case–control studies. Our results suggest that the PCA and graph spectral approaches are effective in inferring genetic ancestry. Incorporation of this information in a logistic regression model has better performance than using a linear regression model, especially when testing highly differentiated SNPs. Although GC and ROADTRIPS can be used to correct for family structure and cryptic relatedness (Thornton and McPeck, 2010), given the simulations performed here, both of them tend to lose power in correcting for stratification, as compared to the PCA and graph spectral methods. For case–control studies without missing genotypes, both methods correct the Armitage trend statistics across the genome by a common inflation factor, though this factor differs between the two methods. Our extensive simulations show that the correction factor used in GC tends to correct more when there is moderate or higher population stratification. As a result, the empirical type I error of ROADTRIPS is larger than that of the GC approach. However, ROADTRIPS has slightly higher power for moderate or higher population stratification. The overestimation of the inflation factor in GC might explain why GC is conservative for random SNPs, when population stratification is present (Astle and Balding, 2009). A variance component model, such as EMMAX, has the advantage of maintaining high power to detect association in the presence of population structure, as well as having potential to correct for family structure, however, cannot properly correct for stratification at unusually differentiated markers. Therefore, it would be of great interest to develop novel, more powerful association methods that simultaneously correct for both population stratification and pedigree structure.

Acknowledgments

The authors thank Mary Sara McPeck for discussion and helpful comments, and two anonymous reviewers for critical comments. This work was supported by grants provided by Verto Institute.

References

Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics*. 1955; 11:375–386.

- Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*. 2009; 24:451–471.
- Balding D, Nichols R. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*. 1995; 96:3–12. [PubMed: 7607457]
- Cavalli-Sforza, LL.; Menozzi, P.; Piazza, A. The history and geography of human genes. Princeton: Princeton University Press; 1994.
- Chen HS, Zhu X, Zhao H, Zhang S. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet*. 2003; 67:250–264. [PubMed: 12914577]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55:997–1004. [PubMed: 11315092]
- Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. *Genet Epidemiol*. 2007; 80:921–930.
- Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, Mckeigue PM. Control of confounding of genetic associations in stratified populations. *Am J Hum Genet*. 2003; 72:1492–1504. [PubMed: 12817591]
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010; 42:348–354. [PubMed: 20208533]
- Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. A randomization test for controlling population stratification in whole-genome association studies. *Am J Hum Genet*. 2007; 81:895–905. [PubMed: 17924333]
- Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*. 1994; 265:2037–2048. [PubMed: 8091226]
- Lee AB, Luca D, Klei L, Devlin B, Roeder K. Discovering genetic ancestry using spectral graph theory. *Genet Epidemiol*. 2010; 34:51–59. [PubMed: 19455578]
- Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol*. 2008; 32:215–226. [PubMed: 18161052]
- Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. On the use of general control samples for genome-wide association studies: Genetic matching highlights causal variants. *Am J Hum Genet*. 2008; 82:453–463. [PubMed: 18252225]
- Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*. 2004; 36:512–517. [PubMed: 15052271]
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat Rev Genet*. 2008; 9:356–369. [PubMed: 18398418]
- Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in europeans. *Science*. 1978; 201:786–792. [PubMed: 356262]
- Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, Shianna KV, Yoon W, Kasperavičiūtė D, Gennarelli M, Strittmatter WJ, Bonvicini C, Rossi G, Jayathilake K, Cola PA, Mcevoy JP, Keefe RSE, Fisher EMC, St. Jean PL, Giegling I, Hartmann AM, Möller H-J, Ruppert A, Fraser G, Crombie C, Middleton LT, St. Clair D, Roses AD, Muglia P, Francks C, Rujescu D, Meltzer HY, Goldstein DB. A genome-wide investigation of snps and cnvs in schizophrenia. *PLoS Genet*. 2009; 5 e1000373.
- Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*. 2008; 40:646–649. [PubMed: 18425127]
- Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006; 2:e190. [PubMed: 17194218]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38:904–909. [PubMed: 16862161]

- Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet.* 2010; 11:459–463. [PubMed: 20548291]
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet.* 2000; 67:170–181. [PubMed: 10827107]
- Rakovski CS, Stram DO. A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors. *PLoS ONE.* 2009; 4:e5825. [PubMed: 19503792]
- Reich D, Price AL, Patterson N. Principal component analysis of genetic data. *Nat Genet.* 2008; 40:491–492. [PubMed: 18443580]
- Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet.* 2001; 68:466–477. [PubMed: 11170894]
- Thornton T, McPeck MS. Roadtrips: Case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet.* 2010; 86:172–184. [PubMed: 20137780]
- Wright S. The genetical structure of populations. *Ann Eugenics.* 1951; 15:323–354.
- Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, De Bakker PIW, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding C-J, Doney ASF, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CNA, Payne F, Perry JRB, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet.* 2008; 40:638–645. [PubMed: 18372903]
- Zhang J, Niyogi P, McPeck MS. Laplacian eigenfunctions learn population structure. *PLoS ONE.* 2009; 4:e7928. [PubMed: 19956572]
- Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol.* 2003; 24:44–56. [PubMed: 12508255]
- Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet.* 2010; 42:355–360. [PubMed: 20208535]
- Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. *Genet Epidemiol.* 2002; 23:181–196. [PubMed: 12214310]

Table 1

Population structure settings in simulation I.

Population structure setting*	Group	Population 1	Population 2	Population 3	Population 4
S1 (<i>a, b</i>)	Case	500 <i>a</i>	500 <i>b</i>		
	Control	500(1 - <i>a</i>)	500(1 - <i>b</i>)		
S2 (<i>p</i> ₁ , <i>p</i> ₀)	Case	500 <i>p</i> ₁	500(1 - <i>p</i> ₁)		
	Control	500 <i>p</i> ₀	500(1 - <i>p</i> ₀)		
S3 (<i>p</i> ₁₁ , <i>p</i> ₁₂ , <i>p</i> ₀₁ , <i>p</i> ₀₂)	Case	500 <i>p</i> ₁₁	500 <i>p</i> ₁₂	500(1 - <i>p</i> ₁₁ - <i>p</i> ₁₂)	
	Control	500 <i>p</i> ₀₁	500 <i>p</i> ₀₂	500(1 - <i>p</i> ₀₁ - <i>p</i> ₀₂)	
S4 (<i>p</i> ₁₁ , <i>p</i> ₁₂ , <i>p</i> ₁₃ , <i>p</i> ₀₁ , <i>p</i> ₀₂ , <i>p</i> ₀₃)	Case	500 <i>p</i> ₁₁	500 <i>p</i> ₁₂	500 <i>p</i> ₁₃	500(1 - <i>p</i> ₁₁ - <i>p</i> ₁₂ - <i>p</i> ₁₃)
	Control	500 <i>p</i> ₀₁	500 <i>p</i> ₀₂	500 <i>p</i> ₀₃	500(1 - <i>p</i> ₀₁ - <i>p</i> ₀₂ - <i>p</i> ₀₃)

Population structure S1 has two subpopulations with each having 500 samples. Population structure S2–S4 each has 500 cases and 500 controls. In S1, the parameters *a* and *b* take values in Table 2. S2 contains two subpopulations, where *p*₁ and *p*₀ denote the proportions of cases and controls from subpopulation 1, respectively. We consider two levels of population stratification: the moderate ancestral difference is defined as (*p*₁, *p*₀) = (0.4, 0.6), and the extreme ancestral difference is defined as (*p*₁, *p*₀) = (0.5, 0). S3 contains three subpopulations. Moderate and extreme ancestral differences are defined as (*p*₁₁, *p*₁₂, *p*₀₁, *p*₀₂) = (0.45, 0.35, 0.35, 0.20) and (0.33, 0.67, 0.0, 0.33), respectively. S4 contains four subpopulations. Parameters (*p*₁₁, *p*₁₂, *p*₁₃, *p*₀₁, *p*₀₂, *p*₀₃) take values (0.35, 0.3, 0.2, 0.15, 0.2, 0.3) and (0.25, 0.25, 0.5, 0.0, 0.5, 0.25), which correspond to moderate and extreme ancestral differences between cases and controls, respectively.

Table 2

Empirical type I error rates and power under population structure S1. Type I error rates that are significantly different from the nominal 0.0001 level are in bold.

<i>a</i>	<i>b</i>	Armitage	GC	EIGENSTART	PCA-L	LAPSTRUCT	ROADTRIPS	EMMAX
Random SNPs								
0.1	0.5	0.02028	0.00004	0.00009	0.00009	0.00009	0.00010	0.00012
0.2	0.5	0.00486	0.00005	0.00009	0.00009	0.00010	0.00010	0.00011
0.3	0.5	0.00082	0.00006	0.00008	0.00008	0.00008	0.00009	0.00009
0.4	0.5	0.00017	0.00008	0.00010	0.00010	0.00008	0.00009	0.00011
0.5	0.5	0.00008	0.00008	0.00008	0.00008	0.00008	0.00008	0.00010
0.5	0.4	0.00018	0.00008	0.00009	0.00009	0.00009	0.00008	0.00010
0.5	0.3	0.00086	0.00008	0.00009	0.00009	0.00008	0.00010	0.00009
0.5	0.2	0.00501	0.00006	0.00009	0.00009	0.00008	0.00010	0.00009
0.5	0.1	0.02036	0.00004	0.00009	0.00010	0.00010	0.00010	0.00010
Differentiated SNPs ($f_1 = 0.1, f_2 = 0.5$)								
0.1	0.5	1.00000	0.91045	0.00040	0.00009	0.00009	0.99201	0.00606
0.2	0.5	0.97915	0.57125	0.00021	0.00009	0.00008	0.87109	0.00143
0.3	0.5	0.38814	0.12184	0.00012	0.00009	0.00010	0.37867	0.00052
0.4	0.5	0.00579	0.00305	0.00011	0.00010	0.00008	0.02320	0.00019
0.5	0.5	0.00010	0.00010	0.00009	0.00009	0.00008	0.00009	0.00009
0.5	0.4	0.00518	0.00266	0.00007	0.00007	0.00008	0.02202	0.00017
0.5	0.3	0.38344	0.11312	0.00006	0.00009	0.00008	0.37250	0.00030
0.5	0.2	0.98706	0.57455	0.00003	0.00008	0.00007	0.89112	0.00044
0.5	0.1	1.00000	0.94689	0.00001	0.00009	0.00009	0.99762	0.00083
Causal SNPs								
0.1	0.5	0.43299	0.03755	0.27510	0.26923	0.26720	0.04281	0.29725
0.2	0.5	0.45957	0.11657	0.38194	0.37861	0.37689	0.11862	0.39774
0.3	0.5	0.48700	0.26874	0.45750	0.45520	0.45375	0.26762	0.46886
0.4	0.5	0.50601	0.44644	0.49994	0.49799	0.49700	0.43940	0.50941
0.5	0.5	0.51422	0.51357	0.51417	0.51236	0.51233	0.51361	0.51697
0.5	0.4	0.50613	0.44239	0.49949	0.49748	0.49730	0.43852	0.50931
0.5	0.3	0.48559	0.26998	0.45589	0.45363	0.45301	0.26730	0.46943

<i>a</i>	<i>b</i>	Armitage	GC	EIGENSTART	PCA-L	LAPSTRUCT	ROADTRIPS	EMMAX
0.5	0.2	0.45947	0.11431	0.38069	0.37724	0.37673	0.11720	0.39735
0.5	0.1	0.42993	0.03664	0.27199	0.26593	0.26610	0.04149	0.29731

Empirical type I error rates and power under population structure S2. Type I error rates that are significantly different from the nominal 0.0001 level are in bold.

Table 3

	Armitage	GC	EIGENSTRAT	PCA-L	LAPSTRUCT	ROADTRIPS	EMMAX
			S2 (0.4, 0.6)				
Random SNPs	0.00087	0.00006	0.00011	0.00009	0.00009	0.00008	0.00011
Differentiated SNPs	$(f_1, f_2) = (0.2, 0.8)$	0.85123	0.49698	0.00011	0.00011	0.96531	0.00093
	$(f_1, f_2) = (0.1, 0.5)$	0.35325	0.10738	0.00015	0.00012	0.35501	0.00034
Causal SNPs	0.51107	0.29663	0.48485	0.48435	0.48433	0.29826	0.49984
			S2 (0.5, 0)				
Random SNPs	0.03684	0.00005	0.00012	0.00010	0.00010	0.00012	0.00014
Differentiated SNPs	$(f_1, f_2) = (0.2, 0.8)$	1.00000	0.00010	0.00008	0.00008	1.00000	0.13213
	$(f_1, f_2) = (0.1, 0.5)$	1.00000	0.93297	0.00038	0.00012	0.98832	0.01903
Causal SNPs	0.50954	0.03351	0.26539	0.25832	0.25812	0.03982	0.31546

Table 4

Empirical type I error rates and power under population structure S3. Type I error rates that are significantly different from the nominal 0.0001 level are in bold.

	Armitage	GC	EIGENSTRAT	PCA-L	LAPSTRUCT	ROADTRIPS	EMMAX
			S3 (0.45, 0.35, 0.35, 0.20)				
Random SNPs	0.00107	0.00010	0.00009	0.00011	0.00010	0.00010	0.00012
Differentiated SNPs	$(f_1, f_2, f_3) = (0.2, 0.8, 0.8)$ 0.01256	0.00034	0.00007	0.00008	0.00007	0.04707	0.00016
	$(f_1, f_2, f_3) = (0.1, 0.5, 0.5)$ 0.00557	0.00020	0.00006	0.00010	0.00007	0.00294	0.00008
Causal SNPs	0.16474	0.06590	0.04381	0.04261	0.04244	0.07138	0.06076
			S3 (0.33, 0.67, 0.0, 0.33)				
Random SNPs	0.06147	0.00006	0.00011	0.00010	0.00008	0.00011	0.00016
Differentiated SNPs	$(f_1, f_2, f_3) = (0.2, 0.8, 0.8)$ 1.00000	0.75792	0.00009	0.00007	0.00008	0.98969	0.42581
	$(f_1, f_2, f_3) = (0.1, 0.5, 0.5)$ 0.97670	0.00288	0.00021	0.00007	0.00006	0.01847	0.03698
Causal SNPs	0.50379	0.01812	0.11261	0.10402	0.10376	0.02428	0.19193

Table 5

Empirical type I error rates and power under population structure S4. Type I error rates that are significantly different from the nominal 0.0001 level are in bold.

	Armitage	GC	EIGENSTRAT	PCA-L	LAPSTRUCT	ROADTRIPS	EMMAX
	S4 (0.35, 0.3, 0.2, 0.15, 0.2, 0.3)						
Random SNPs	0.00123	0.00007	0.00008	0.00008	0.00007	0.00009	0.00009
Differentiated SNPs	$(f_1, f_2, f_3, f_4) = (0.2, 0.8, 0.2, 0.8)$	0.01046	0.00022	0.00008	0.00008	0.03304	0.00017
	$(f_1, f_2, f_3, f_4) = (0.1, 0.5, 0.1, 0.5)$	0.00523	0.00014	0.00007	0.00007	0.00260	0.00012
Causal SNPs	0.1216	0.02831	0.03641	0.03610	0.03460	0.02924	0.04503
	S4 (0.25, 0.25, 0.5, 0.0, 0.5, 0.25)						
Random SNPs	0.00855	0.00008	0.00011	0.00010	0.00008	0.00009	0.00013
Differentiated SNPs	$(f_1, f_2, f_3, f_4) = (0.2, 0.8, 0.2, 0.8)$	1.00000	1.00000	0.00010	0.00012	1.00000	0.70702
	$(f_1, f_2, f_3, f_4) = (0.1, 0.5, 0.1, 0.5)$	1.00000	0.99996	0.00010	0.00010	1.00000	0.16459
Causal SNPs	0.04989	0.00038	0.02131	0.02099	0.02070	0.00045	0.02796

Empirical type I error rates and power under an admixed population. Type I error rates that are significantly different from the nominal 0.0001 level are in bold.

Table 6

	Armitage	GC	EIGENSTRAT	PCA-L	LAPSTRUCT	ROADTRIPS	EMMAX
Ancestry risk $r = 2$							
Random SNPs	0.00023	0.00008	0.00010	0.00010	0.00008	0.00009	0.00010
Differentiated SNPs ($f_1 = 0.1, f_2 = 0.5$)	0.03132	0.01582	0.00011	0.00010	0.00010	0.02589	0.00058
Causal SNPs	0.51822	0.44147	0.49038	0.48853	0.48761	0.43895	0.50869
Ancestry risk $r = 3$							
Random SNPs	0.00062	0.00008	0.00010	0.00011	0.00010	0.00008	0.00011
Differentiated SNPs ($f_1 = 0.1, f_2 = 0.5$)	0.27209	0.11400	0.00013	0.00011	0.00011	0.15964	0.00176
Causal SNPs	0.51717	0.34430	0.45171	0.44943	0.44854	0.33997	0.48056