Genetic Epidemiology 23:181–196 (2002)

Association Mapping, Using a Mixture Model for Complex Traits

Xiaofeng Zhu,[1]* ShuangLin Zhang,[2,3] Hongyu Zhao,[4] and Richard S. Cooper[1]

[1]Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, Illinois
[2]Department of Mathematical Science, Michigan Technological University, Houghton, Michigan
[3]Department of Mathematics, Heilongjiang University, Harbin, China
[4]Department of Epidemiology and Public Heath, Yale University School of Medicine, New Haven, Connecticut

Association mapping for complex diseases using unrelated individuals can be more powerful than family-based analysis in many settings. In addition, this approach has major practical advantages, including greater efficiency in sample recruitment. Association mapping may lead to false-positive findings, however, if population stratification is not properly considered. In this paper, we propose a method that makes it possible to infer the number of subpopulations by a mixture model, using a set of independent genetic markers and then testing the association between a genetic marker and a trait. The proposed method can be effectively applied in the analysis of both qualitative and quantitative traits. Extensive simulations demonstrate that the method is valid in the presence of a population structure. Genet. Epidemiol. 23:181–196, 2002. © 2002 Wiley-Liss, Inc.

Key words: population structure; case-control design; principal component.

INTRODUCTION

Association studies have attracted considerable attention in the genetic dissection of complex diseases because they are potentially more powerful than

Grant sponsor: NIH; Grant numbers: HL53353, HL65702, GM59507; Grant sponsor: Donald W. Reynolds Cardiovascular Clinical Research Center.
*Correspondence to: Xiaofeng Zhu, Ph.D., Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, IL 60153. E-mail: xzhu1@lumc.edu
Received for publication 5 February 2002; Revision accepted 8 April 2002
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/gepi.0210

family-based linkage studies [Risch and Merikangas, 1996; Risch, 2000]. The approach often used in association studies for qualitative traits is the case-control design, in which allele frequencies are compared between unrelated individuals that are affected (cases) and unaffected (controls). However, a fundamental problem confronting this design is the selection of controls, as a false-positive association may appear if they are chosen with bias. For example, population stratification across different subgroups can lead to "spurious" associations as a result of confounding [Lander and Schork, 1994]. The most likely source of confounding is ethnicity, in which both disease prevalence and allele frequencies are different by ethnicity, and cases and controls are not well-matched in terms of ethnic origin [Risch, 2000]. The same problem is also inherent in association mapping of quantitative trait loci (QTLs) [Zhu and Elston, 2001].

To overcome the difficulty caused by population structure, Spielman et al. [1993] proposed the transmission-disequilibrium test (TDT), which compares the frequencies of genetic marker alleles that are transmitted from heterozygous parents to affected children against those that are not transmitted. In this design, the ethnic background of cases and controls is necessarily matched. The TDT method has been extended to include a variety of genetic models for qualitative traits [Schaid, 1996; Horvath and Laird, 1998; Spielman and Ewens, 1998; Boehnke and Langefeld, 1998; Zhao et al., 2000]. A number of quantitative trait TDT methods have also been developed to control for the effect of population stratification in mapping QTLs [Allison, 1997; Allison et al., 1999; Rabinowitz, 1997; Rabinowitz and Laird, 2000; Xiong et al., 1998; George et al., 1999; Zhu and Elston, 2001; Zhu et al., 2001]. However, compared with the case-control design, TDT-based methods have substantial practical disadvantages. TDT methods require the collection of DNA samples from family members, which is more difficult than from unrelated controls, especially in the case of late-onset diseases. Furthermore, only part of the sample is actually used in TDT approaches. Compared with the case-control design, the TDT is a less efficient method.

An alternative strategy to eliminate the effect of population stratification is to use independent genetic markers typed in cases and controls. Devlin and Roeder [1999] proposed a "genomic control" (GC) method relying on this technique for association mapping to control population stratification when genome screen data are available. The underlying assumption is that under the null hypothesis, allele frequency differences between cases and controls come from the same distribution at all loci. This assumption may not be valid in reality, however, because of selection; the false-positive rate may therefore be inflated [Pritchard et al., 2000a]. Pritchard et al. [2000a] proposed an approach called "structured association" (SA), which contrasts with the GC method [Pritchard and Donnelly, 2001]. SA uses a set of independent genetic markers to estimate the number of subpopulations based on a Markov chain Monte Carlo (MCMC) method and the ancestry probabilities of individuals from putative "unstructured" subpopulations [Pritchard et al., 2000b]. This information is then used to test for association. When the number of subpopulations is large, the simulation-based test statistic becomes very computationally intensive, especially for genome-wide association analysis. Other methods have also been published [Satten et al., 2001; Zhang et al., 2001].

In this paper, we propose an approach based on a mixture model to infer the number of subpopulations, using a set of independent genetic markers, and then test for the association between a marker and a trait. This approach is valid in the presence of a population structure and can be applied equally to the analysis of qualitative and quantitative traits, although we focus here on qualitative traits.

METHODS

Consider an association study with N unrelated individuals sampled from a population in which structure exists (for case-control design, $N=N_d+N_h$, where N_d is the number of diseased individuals and N_h is the number of unaffected individuals). Let y_i represent the phenotypic value for the i^{th} individual, with $y_i=1$ if affected and 0 if unaffected, and g_i the genotypic value for a candidate gene which we wish to test for association with the trait. We assume there are M independent markers that have been genotyped. Define x_{im} as the genotype value of the m^{th} marker for the i^{th} individual. In this paper, we focus only on biallelic markers, and assume the two alleles to be *1* and *2*. The genotypic value can be defined as 0, 1, and 2 for marker genotypes *11*, *12*, and *22*, respectively. Define $X=(X_1, X_2,\ldots, X_N)'$, where $X_i=(x_{i1}, x_{i2},\ldots,x_{iM})'$, $i=1, 2,\ldots,N$, and the prime represents a transpose of a vector or matrix. Now considering X_i as a random variable, we have an $M \times M$ sample covariance matrix $\Sigma=\mathrm{cov}(X)$. Let e_j be the j^{th} eigenvector corresponding to the j^{th} largest eigenvalue of Σ. (All the eigenvalues are positive because Σ is a nonnegative definite matrix.) We can then calculate the j^{th} principle component as $t_{ij} = (X_i - \bar{X})'e_j$, where $\bar{X}$ is the average of X_i, $i=1, 2,\ldots,N$. Let $T_i = (t_{i1}\ldots t_{iM})$ be the M principal component values for the i^{th} individual. Because the principal components are the linear combinations of the independent random variables, it is reasonable to assume that the t_{ij} follow approximately a mixture of K normal distributions when the sample size is large and the genetic markers originate from K subpopulations (see Apprendix). Furthermore, conditional on the k^{th} subpopulation, we can assume the distribution of T_i is the product of M normal distributions:

$$f(T_i|k) = \prod_{j=1}^{M} N(t_{ij}|\mu_{kj}, \sigma^2_{kj}),$$

where

$$N(t_{ij}|\mu_{kj}, \sigma^2_{kj}) = \frac{1}{\sqrt{2\pi\sigma^2_{kj}}} \exp\left[\frac{(t_{ij}-\mu_{kj})^2}{2\sigma^2_{kj}}\right]$$

is a normal density function. The distribution of T_i is

$$f(T_i) = \sum_{k=1}^{K} \lambda_k \prod_{j=1}^{M} N(t_{ij}|\mu_{kj}, \sigma^2_{kj}),$$

where λ_k is the probability that an individual originates from the k^{th} subpopulation, with the restriction $\Sigma\lambda_k=1$.

Given an individual is from the k^{th} subpopulation, we use logistic regression to model the association between a candidate gene and a trait:

$$\log\left[\frac{\Pr[y_i=1|g_i, k]}{\Pr[y_i=0|g_i, k]}\right] = \mu + \beta g_i + \delta_k \quad (1)$$

where δ_k indicates the effect of the k^{th} subpopulation subject to the restriction that $\delta_k = 0$, and β represents the effect of the candidate gene. It is assumed that the effect of the candidate gene is the same across subpopulations, but this is not a necessary assumption because we can use a population-specific β. This model is also used by Satten et al. [2001]. For the sake of simplicity, we do not include any covariates in equation (1), but this is not a necessary restriction. Under equation (1), the probability that an individual has phenotype y_i, given genotypic value g_i and the principal component T_i, is

$$\Pr[y_i|g_i,\ T_i] = \sum_{k=1}^{K} \Pr[y_i|g_i,\ T_i,\ k] \Pr[k|g_i,\ T_i].$$

We further assume that the probability that an individual belongs to the k^{th} subpopulation is only dependent on the principal components, which are obtained from the marker data. Thus we have

$$\Pr[y_i|g_i,\ T_i] = \sum_{k=1}^{K} \Pr[y_i|g_i,\ k] f(T_i|k)/f(T_i).$$

The likelihood of the observed data under equation (1) is then

$$L_{LOGISTIC} = \prod_{i=1}^{N} \sum_{k=1}^{K} \lambda_k \Pr[y_i,\ k] f(T_i|k)/f(T_i). \tag{2}$$

When K is known, we can maximize the likelihood function (2) to estimate the parameters. To test the null hypothesis $\beta = 0$, we use the statistic $Z_\beta = \hat{\beta}/SE(\hat{\beta})$, where $\hat{\beta}$ and $SE(\hat{\beta})$ are the estimates of β and its standard error by maximizing the likelihood function (2). According to asymptotic theory, the statistic Z_β has approximately a normal distribution.

Due to the numerous parameters, directly maximizing the likelihood function (2) is difficult. Therefore, we propose to estimate the number of subpopulations K first. Biernacki and Govaert [1999a] and Biernacki et al. [1999b] discussed several criteria for choosing the number of component K in a mixture model, including the Akaike information criterion (AIC) [Akaike, 1974] and the Bayesian information criterion (BIC). Both AIC and BIC choose K to maximize a penalized likelihood function, where the penalty of BIC is much greater than that of AIC when the number of parameters increases in the model. In other words, BIC favors a model with fewer parameters. Based on extensive simulations, Biernacki et al. [1999b] concluded that the BIC criterion behaves better than the AIC criterion. Therefore, we use the BIC criterion to choose K. Because K is assumed to be dependent on the marker data only, we use the likelihood function of observing the principal components $T_i, \ldots T_N$, i.e.,

$$L_{POP} = \prod_{i=1}^{N} \sum_{k=1}^{K} \lambda_k f(T_i|k), \tag{3}$$

to infer the number of subpopulations K. The BIC is defined as

$$BIC(K) = -2 \log(L_{POP}(K)) + \log(N) p(K),$$

where $L_{POP}(K)$ is the maximized likelihood for a given K and $p(K)$ is the number of parameters in the mixture model. In practice, we can calculate the BIC values for different K and select the K with the minimum BIC value. We can begin with $K=1$ until the BIC value starts to increase.

To estimate the BIC value given K, we maximize the likelihood function (3) by using the E-M algorithm for a mixture of normal distributions, as described by Celeux and Govaert [1995]. We select the initial values of $(\mu_{kj}, \sigma^2_{kj})$ as follow. Given j, we first normalize the value of t_{ij} for the i^{th} individual to the interval [0, 1]. The initial values are selected as

$$\lambda_k = \frac{1}{K},\ \mu_{kj} = \frac{k}{K},\ \sigma^2_{kj} = \frac{1}{K},\ k = 1, 2, \ldots, K,\ j = 1, 2, \ldots, M.$$

E-step

1. Estimate the probability that the i^{th} individual originated from k^{th} mixture component given T_i by

$$\lambda_{ik} = \frac{\lambda_k f(T_i|k)}{\sum_{k=1}^{K} \lambda_k f(T_i|k)},\ i = 1, \ldots, N, k = 1, \ldots, K.$$

2. Define indicator variables I_{ik}, $i = 1, \ldots, N, k = 1, \ldots, K$, as

$$I_{ik} = \begin{cases} 1; & \text{if } \lambda_{ik} = \max\{\lambda_{i1}, \ldots, \lambda_{iK}\} \\ 0; & \text{otherwise} \end{cases}$$

3. The estimate λ_k is given by $\lambda_k = 1/N \sum_{i=1}^{N} I_{ik}$, which is the proportion of the samples that originate from k^{th} subpopulation.

M-step

The M-step determines new estimates of the parameters $(\mu_{kj}, \sigma^2_{kj})$ by

$$\mu_{kj} = \frac{\sum_{i=1}^{N} t_{ij} I_{ik}}{\sum_{i=1}^{N} I_{ik}} \text{ and } \sigma^2_{kj} = \frac{\sum_{i=1}^{N} (t_{ij} - \mu_{kj})^2 I_{ik}}{\sum_{i=1}^{N} I_{ik}}.$$

We repeat both the E-step and M-step until the absolute difference of the current and previous log-likelihood values is less than a predefined value. Since the E-M algorithm above may be sensitive to the selection of initial values, we then restart the E-M algorithm by permuting the initial values of $(\mu_{1j}, \ldots, \mu_{Kj})$ for a given j.

After estimating K, we can estimate μ, β, δ_k by maximizing likelihood (2), using the E-M algorithm above with a little modification:

1. In the E-step, we calculate the probability that the i^{th} individual originated from the k^{th} mixture component by

$$\lambda_{ik} = \frac{\lambda k \Pr[y_i|g_i, k] f(T_i|k)}{\sum_{k=1}^{K} \lambda_k \Pr[y_i|g_i, K] f(T_i|k)};$$

2. In the M-step, we further maximize likelihood (2) to obtain μ β, and δ_k after obtaining μ_{kj} and σ^2_{kj}.

The total number of parameters to maximize likelihood function (3) is $K(2M+1)+1$ when all the M principal components are used. To ensure the global maximum is achieved, we restart the E-M algorithm with different sets of initial values. However, the computational time is a problem. According to our experience, we suggest using 3 or 4 principal components to infer the number of subpopulations K. After estimating K, we use $K-1$ principal components in likelihood function (2) to estimate μ β and δ_k The results obtained from maximizing likelihood function (3) can be used for the initial values when maximizing likelihood function (2), and additional sets of initial values are then not necessary.

In our simulation studies later, we compared our methods to the results from STRUCTURE/STRAT software [Pritchard et al. 2000a,b], which can be downloaded via website (http://pritch.bsd.uchicago.edu).

SIMULATION STUDIES

To validate the proposed method, we conducted simulation studies under a variety of population structure models. We simulated data first by coalescent techniques [Hudson, 1990], and then by known allele frequencies in four subpopulations extracted from ALFRED [Cheung et al., 2000].

Model A: Data Simulated by a Coalescent Process

By using the method of Zhang et al. [2001], we simulated the data only under the null hypothesis of no association between a genetic marker and a disease trait. We assumed that the sizes of the two subpopulations were 100 and 10,000 individuals when they divided, and that the current sizes are 10^7 and 5×10^7, respectively. Two population divergence times between the two subpopulations were considered: 1) 500 generations, and 2) 1,500 generations, with the latter corresponding to the likely divergence time between African and European populations [Reich et al., 2001]. The population structure was created by drawing 50 cases and 30 controls from the first population, and 50 cases and 70 controls from the second population. We simulated 100 independent markers that are not associated with the disease phenotype. A mutation rate of 5×10^{-7} per marker per generation was assumed, and we only selected markers with minor allele frequencies in the sample greater than 0.2.

Model B: Test Sample Generated From Empirical Population Data

An allele frequency database modeled on human population samples (“ALFRED”) was recently designed by Cheung et al. [2000], and it is available to the public via website (http//info.med.yale.edu/genetics/kkidd). ALFRED provides both SNPs and microsatellite markers. From ALFRED, we extracted 147 markers across four populations (Danes, San Francisco Chinese, Biaka, and Maya), representing different continental populations. For the microsatellite markers, we pooled the alleles to form a biallelic marker, with allele frequencies between 10–90%. All markers were assumed to be independent. We then simulated marker genotypes for each individual conditional on the allele frequencies in the individual’s original

population. Since all frequencies were obtained from real data, the genotypes should better represent the marker distributions of the population to which he/she belongs. An admixed population was created from these four subpopulations.

Model B1: discrete admixture of two subpopulations

We sampled 75 individuals as controls from each of the two subpopulations, but 100 individuals of the cases were drawn from subpopulation 1 and 50 from subpopulation 2. Thus, we actually assigned the risk of disease in subpopulation 1 to be two times higher than in subpopulation 2 [Pritchard and Rosenberg, 1999]. One hundred biallelic markers were simulated independently, based on the allele frequencies.

Model B2: discrete admixture of four subpopulations

We sampled 38, 38, 37, and 37 individuals as controls from each of the four subpopulations, but 60, 45, 30, and 15 cases from subpopulations 1, 2, 3, and 4, respectively. We designated the risk of disease as 4:3:2:1 for populations 1, 2, 3, and 4, respectively [Pritchard and Rosenberg, 1999]. One hundred biallelic markers were independently simulated.

Model B3: continuous admixture of two subpopulations

We simulated a data set using model C in Pritchard et al. [2000a]. Briefly, the data include 400 cases and 400 controls with 200 biallelic markers. Among the 200 markers, the first 53 marker frequencies were used twice. Under the assumption that there are two underlying subpopulations, we generated two datasets where the disease prevalence in one subpopulation is 8- and 4-fold higher than in the other subpopulation, respectively. For each control, we first simulated λ from a beta distribution with parameters (1, 4). This λ represents the fraction of the ancestry from the first subpopulation. Next we simulated this individual's two alleles at each marker with probability λ from the first subpopulation, and probability $1 - \lambda$ from the second population. To simulate cases, a rejection sampling method was used. For each case a probability λ was drawn from the beta distribution and was accepted with probability $(1 + 7\lambda)/8$, or $(1 + 3\lambda)/4$, respectively. If λ was rejected, we drew a new λ from the beta distribution. This procedure was repeated until a selected K was accepted.

Simulation of Candidate Loci

For data set A, we used the first marker as the candidate marker. For data set B, we used the same method as in Pritchard et al. [2000b] to generate biallelic candidate loci for each individual. Briefly, we assumed that the candidate locus has two alleles A and a and allele A has frequency p_k in the k^{th} subpopulation. Under the null hypothesis of no association between the candidate locus and the phenotype, a candidate genotype was randomly assigned to each of the cases and controls, conditional on the subpopulation from which the individual originates, and the frequencies of allele A and a in that subpopulation. Under the alternative

hypothesis, the probability that a case has genotype g at the candidate locus in the k^{th} population is

$$\Pr[g|affected,\ k] = \frac{R_g \Pr[g|k]}{\sum_g R_g \Pr[g|k]}, \quad (4)$$

where summation is over (AA, Aa, and aa), and R_g is a relative risk associated with genotype g that is population-specific. For the multiplicative model, the relative risk associated with a genotype was assumed to be the product of the relative risk factors of the individual's two alleles. For example, the relative risk for an individual with genotype AA, Aa, and aa is R_AR_A, R_AR_a, and R_aR_a, respectively. For the dominant model, we assumed $R_{AA} = R_{Aa}$. Then, the affected individual's genotype at the candidate locus was simulated based on the probability in equation (4). The method of simulating the candidate genotypes can be interpreted to mean that the candidate locus is either the functional site itself or is a marker in linkage disequilibrium with a functional site [Pritchard et al., 2000b].

RESULTS

Validity

We simulated a large number of datasets under both model A (the coalescent model) and model B (the admixture of 2 or 4 discrete real populations, or continuous admixture of two populations). To reduce the computation time, we assumed K was known in our estimation and was the same as in the simulations. Tables I and II summarize type I error rates for statistic Z_β under the null hypothesis of no association between the candidate marker and the phenotype. For comparison, we also calculated type I error rates using the logistic regression model without adjustment for population stratification and the STRAT by Pritchard et al. [2000a].

In the case of the coalescent models in which two subpopulations are involved (Table I), the results show that Z_β is associated with a somewhat conservative type I error (at P values of 0.05, 0.01, and 0.001), while logistic regression without adjustment for stratification greatly inflates the type I error. Our estimated P values suggest that our procedure performs well, even when 100 independent biallelic markers are used to control population stratification.

TABLE I. Type 1 Error Rates for Both Z_β and Logistic Regression Model Without Adjusting for Population Stratification Under Coalescent Models[a]

No. of generations	P-value	Z_β	Logistic regression
500	0.05	0.0497	0.139
	0.01	0.0093	0.041
	0.001	0.0007	0.005
1,500	0.05	0.0485	0.275
	0.01	0.0085	0.103
	0.001	0.0009	0.014

[a]Results are based on 10,000 replications. For each replication, the sample consists of 100 cases and 100 controls. A total of 100 independent markers was used to make inferences about the population structure.

TABLE II. Type I Error Rates for Z_β and a Logistic Regression Model Without Adjusting for Population Stratification, and STRAT Using Empirical Population Genetic Data[a]

Model B1				Model B2				Model B3			
	Z_β	Logistic regression	STRAT[b]		Z_β	Logistic regression	STRAT[b]		Z_β	Logistic regression	STRAT
					$P=0.05$						
p_1, p_2[c]				p_1, p_2, p_3, p_4[d]				R[e]			
0.5, 0.1	0.049	0.351	0.06	0.9, 0.7, 0.30, 0.2	0.051	0.811	0.059	4	0.058	0.140	0.097
0.9, 0.1	0.045	0.790	0.064	0.1, 0.3, 0.70, 0.9	0.047	0.728	0.065	8	0.068	0.255	0.159
					$P=0.01$						
0.5, 0.1	0.01	0.118	0.01	0.9, 0.7, 0.3, 0.1	0.0085	0.521	0.012	4	0.013	0.043	0.029
0.9, 0.1	0.008	0.984	0.01	0.1, 0.3, 0.7, 0.9	0.0085	0.38	0.014	8	0.015	0.102	0.058
					$P=0.001$						
0.5, 0.1	0.001	0.0214		0.9, 0.7, 0.3, 0.1	0.005	0.178		4	0.0013	0.0076	0.0038
0.9, 0.1	0.005	0.0559		0.1, 0.3, 0.7, 0.9	0.0006	0.106		8	0.0020	0.0228	0.018

[a]Results are based on 10,000 replications. For each replication, the sample consists of 150 cases and 150 controls. A total of 100 independent markers was used to infer the population structure for model B1 and B2, but 200 markers for B3.
[b]Indicates the results based on 1,000 replications.
[c]p_1, p_2 represent the testing marker allele frequencies in subpopulations 1 and 2, respectively.
[d]p_1, p_2, p_1, p_2 represent the testing marker allele frequencies in subpopulations 1, 2, 3, and 4, respectively.
[e]R represents the relative disease prevalence for two subpopulations.

We calculated the type I error rates using the empirical population genetic data. Table II summarizes the results for the mixture of 2 and 4 subpopulations (models B1 and B2). When the admixed population consists of 2 or 4 subpopulations, the type I error rate was still on the conservative side, but the unadjusted logistic regression had a large type I error. STRAT also gives a reasonable type I error rate.

In the case of model B3, in which a continuous mixture model of two subpopulations is assumed, the type I error of Z_β is still close to the nominal level. In contrast, both STRAT and logistic regression model have a higher type I error rate (Table II).

To view how well our method can infer the hidden population structure, we examined the number of subpopulations estimated by BIC. For each simulated data set, we calculated the BIC values from $K=1$–6. The K corresponding to the smallest BIC value was the estimated number of subpopulations. Table III presents the results based on 5,000 replications for the mixture model. In the cases of admixed populations simulated by coalescent models (number of generations equal to 500 and 1,500, $K=2$) and by four empirical population distributions (model B2), the mixture model predicted the number of subpopulations more than 99.7% correctly. Furthermore, every individual could be assigned to the right subpopulation membership with more than 99% probability when the number of subpopulation was correctly inferred. When the number of markers to control population stratification was dropped to 50, only less than 1% individuals were assigned to the wrong subpopulations.

The performance of the mixture model was further examined in the case of continuous admixture, such as occurs with model B3, after increasing the number of

TABLE III. Frequencies of Number of Subpopulations Estimated by Mixture Models[a]

Frequencies of the no., of subpopulations inferred					
	Simulation model				STRUCTURE[b]
No. of subpopulations to be inferred	Coalescent 500 generations, $K=2$	Coalescent 1,500 generations, $K=2$	B2, $K=4$	B3, $K=2$	B3, $K=2$
1	0	0	0	250	0
2	5,000	4,999	0	4,699	46
3	0	1	9	50	56
4	0	0	4,987	1	89
5	0	0	4	0	106
6	0	0	0	0	123

[a]K is the number of subpopulations used in simulations. One hundred markers were used to infer population structure for the coalescent model and B2, but 200 markers for B3. For the coalescent model and B2, we sampled 150 cases and 150 controls, but 400 cases and 400 controls for model B3. All results were based on 5,000 replications, except that by STRUCTURE.
[b]Result was based on 500 replications.

independent markers controlling for population structure to 200. In this case, it is much more difficult to infer the population structure because each individual is admixed. Under these circumstances, we found that the mixture model could still correctly predict the number of subpopulations 94% of the time (Table III). The correlation between the estimate of the amount of ancestry and the actual ancestry is about 0.78. For comparison, we used the program STRUCTURE [Pritchard et al., 2000b] to estimate the number of subpopulations, based on 500 replications to reduce computation time. The results show that the number of subpopulations was overestimated by STRUCTURE (last column in Table III), and only in 53 out of 500 simulation runs were the correct numbers of subpopulations predicted.

Power

We examined the power of the statistics Z_β under a variety of alternative models and compared it with STRAT. We assumed that the number of subpopulations was known. In all power calculations, we assigned the genetic value g to be 1 if the allele associated with a trait appeared in the candidate marker, and otherwise, to be 0. Table IV summarizes the power of these statistics under different models. For the multiplicative model, STRAT is usually more powerful than Z_β. The difference in the power estimates between STRAT and Z_β becomes larger when the relative risk for the same allele varies in different subpopulations. However, Z_β is more powerful than STRAT for a dominant model. The reason that Z_β lacks power in the multiplicative setting is that we assumed that the genetic effects are fixed across subpopulations, and our coding method for the genetic value is only more powerful for the dominant model. We can improve the power of Z_β, for example, by coding different genetic values for homozygote and heterozygote genotypes [Zhu and Elston, 2001].

TABLE IV. Power Comparisons of Statistics Z_β and STRAT, Using Empirical Population Genetic Data From ALFRED[a]

	Multiplication $R_{A1} = R_{A2} = 1.5$		Multiplication $R_{A1} = 1.0$, $R_{A2} = 2$		Dominant, $R_{A1A1} = R_{A1a1} = R_{A2A2} = R_{A2a2} = 2.0$		Dominant, $R_{A1A1} = R_{A1a1} = 1.5$, $R_{A2A2} = R_{A2a2} = 2.0$	
	Z_β	STRAT	Z_β	STRAT	Z_β	STRAT	Z_β	STRAT
p_1, p_2 [b]					$P = 0.05$			
0.1, 0.5	0.40	0.43	0.22	0.62	0.68	0.46	0.42	0.25
0.1, 0.9	0.21	0.27	0.04	0.19	0.48	0.38	0.18	0.13
					$P = 0.01$			
0.1, 0.5	0.17	0.22	0.07	0.38	0.41	0.25	0.20	0.10
0.1, 0.9	0.08	0.10	0.006	0.07	0.25	0.18	0.07	0.04

[a]Results are based on 2,000 replications for Z_β and 1,000 replications for STRAT. A total of 147 biallelic markers was used to control the effect of population structure. R_{A1} and R_{A2} are relative risk factors, where the subscript represents the allele that is inherited from an ancestor in population 1 or 2. In all models, we set $R_{a1} = R_{a2} = 1$. For STRAT, we assume that the number of subpopulations is known, and equal to 2, but we estimate it for Z_β. A, a represent the two alleles of a candidate marker; 1, 2 represent subpopulations 1 and 2.

[b]p_1, p_2 represent the testing marker allele frequencies in subpopulation 1 and 2, respectively.

DISCUSSION

We have presented a method for association mapping with a structured population sample. In our method we first infer K, i.e., the number of subpopulations based on the marker information, and then we conduct the association test conditional on the number of subpopulations. Extensive simulations show that our method is robust under different population structures, although more SNPs are required to adequately control for the effect of population structure in the case of continuous admixture. Our simulation studies also showed that K could be estimated correctly by the mixture model, using the BIC information. In our method, the marker information is summarized by the principal components. Although part of the marker information may be lost because we use only the first $K - 1$ principal components, this effect will be minimal compared to adding a large number of additional parameters into the model. In our simulation studies of equation 3, we added the second and the third principal components, but the correlation between estimated ancestry and actual ancestry is almost the same. When we used all the principal components, we found that the maximum of the likelihood function often changes with different initial values, indicating that many local maximums of the likelihood function exist. Hence the global maximum likelihood is much more difficult to achieve when there are too many parameters. In a similar method, using the latent-class model proposed by Satten et al. [2001], all the marker information is used, and the inference of population structure and the estimation of the genetic effect are performed simultaneously. However, the number of parameters in their model is substantially larger, and maximizing the likelihood function may be problematic. This may explain their simulation results, where the number of subpopulations was underestimated for the admixture of distinct subpopulations

[Tables 2 and 3 in Satten et al., 2001], and dependent on sample size for the continuous admixture [Table 5 in Satten et al., 2001]. Underestimating the number of subpopulations would invalidate the test statistics for the null hypothesis, inflating the type I error. Overestimating the number of subpopulations may not inflate the type I error, but it will reduce power. Satten et al. [2001] did not perform simulation studies regarding type I error rates, although they mentioned several possible test statistics, such as the likelihood ratio test or permutation tests; their validity should be studied further before they are used to test the null hypothesis. As noted by Satten et al. [2001], as well as by Pritchard and Donnelly [2001], the simulation study by Satten et al. [2001] was too small to permit a conclusion about coverage properties. In contrast, we provide a method for both parameter estimation and hypothesis testing.

In our simulations, we used 100 independent markers to control for the effect of population stratification for discrete admixture, and 200 markers for continuous admixture. In general, it is difficult to predict the number of markers required to adequately control for the effect of population structure, which is dependent on degree of differentiation of the subpopulations, the number of subpopulations, and the extent of admixture [Pritchard et al., 2000a]. When the number of markers controlling for the population stratification was dropped to 50, the mixture model still predicted an individual's subpopulation membership correctly for the discrete admixture population. However, many more markers are necessary for a continuously admixed population. In practice, we can test association between a trait and each marker that is used for controlling for the effect of population structure. A positive rate above the nominal level indicates that additional markers are required to control for the effect of population structure [Zhang et al., unpublished findings].

We suggest using the first $K - 1$ principal components to avoid too many parameters in the model. After the first $K-1$ principal components, little information can be gained. For example, we estimated the amount of ancestry for each individual for a simulated dataset according to model B3. The correlations between the estimated ancestry and the actual ancestry are 0.782, 0.788, and 0.787, corresponding to 1, 2, and 3 principal components used in the model, respectively. In practice, K is usually unknown and should be estimated. We suggest using at least 3 or 4 principal components to infer the number of subpopulations.

In our type I error studies, we assumed that K was known and was the same as what we used in the simulation data. This may have some limitations. If we used the estimated K, the type I error rate might have been higher than what we observed in our simulations.

A similar method was proposed by Pritchard et al. [2000a]. This method first infers K and the probabilities that each individual belongs to a subpopulation using the MCMC method, and then uses this information for association mapping. In our simulation study, the method proposed by Pritchard et al. [2000a] tends to overestimate the number of subpopulations. While this may not inflate type I error rate, power could be diminished and the computational time could increase dramatically. We also noted that the type I error of STRAT in our simulations is higher than in Pritchard et al. [2000a]. Perhaps 200 biallelic markers are not enough to control well for the effect of population structure in our simulations.

Devlin and Roeder [1999] introduced a GC method to correct for the effect due to population structure. Reich and Goldstein [2001] showed that the GC method is robust to population stratification by simulations. However, the GC method assumes that the allele frequency differences at all loci under the null hypothesis come from the same distribution. Such an assumption could be violated when selection is involved, and therefore lead to high type I error rates [Pritchard et al., 2000a]. The power of the GC method can be decreased if the genetic effect is different across different subpopulations [Pritchard et al., 2001]. Our proposed method can still have power, providing the population-specific parameter β is used.

When the target population is a mixture of two subpopulations, we found that adding the first principal component as a covariate to a simple logistic regression can adjust for the effect of population structure (data not shown). This property could be useful in association mapping in a population with a demographic history similar to that of African Americans, based on admixture of two continental populations. To assess the statistical properties, the routine SAS package can be used.

Although our method is based on biallelic markers, it can also be extended to the case of microsatellite markers, using the same idea as proposed by Zhang et al. [unpublished findings]. Specifically, suppose there are L_m alleles for m^{th} marker, and that the alleles are indicated by 1, 2,. . .,L_m. Let $z_{im}^{(1)}$ and $z_{im}^{(2)}$ be the two alleles of the m^{th} marker genotype for the i^{th} individual. We then define a $2L_m$ dimensional vector $X_{im} = (x_1^{(1)}, \ldots, x_{Lm}^{(1)}, x_1^{(2)}, \ldots, x_{Lm}^{(2)})$ as

$$x_k^{(1)} = \begin{cases} 1, & \textit{if } z_{ik}^{(1)} = k, \\ 0, & \textit{otherwise}, \end{cases} \quad \text{and} \quad x_k^{(2)} = \begin{cases} 1, & \textit{if } z_{ik}^{(2)} = k, \\ 0, & \textit{otherwise}, \end{cases}$$

where $k = 1, \ldots, L_m$. We further define a $2\sum_{i=1}^{M} L_m$ vector $X_i = (X_{i1}, \ldots, X_{iM})$, $i = 1, \ldots, N$. The principal component analyses will be performed on the vectors $X_1, \ldots, X_N$. This definition does not depend on the code of an allele; it then can adjust the allele effect properly.

Zhang and Zhao [2001] proposed a quantitative similarity-based association test to identify association between a candidate marker and a quantitative trait using unrelated individuals. Our method can also be extended to quantitative traits. For example, we can replace the logistic with a linear model. That is, conditional on subpopulation k, we assume

$$y_i = \mu + \beta g_i + \delta_k + \varepsilon_i,$$

where y_i now represents the quantitative trait of the i^{th} individual, and ε_i is a random error term from a normal distribution $N(0, \sigma^2)$. Then the likelihood function (2) becomes

$$L = \prod_{i=1}^{N} \sum_{k=1}^{K} \lambda_{ik} N(y_i | \mu + \beta g_i + \delta_k, \sigma^2) f(T_i/k)/f(T_i).$$

Similarly, we maximize this likelihood and use the statistic Z_β to test the null hypothesis of no association between a trait and a candidate marker. When the target population is an admixture of two subpopulations, a simple linear regression model

$$y_i = \mu + \beta g_i + \gamma t_i + \varepsilon_i$$

can be applied. This idea can also be extended to deal with the case of more than two subpopulations by using a partial linear model approach [Zhang et al., unpublished findings].

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments.

REFERENCES

Akaike H. 1974. A new look at the statistical identification model. IEEE Trans Auto Control 19: 716–23.

Allison DB. 1997. Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60:676-690.

Allison DB, Heo M, Kaplan N, Martin ER. 1999. Sibling-based tests of linkage and association for quantitative traits. Am J Hum Genet 64:1754–64.

Biernacki C, Govaert G. 1999a. Choosing models in model-based clustering and discriminant analysis. J Stat Comput Sim 64:49–71.

Biernacki C, Celeux G, Govaert G. 1999b. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. Pattern Recogn Lett 20:267–72.

Boehnke M, Langefeld CD. 1998. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–61.

Celeux G, Govaert G. 1995. Gaussian parsimonious clustering model. Pattern Recogn 28:781–91.

Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. 2000. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res 29:361–3.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55:997–1004.

George VT, Tiwari H, Zhu X, Elston RC. 1999. A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. Am J Hum Genet 65:236–45.

Horvath S, Laird NM. 1998. A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–97.

Hudson RR. 1990. Gene genealogies and the coalescent process In: Futuyma D, Antonovicsv J, editors. Oxford surveys in evolutionary biology. Oxford: Oxford University Press. p 1–44.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. Science 265:2037–48.

Pritchard JK, Donnelly P. 2001. Case-control studies of association in structured or admixed populations. Theor Pop Biol 60:227–37.

Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–8.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000a. Association mapping in structured populations. Am J Hum Genet 67:170–181.

Pritchard JK, Stephens M, Donnelly P. 2000b. Inference of population structure using multilocus genotype data. Genetics 155:945–59.

Rabinowitz D. 1997. a transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342–50.

Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50:211–23.

Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4–16.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadlan SF, Ward R, Lander ES. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Risch N. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–56.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273: 1516–7.

Satten GA, Flanders WD, Yang Q. 2001. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68:466–77.
Schaid DJ. 1996. General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–49.
Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–8.
Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–16.
Xiong MM, Krushkal J, Boerwinkl E. 1998. TDT statistics for mapping quantitative trait loci. Ann Hum Genet 62:431–52.
Zhang S, Zhao H. 2001. Quantitative similarity-based association tests using population samples. Am J Hum Genet 69:601–14.
Zhang S, Kidd KK, Zhao H. 2001. Detecting genetic association in case-control studies using similarity-based association tests. Stat Sin 12:337–59.
Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenaur D, Sun F, Kidd KK. 2000. Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 67:936–46.
Zhu X, Elston RC. 2001. Transmission/disequilibrium tests for quantitative traits. Genet Epidemiol 20:57–74.
Zhu X, Elston RC, Cooper RS. 2001. Testing quantitative traits for association and linkage in the presence or absence of parental data. Hum Hered 51:183–91.

APPENDIX

In this appendix, we give a proof that under certain conditions the distribution of the principal component is a normal mixture. For simplicity, we only give the proof for the case of a mixture of two discrete subpopulations.

Suppose we sample N_1 individuals from the first subpopulation and N_2 individuals from the second subpopulation ($N_1 + N_2 = N$). Using the same notation in the paper, let $t_{ij} = \sum_{m=1}^{M} x_m e_{mj}$ be the j^{th} principal component of the i^{th} individual, where $(e_{1j}, e_{2j}, \ldots, e_{Mj})$ is the eigenvector corresponding to the j^{th} eigenvalue and x_{im} is the marker genotypic value. Given the j^{th} principal component, if we can prove that t_{ij} is approximately normally distributed with mean μ_{1j} and variance σ_{1j}^2 for the individuals from the first subpopulation and mean μ_{2j} and variance σ_{2j}^2 for the individuals from the second subpopulation, then t_{ij} is a sample from a normal mixture.

Let q_{1m} denote the frequency of allele 2 at the m^{th} marker in the first subpopulation, with $\eta \le q_{1m} \le 1 - \eta$ for a small η, then we have the following proposition.

Proposition

Under the condition that $\max_{1 \le m \le M} e_{mj} \to 0$ as $M \to \infty$, t_{ij} is approximately normally distributed with mean μ_{1j} and variance σ_{1j}^2 in the first subpopulation, where

$$\mu_{1j} = \lim_{M \to \infty} \sum_{m=1}^{M} 2e_{mj} q_{1M} \quad \text{and} \quad \sigma_{1j}^2 = \lim_{M \to \infty} \sum_{m=1}^{M} 2e_{mj}^2 q_{1m}(1 - q_{1m}).$$

Proof

Let $y_{mj} = e_{mj} x_{im}$. Then $y_{1j}, y_{2j}, \ldots, y_{Mj}$ are independent random variables. Further, let $S_j^2 = \sum_{m=1}^{M} \mathrm{var}(y_{mj}) = 2\sum_{m=1}^{M} e_{mj}^2 q_{1m}(1 - q_{1m})$ and $a_{mj} = E(y_{mj}) = 2e_{mj} q_{1m}$. To prove the proposition, according to the Lindeberg-Feller theorem, we

only need to check the following condition: for any $\varepsilon > 0$,

$$g_M(\varepsilon) = \frac{1}{S_j^2} \sum_{m=1}^{M} \int_{|y-a_{mj}| \geq \varepsilon S_j} |y - a_{mj}|^2 dF_{y_{mj}} \to 0 \text{ as } M \to \infty,$$

where $F_{y_{mj}}$ is the distribution function of y_{mj}. Note that

$$g_M(\varepsilon) = \frac{1}{S_j^2} \sum_{m=1}^{M} \int_{|x-2q_{1m}| \geq \varepsilon e_{mj} S_j} e_{mj}^2 |x - 2q_{1m}|^2 \, dF_{x_m}$$

and the possible values of x are 0, 1, and 2. Further, noting that $|2q_{1m}| \geq 2\lambda$, $|2 - 2q_{1m}| \geq 2(1 - \lambda)$, $\max_{1 \leq m \leq M} e_{mj} \to 0$ as $M \to \infty$ and $\sum_{m=1}^{M} e_m^2 = 1$, it is easy to verify that $g_M(\varepsilon) \to 0$ for any $\varepsilon > 0$. Thus, we complete the proof of the proposition.

Using the same method, we can prove that t_{ij} is approximately normal distributed with mean μ_{2j} and variance σ_{2j}^2 for the individuals from the second subpopulation. Hence, t_{ij} is a sample from a mixture of two normal distributions.

Note

Intuitively, the condition of $\max_{1 \leq m \leq M} e_{mj} \to 0$ as $M \to \infty$ indicates that each of the markers contributes a small part of the information to the j^{th} principal component, a summary quantity measuring the difference between the two subpopulations. That is to say that if population stratification exists, we should expect to find a consistent pattern of the allele-frequency differences at many markers across the genome. In fact, this assumption is the basis of all the methods using genomic markers to control the effect of population stratification [Pritchard et al., 2000a,b; Devlin and Roeder, 1999].