

On a Semiparametric Test to Detect Associations Between Quantitative Traits and Candidate Genes Using Unrelated Individuals

Shuanglin Zhang,^{1–3} Xiaofeng Zhu,⁴ and Hongyu Zhao^{1*}

¹Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut

²Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan

³Department of Mathematics, Heilongjiang University, Harbin, China

⁴Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, Illinois

Although genetic association studies using unrelated individuals may be subject to bias caused by population stratification, alternative methods that are robust to population stratification such as family-based association designs may be less powerful. Recently, various statistical methods robust to population stratification were proposed for association studies, using unrelated individuals to identify associations between candidate markers and traits of interest (both qualitative and quantitative). Here, we propose a semiparametric test for association (SPTA). SPTA controls for population stratification through a set of genomic markers by first deriving a genetic background variable for each sampled individual through his/her genotypes at a series of independent markers, and then modeling the relationship between trait values, genotypic scores at the candidate marker, and genetic background variables through a semiparametric model. We assume that the exact form of relationship between the trait value and the genetic background variable is unknown and estimated through smoothing techniques. We evaluate the performance of SPTA through simulations both with discrete subpopulation models and with continuous admixture population models. The simulation results suggest that our procedure has a correct type I error rate in the presence of population stratification and is more powerful than statistical association tests for family-based association designs in all the cases considered. Moreover, SPTA is more powerful than the Quantitative Similarity-Based Association Test (QSAT) developed by us under continuous admixture populations, and the number of independent markers needed by SPTA to control for population stratification is substantially fewer than that required by QSAT. *Genet Epidemiol* 24:44–56, 2003. © 2003 Wiley-Liss, Inc.

Key words: coalescent model; partial linear model; population genetics; population stratification; quantitative traits; semi-parametric model; smoothing method

Grant Sponsor: National Institutes of Health; Grant number: GM59507; Grant sponsor: National Science Foundation of China; Grant number: 100710011.

*Correspondence to: Hongyu Zhao, Ph.D., Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College St., New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

Received for publication 16 April 2002; Revision accepted 20 May 2002

Published online in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/gepi.10196

INTRODUCTION

Several statistical methods were recently proposed to utilize genomic markers to control for population stratification that may be present in a sample of unrelated individuals, both in the analysis of qualitative traits [Devlin and Roeder, 1999; Bacanu et al., 2000; Devlin et al., 2001; Pritchard et al., 2000b; Reich and Goldstein, 2001; Satten et al., 2001; Zhang et al., 2002] and in the analysis of quantitative traits [Zhang and Zhao, 2001; Bacanu et al., 2002]. These approaches are

promising because they may have greater power than family-based association designs, and they may be robust against potential population stratification. These methods can be broadly divided into two classes. The first, which consists of the methods proposed by Devlin and Roeder [1999], Bacanu et al. [2000, 2002], Devlin et al. [2001], and Reich and Goldstein [2001], adjusts the ordinary chi-square test statistic χ^2 to χ^2/λ , where λ can be estimated using genomic makers. These methods assume that the parameter λ is constant across the genome. If the actual distribution has thicker tails than predicted, this could lead to a high type I error rate. For instance, selection at a candidate

locus (or at nearby loci) could drive the alleles to different frequencies in different subpopulations, making large values of the statistic more likely [Pritchard and Donnelly, 2001]. The second class of methods, including those proposed by Pritchard et al. [2000b, 2001], Zhang et al. [2002], and Zhang and Zhao [2001], involves two-step procedures. In the first step, inference is made on the population structure in the sampled individuals, using a set of independent markers. Statistical tests between candidate markers and traits of interest are then performed, using the inferred population structures in the second step. Satten et al. [2001] developed a similar approach by integrating the two steps into a single likelihood formulation. Zhu et al. [2002] recently proposed a one-step method based on the mixture model. Simulation results have found that these methods generally perform well under discrete subpopulation models. However, the population under study may be a continuous mixture of ancestral populations. In this situation, the sampled individuals may be divided into many subpopulations. Although it is not clear how such scenarios will affect the type I error and power of the methods for qualitative traits [Pritchard et al., 2000b, 2001; Zhang et al., 2002; Satten et al., 2001; Zhu et al., 2002], if the number of subpopulations identified by these methods is large, our simulation results showed that the power of such methods for quantitative traits [Zhang and Zhao, 2001] is quite low due to small sample sizes within each subpopulation.

In this article, we develop a semiparametric test for association (SPTA) to examine associations between candidate markers and quantitative traits of interest in a sample of unrelated individuals. The SPTA controls for population stratification through a set of genomic markers by first deriving a genetic background variable for each sampled individual through his/her genotypes at a series of independent markers, and then modeling the relationship between trait values, genotypic scores at the candidate marker, and a genetic background variable through a semiparametric model, where the trait value is treated as the dependent variable. The exact form of relationship between the trait value and the genetic background variable is estimated through smoothing techniques. To test whether the given set of genetic markers allows us to control for population stratification, we propose to apply SPTA to every genetic marker to obtain the corresponding P -value, and then examine if the estimated P -values have a uniform distribu-

tion, which is expected when population stratification is well controlled-for. We evaluate the performance of SPTA through simulations, both with discrete subpopulation models and with continuous admixture population models. The simulation results suggest that our procedure has a correct type I error rate in the presence of population stratification, and is more powerful than statistical association tests for family-based association designs [Monks and Kaplan, 2000; Sun et al., 2000] in all the cases considered. In addition, SPTA is more powerful than Quantitative Similarity-Based Association Test (QSAT) [Zhang and Zhao, 2001] under continuous admixture populations. Furthermore, the number of independent markers needed by SPTA to control for population stratification is substantially fewer than that required by QSAT.

METHODS

NOTATION AND STATISTICAL MODELS

We assume that the candidate marker is biallelic, with two alleles M and m . There are three genotypes at this marker: MM , Mm , and mm . For an individual, we use A to denote the additive genotypic score, with A being 1, 0, and -1 for genotypes MM , Mm , and mm , respectively. We use D to denote the dominance genotypic score, with D being 0, 1, and 0 for genotypes MM , Mm , and mm , respectively. Let y_i , A_i , and D_i denote the quantitative trait value, additive genotypic score, and dominance genotypic score of the i th individual, respectively.

For a homogeneous population, genetic association between the candidate marker and the quantitative trait can be studied through the following regression model:

$$y_i = \mu + \alpha A_i + \beta D_i + e_i, \quad (1)$$

where the e_i are assumed to be independent of each other and independent of the values of A_i and D_i , with mean 0 and variance σ^2 . In this regression model, α and β are the additive and dominance genetic values. The least squares estimates of α and β , denoted by $\hat{\alpha}$ and $\hat{\beta}$, respectively, are unbiased estimators of α and β . The null hypothesis is that there is no association between the candidate marker and the trait of interest, i.e., $\alpha = \beta = 0$. If we assume a normal distribution for the trait values, the standard F test can be performed to identify deviation from the null hypothesis. Without the assumption of

normal distribution and possible different trait variances for different genotypes, we use the statistic $T = \hat{\eta}^T V^{-1} \eta$ in this article and a permutation procedure to evaluate the P -value of the observed association, where $\hat{\eta}^T = (\hat{\alpha}, \hat{\beta})$ and V is the variance-covariance matrix of $\hat{\eta}$ up to a constant. Note that we can also use the F test combined with the permutation test to assess P -values.

The above regression method may be invalid in the presence of population stratification [Zhang and Zhao, 2001]. In a nonhomogeneous population, both phenotype mean (μ) and genetic scores (α and β) may vary across different genetic backgrounds. If there is a variable, denoted t , that can characterize the differences among individuals with different genetic backgrounds, we may model the relationship between trait values y and genotypic scores A and D through the following model

$$y_i = \mu(t_i) + \alpha(t_i)A_i + \beta(t_i)D_i + e_i, \quad (2)$$

where t_i is the genetic background variable t of the i th individual, and $\mu(\cdot)$, $\alpha(\cdot)$, and $\beta(\cdot)$ are all unknown functions. We expect that if the true genetic backgrounds of the i th and j th individuals are similar, their estimated genetic background variables t_i and t_j should also be similar. Under model (2), the null hypothesis of no association between the candidate marker and the trait of interest is $H_{01} : \alpha(t) = \beta(t) = 0$ for all t . Let us discuss two examples which can be considered special cases of model (2).

Example 1: discrete subpopulations. Assume that there are k subpopulations and each subpopulation is homogeneous. Let μ_i denote the phenotypic mean in the i th subpopulation, y_{ij} denote the trait value of the j th individual in the i th subpopulation, and A_{ij} and D_{ij} denote the additive and dominance genotypic scores of the j th individual in the i th subpopulation. A statistical model describing the relationship between trait values and genotypes is

$$y_{ij} = \mu_i + \alpha_i A_{ij} + \beta_i D_{ij} + e_{ij}. \quad (3)$$

If the genetic background variable t takes different values in different subpopulations and takes the same value within each subpopulation, it is easy to see that model (3) is a special case of model (2).

Example 2: admixture population. Consider a population that is an admixture of two ancestral populations, and suppose that an allele at a marker in the i th individual comes from the first

subpopulation with probability P_i and comes from the second subpopulation with probability $1 - P_i$. For different P_i , the phenotypic mean and the effects of genotypic scores may be different. We can model the relationship between trait values and genotypic scores by

$$y_i = \mu_0(P_i) + \alpha_0(P_i)A_i + \beta_0(P_i)D_i + e_i, \quad (4)$$

where $\mu_0(\cdot)$, $\alpha_0(\cdot)$, and $\beta_0(\cdot)$ are unknown functions of P . If the estimated genetic background variable t is a smooth function of P , model (4) is equivalent to model (2).

We propose to use the following semiparametric model, also called a partial linear model, to describe the relationship between trait values and genotypic scores:

$$y_i = \mu(t_i) + \alpha A_i + \beta D_i + e_i, \quad (5)$$

where $\mu(\cdot)$ is an unknown smooth function of the genetic background variable t , and the e_i are independent of each other and independent of t_i , A_i , and D_i . The assumption that the function $\mu(\cdot)$ is smooth is based on the consideration that similar genetic backgrounds should lead to similar phenotypic means.

This model is more general than the model discussed in Zhang and Zhao [2001], which generalized the idea of decomposing the genotypic scores into two orthogonal components, a between-family (b) component and a within-family (w) component, in the context of family-based association studies [Fulker et al., 1999; Abecasis et al., 2000; Sham et al., 2000], to study associations between candidate markers and traits of interest using population samples. The model in Zhang and Zhao [2001] has the form of

$$y_i = \mu + \alpha_b A_{bi} + \alpha_w A_{wi} + \beta_b D_{bi} + \beta_w D_{wi} + e_i, \quad (6)$$

where A_{bi} and D_{bi} are the mean of the additive genotypic score and the mean of the dominance genotypic score among the individuals in the same subpopulation as the i th individual, and $A_{wi} = A_i - A_{bi}$ and $D_{wi} = D_i - D_{bi}$ are within-subpopulation components. The genotypes on a set of independent markers can be used to infer the population structure. The null hypothesis of no association between candidate marker and trait value can be tested by examining $H_{02} : \alpha_w = \beta_w = 0$, based on model (6), and the test has been found to have a correct type I error rate even if population stratification exists. However, when the study population is a continuous admixture population, the sampled individuals may be divided into many subpopulations, based

on this approach. Although the procedure would have a correct type I error rate, the power of the test may be very low due to small sample size within each subpopulation. This phenomenon was also described by Satten et al. [2001] and similarly observed by Rosenbaum and Rubin [1984] in another context. It can be seen that model (6) used by Zhang and Zhao [2001] is a special case of the partial linear model (5) if we suitably choose the genetic background variable t . In fact, model (6) can be rewritten as

$$y_i = \mu + (\alpha_b - \alpha_w)A_{bi} + (\beta_b - \beta_w)D_{bi} + \alpha_w A_i + \beta_w D_i + e_i.$$

Both A_{bi} and D_{bi} depend on the genetic background variable of the subpopulation that the i th individual belongs to. For example, we may define functions g and f such that $A_{bi} = g(t_i)$ and $D_{bi} = f(t_i)$. Let $\mu(t_i) = \mu + (\alpha_b - \alpha_w)g(t_i) + (\beta_b - \beta_w)f(t_i)$; then we can see that model (6) is a special case of model (5).

Genetic background variable. In the above formulation, one key component is the genetic background variable which is defined for each sampled individual. We propose to use principal component analysis on a set of independent marker genotypes to estimate genetic background variables. More specifically, suppose that there are L independent markers A_l , where $l = 1, \dots, L$, and there are m_l alleles, denoted by $1, 2, \dots, m_l$, at marker A_l . We further assume that there are n individuals in our sample, and let $z_{il}^{(1)}$ and $z_{il}^{(2)}$ ($z_{il}^{(1)} \geq z_{il}^{(2)}$) denote the two alleles of the genotype of the i th individual at the l th marker, where $i = 1, \dots, n$ and $l = 1, \dots, L$. The following two methods can be used to estimate the genetic background variable.

Method 1. Let $x_{il}^{(1)} = (x_{il1}^{(1)}, \dots, x_{ilm_l}^{(1)})$ and $x_{il}^{(2)} = (x_{il1}^{(2)}, \dots, x_{ilm_l}^{(2)})$ denote two m_l -dimensional vectors and

$$\begin{aligned} x_{ilk}^{(1)} &= \begin{cases} 1, & \text{if } z_{il}^{(1)} = k, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and } x_{ilk}^{(2)} \\ &= \begin{cases} 1, & \text{if } z_{il}^{(2)} = k, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

where $i = 1, \dots, n$, $l = 1, \dots, L$ and $k = 1, \dots, m_l$. Let $X_i = (x_{i1}^{(1)}, x_{i1}^{(2)}, \dots, x_{iL}^{(1)}, x_{iL}^{(2)})^T$ be a $2m = 2 \sum_{l=1}^L m_l$ dimensional vector, where $i = 1, \dots, n$. Therefore, we have transformed the genotypes at L independent markers for the i th individual into a numerical vector X_i . Principal component analyses are then performed on the numerical vectors X_1, \dots, X_n . Let $\Sigma = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$ denote the variance-covariance matrix of

X_1, \dots, X_n and let q , a $2m$ dimensional vector, denote the eigenvector corresponding to the largest eigenvalue of Σ . Then, we use $t_i = q^T X_i$, the first principal component, to estimate the genetic background variable value for i th individual. In our analysis, the values of the t_i are standardized so that all $t_i \in [0, 1]$. Note that if the genotypes of the i th individual and the j th individual are similar, t_i and t_j should also be similar.

Method 2. The above approach needs to calculate the eigenvalues and eigenvectors of a $2m \times 2m$ matrix, which can be very large and may lead to computational problems. Therefore, in the second approach, we use a two-step procedure to assign a genetic background variable to each individual. In the first step, we only consider one marker at a time and obtain the first principal component score for every sampled individual at this marker. Using the notation given above, let $x_{il} = (x_{il}^{(1)}, x_{il}^{(2)})^T$ denote the numerical expression of the genotype of the i th individual at the l th marker, and

$$\Sigma_l = \sum_{i=1}^n (x_{il} - \bar{x}_l)(x_{il} - \bar{x}_l)^T$$

denote the variance-covariance matrix for x_{i1}, \dots, x_{in} . We further let q_l , a $2m_l$ dimensional vector, denote the eigenvector of Σ_l corresponding to the largest eigenvalue. Then $s_{il} = q_l^T x_{il}$ is the first principal component score for the i th individual at the l th marker. In the second step, we perform principal component analysis using the scores at all the L markers and assign the first principal component as the genetic background variable value. Let $S_i = (s_{i1}, \dots, s_{iL})^T$, $\Sigma = \sum_{i=1}^n (S_i - \bar{S})(S_i - \bar{S})^T$ be the variance-covariance matrix of S_1, \dots, S_n and q denote the eigenvector corresponding to the largest eigenvalue of Σ . Then $t_i = q^T S_i$ is the first principal component score of the i th individual. After standardization, we use t_i to represent the genetic background variable value of the i th individual.

Our simulations show (data not shown) that the first method is slightly better than the second method, but the difference is not significant. Because of computational concerns, we use the second method to calculate the genetic background variable for each individual in this article.

Semi parametric test of association. Having defined the genetic background variables, we now introduce the semi-parametric test of association, based on model (6). Various statistical

methods have been proposed to estimate α , β , and function $\mu(\cdot)$, including the penalized least squares [Wahba, 1984; Green et al., 1985; Engle et al., 1986; Shiau et al., 1986], the kernel smoothing method [Speckman, 1988], and the local linear method [Hamilton and Truong, 1997]. These methods differ on how to estimate the nonparametric part $\mu(\cdot)$. In this article, we use the kernel smoothing method proposed by Speckman [1988] for computational simplicity and for the well-understood statistical properties of the estimates.

To estimate α , β , and the unknown function $\mu(\cdot)$, let $y_i^* = y_i - \alpha A_i - \beta D_i$, and then equation (6) can be written as

$$y_i^* = \mu(t_i) + e_i. \quad (7)$$

If α and β are known, equation (7) is a standard nonparametric regression model, and we estimate $\mu(\cdot)$ using the kernel estimator of μ :

$$\hat{\mu}(t) = \frac{\sum_{i=1}^n y_i^* K\left(\frac{t_i-t}{h}\right)}{\sum_{j=1}^n K\left(\frac{t_j-t}{h}\right)}, \quad (8)$$

where $K(\cdot)$ is a kernel function with mode at 0, and h is the smoothing parameter. Intuitively, $\hat{\mu}(t)$ is the weighted mean of y_i^* for those individuals whose genetic background values are similar to t . We will discuss the choice of h below, and we assume the value of h is given for the moment. In this article, we use the quartic kernel [see Speckman, 1988]

$$K(t) = \begin{cases} \frac{15}{16}(1-t^2)^2, & |t| \leq 1, \\ 0, & |t| > 1. \end{cases}$$

Because the values of α and β are unknown, we cannot calculate $\hat{\mu}(t)$. However, there is no need to explicitly calculate $\hat{\mu}(t)$ to estimate α and β . Let

$$w_i(t) = \frac{K\left(\frac{t_i-t}{h}\right)}{\sum_{j=1}^n K\left(\frac{t_j-t}{h}\right)},$$

then $\hat{\mu}(t) = \sum_{i=1}^n w_i(t) y_i^*$ with $\sum_{i=1}^n w_i(t) = 1$ for any t . Replacing $\mu(t_i)$ in equation (5) by $\hat{\mu}(t_i)$, we have, after some simplifications,

$$\tilde{y}_i = \alpha \tilde{A}_i + \beta \tilde{D}_i + e_i. \quad (9)$$

where $\tilde{y}_i = y_i - \sum_{j=1}^n w_j(t_i) y_j$, $\tilde{A}_i = A_i - \sum_{j=1}^n w_j(t_i) A_j$, and $\tilde{D}_i = D_i - \sum_{j=1}^n w_j(t_i) D_j$.

Based on model (9), α and β can be estimated using standard least squares estimates (LSEs):

$$\hat{\alpha} = \frac{V_D C_{Ay} - C_{AD} C_{Dy}}{V_A V_D - C_{AD}^2} \quad \text{and} \quad \hat{\beta} = \frac{V_A C_{Dy} - C_{AD} C_{Ay}}{V_A V_D - C_{AD}^2}, \quad (10)$$

where

$$V_A = \sum_{i=1}^n \tilde{A}_i^2, \quad V_D = \sum_{i=1}^n \tilde{D}_i^2, \quad C_{Ay} = \sum_{i=1}^n \tilde{A}_i \tilde{y}_i, \\ C_{Dy} = \sum_{i=1}^n \tilde{D}_i \tilde{y}_i, \quad \text{and} \quad C_{AD} = \sum_{i=1}^n \tilde{D}_i \tilde{A}_i.$$

The estimate of $\hat{\mu}(t)$ can be obtained from equation (8) by replacing α and β with $\hat{\alpha}$ and $\hat{\beta}$, respectively.

In the Appendix, we show that $\hat{\alpha}$ and $\hat{\beta}$ are asymptotically unbiased estimates of α and β under model (5). Therefore, under the null hypothesis of no association, $E(\hat{\alpha}) = E(\hat{\beta}) = 0$ under the more general model (2). So, any valid test based on $\hat{\alpha}$ and $\hat{\beta}$ for the null hypothesis $H_0 : \alpha = \beta = 0$ under model (5) is still a valid test for the hypothesis $H_{01} : \alpha(t) = \beta(t) = 0$ for any t , under model (2). Let

$$T = \hat{\eta}^T V \hat{\eta}, \quad \hat{\eta}^T = (\hat{\alpha}, \hat{\beta}), \quad V = \begin{pmatrix} V_{A_w} & C_{A_w D_w} \\ C_{A_w D_w} & V_{D_w} \end{pmatrix}.$$

In this article, we propose to use T as our semiparametric test for association (SPTA).

To assess the statistical significance of the observed association, we permute the adjusted trait values of the sampled individuals to derive an empirical distribution for the test statistic. Note that \tilde{y}_i is the difference between the trait value of the i th individual and the weighted local mean of those individuals having a similar genetic background as the i th individual. This step is needed to remove the population stratification effect. For each simulation, we permute $\tilde{y}_1, \dots, \tilde{y}_n$ among the sampled individuals and recalculate the test statistic T based on model (9). We repeat this procedure many times to obtain an empirical sample of T . The statistical significance level of the observed test statistic can then be estimated from this empirical sample.

Smoothing parameter. We have so far assumed a given smoothing parameter in the kernel estimate. Although different kernels have little effect on the function estimation [Hart, 1997; Simonoff, 1996], the effects of the smoothing parameter h can be strong. We propose a novel method to choose h , based on the genotypes at a set of independent markers. This method can also be used to examine whether population stratification has been reasonably controlled for, using the given set of independent markers. For the simplicity of our discussion, we only consider associations between biallelic markers and traits of interest. For associations between candidate genes with multiple alleles and traits of interest,

we can either pool the alleles into two groups or use the model described in the Discussion. Suppose there are L biallelic independent markers \mathcal{A}_l where $l = 1, \dots, L$. When we perform SPTA test for all the markers, we expect that the associated P -values P_1, \dots, P_L for marker $\mathcal{A}_1, \dots, \mathcal{A}_L$ should follow a uniform distribution under the null hypothesis of no association, if population stratification is well controlled-for. We can use the Kolmogorov test to test this null hypothesis. Let F_n be the empirical distribution function of the P -values P_1, \dots, P_L and F be the uniform distribution function. The test statistic of the Kolmogorov test is $M(h) = \max_x |F_n(x) - F(x)|$, where the test statistic depends on h , and we reject the null hypothesis when $M(h)$ is large. We propose to choose h^* that minimizes the test statistic, i.e.,

$$h^* = \arg \min_h M(h). \quad (11)$$

This procedure also provides a statistical test for the effect of population stratification using the set of independent markers. If the P -value of the test statistic $M(h^*)$ is greater than a specified significance level, e.g., 0.05, we may consider that the population stratification has been well controlled-for. For the 0.05 significance level, the test statistic is not significant if $\sqrt{n}M(h^*) \leq 1.36$ [Nguyen and Roger, 1989, p. 373].

SIMULATION MODELS

Here, we discuss the simulation models used to assess whether SPTA is robust to population stratification and to compare its power with other association tests. In our simulation studies, we either generate the data through discrete subpopulation models or continuous admixture population models, and we vary genetic models in our simulations.

Discrete subpopulation models. Under discrete subpopulation models, we either generate data through coalescent models or through empirical population genetics data [Zhang and Zhao, 2001; Zhang et al., 2002].

For coalescent models, we consider three population divergence times between the two subpopulations: 1) $T = 500$ generations, 2) $T = 1,500$ generations, and 3) $T = 4,500$ generations. The sample consists of 25 individuals from the first subpopulation and 125 individuals from the second subpopulation. Based on the genotype at the candidate locus, the trait values are generated according to the following model:

$$y_{ij} = \mu_i + \alpha_i A_{ij} + \beta_i D_{ij} + e_{ij}, \quad (12)$$

where $\mu_i = \mu_{00} \times R_i$, $\alpha_i = \beta_i = \mu_0 \times R_i$, and e_{ij} is a normal random variable or a log-normal variable with mean 0 and variance 1. In our simulations, we set $R_1 = 1$ for individuals from the first subpopulation, $R_2 = 1/4$ for individuals from the second subpopulation, and $\mu_{00} = 2$. Furthermore, we set $\mu_0 = 0$ and $\mu_0 = 2$ for type I error examination and power comparison, respectively.

For simulations based on empirical population genetics data, we extract the allele frequencies of 20 SNPs across four populations, including Danes, San Francisco Chinese, Biaka, and Maya, from a population genetics database ALFRED [Cheung et al., 2000; <http://info.med.yale.edu/genetics/kkidd>] that provides allele frequencies for both SNPs and microsatellite markers in different populations. We consider different numbers of markers by using the 20 markers multiple times to infer the genetic background variable. We also assume different trait value distributions and different schemes to assign alleles conferring high trait values. We generate 250 replications, with each replication consisting of the genotype of a total of n individuals at 20 markers. Other parameters are the same as those in Zhang and Zhao [2001].

Continuous admixture population models. Under these models, we assume that the population under study is a continuous admixture of two ancestral populations. In our simulations, we use Danes and Biaka as the two ancestral populations, and extract the allele frequencies of 20 SNPs from ALFRED and repeat these 20 SNPs 2, 3, 4, ... times as if we had 40, 60, 80, ... SNPs. Let P_i represent the probability that allele M of the i th individual is from Biaka and assume P_i is uniformly distributed. Therefore, the allele frequency of allele M at locus l can be written as $P_i p_{l1} + (1 - P_i) p_{l2}$, where p_{l1} and p_{l2} are the allele frequencies of allele M at the l th marker in Biaka and Danes, respectively. The trait values are generated according to the following model:

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

where $\mu_i = \mu_{00} \times P_i$, $\alpha_i = \beta_i = \mu_0 \times P_i$, and e_i is a normal random variable or a log-normal variable with mean 0 and variance 1. In our simulations, we set $\mu_{00} = 2$. Furthermore, we set $\mu_0 = 0$ and $\mu_0 = 2$ for type I error examination and power comparison, respectively. Most genetic studies define ethnic background with five groups: Africans (or African-Americans), Europeans, Asians, Hispanic-Latinos, and American Indians. Empirical population genetics studies suggest that

populations within each group, with the exception of Africans, are more homogeneous than across these broadly defined groups. Moreover, Hispanic-Latinos are a culturally defined, rather than genetically defined, group. Therefore, we have not considered more than four populations in our simulation studies.

Other association tests considered. In addition to SPTA, we also consider three other association tests in our simulations. The first test, denoted by T_k , ignores potential population stratification. The difference between this test and SPTA is that we always treat the sampled individuals as if they were from a homogeneous population. The second test is the QSAT proposed by Zhang and Zhao [2001]. Using either discrete subpopulation models or continuous admixture population models, we also simulate a set of family triads and apply a family-based association test proposed by Monks and Kaplan [2000] and Sun et al. [2000] to determine whether there is an association between the marker and the trait. We denote this test by TDT_{MK} . In power comparisons, we simulate $n/3$, $2n/3$, and n trios, respectively, in the family-based association design, where n is the total number of individuals in the sample of unrelated individuals. We cover a range of sample sizes in the power comparisons because the amount of phenotyping and genotyping is different between the two designs for the same number of individuals. Therefore, it is difficult to select a fixed sample size to make the comparison fair. For each simulation model, we first generate $2n/3$, $4n/3$, and $2n$ individuals, respectively, in the total population as parents, and generate the children's genotypes according to their parent's genotypes. The P -values of TDT_{MK} are also evaluated by the simulations. We have not compared our approach to that of Pritchard et al. [2000b], because their proposed test only applies to qualitative traits.

RESULTS

THE GENETIC BACKGROUND VARIABLE

The genetic background variable estimated by the first principal component is a key component in our method. Our simulations based on the empirical population genetics data (results not shown) show that the first principal component can be used to well separate the subpopulations of Chinese, Danes, and Biaka, but some of the Chinese and Mayas cannot be separated, and the first principal components provide a good estima-

tion of the genetic background variable (the probability that one of the two alleles at each marker is from Biaka) under the continuous admixture model. Although there may be an inaccurate subpopulation assignment for a few individuals in each simulated sample, the simulation results showed that our method still has a correct type I error rate.

Test whether population stratification is reasonably controlled for. The first step in SPTA is to evaluate whether the population stratification can be well controlled-for with the given set of genomic markers. Our simulation results based on the empirical population genetics data showed that, for both normal and log-normal distributions, the distribution of P -values of SPTA at the independent markers is not significantly different from the uniform distribution, provided the number of markers is 120 or more. This observation is consistent with the results given in Figures 1 and 2, where we compare the type I error rates of various statistical tests. These results suggest that the Kolmogorov test described above has good utility in the determination of whether the set of genomic markers can control for population stratification.

Type I error rates. Figures 1–3 summarize the type I error rates for the four test statistics by using different numbers of markers in simulations through the coalescent models, empirical population genetics data, and continuous admixture models, respectively. The results are based on 5,000 replications (250 replications for each of 20 markers, as if there were $250 \times 20 = 5,000$ replications), with each replication consisting of $n = 150$ randomly sampled individuals for all four tests ($n/3$ triads for TDT_{MK} type tests). The 95% confidence interval of the type I error is (0.0438, 0.0562). It is apparent from the figures that the estimated type I error rates of TDT_{MK} are not statistically significant from the nominal levels in all the cases considered. In contrast, association tests based on T_k , which ignores potential population stratification, may have a type I error rate that is substantially higher than the nominal level in the presence of population stratification (in all the cases considered here). If there are enough genomic markers, the type I errors of both SPTA and QSAT are within the 95% confidence interval of the nominal type I error rate. However, in order to control the false-positives, the number of markers needed by QSAT is much larger than that needed by SPTA. It suggests that the first principal component method can extract more

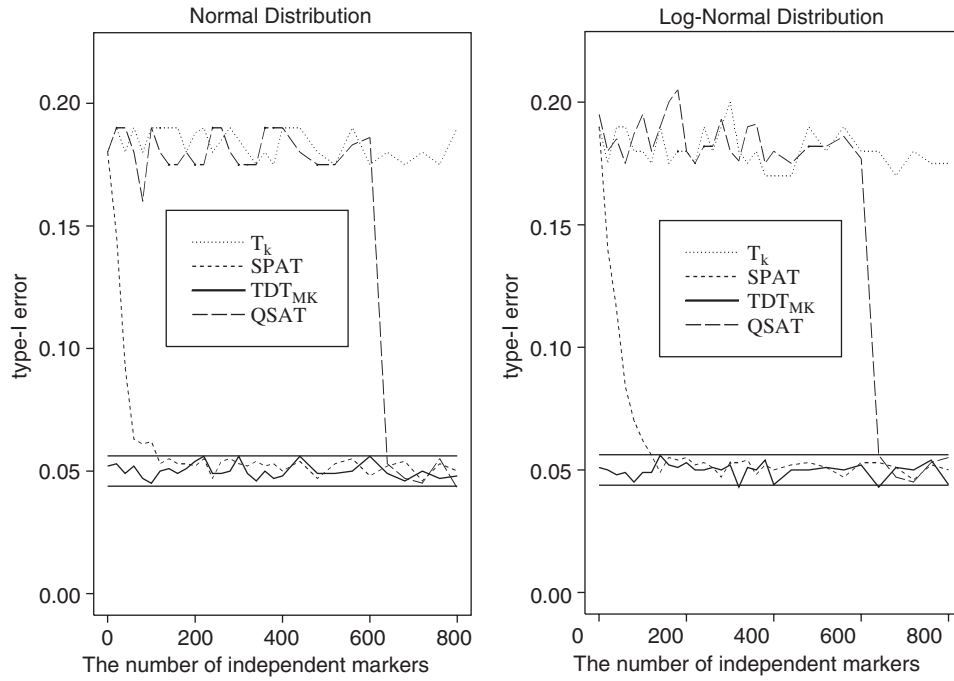


Fig. 1. Type I error comparisons of four tests (T_k , SPAT, TDT_{MK} and QSAT) at nominal value of 5% through empirical population genetics data-based simulations, using four subpopulations. Sample size is $n=150$ for T_k , SPAT, and QSAT, and 50 triads for TDT_{MK} . Two solid lines around 0.05 (nominal type I error) form the 95% confidence interval of the P -value under the null hypothesis.

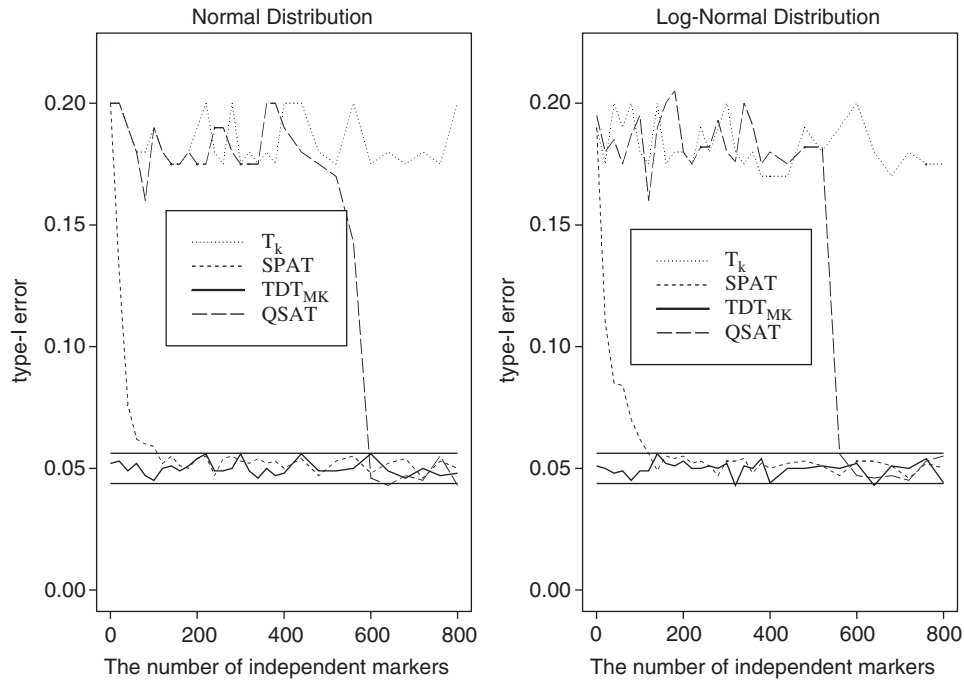


Fig. 2. Type I error comparisons of four tests (T_k , SPAT, TDT_{MK} and QSAT) at nominal value of 5% through empirical population genetics data-based simulations, under continuous admixture population models. Sample size is $n=150$ for the T_k , SPAT, and QSAT, and 50 triads for TDT_{MK} . Two solid lines around 0.05 (nominal type I error) form the 95% confidence interval of the P -value under the null hypothesis.

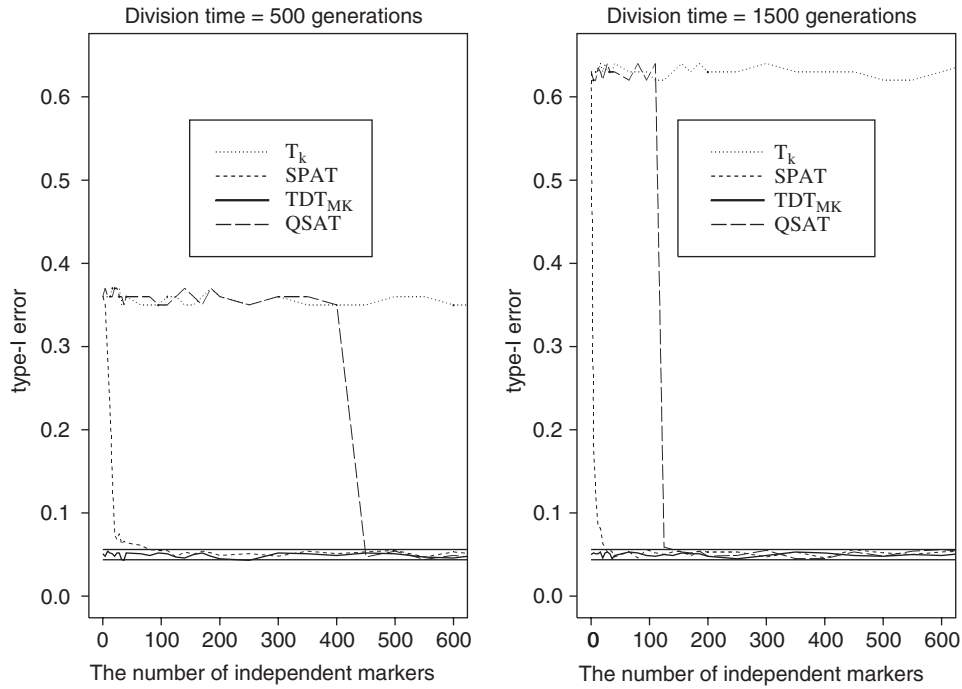


Fig. 3. Type I error comparisons of four tests (T_k , SPAT, TDT_{MK} and QSAT) at nominal value of 5% under coalescent models. Sample size is $n=150$ for T_k , SPAT, and QSAT, and 50 triads for TDT_{MK} . Two solid lines around 0.05 (nominal type I error) form the 95% confidence interval of the P -value under the null hypothesis.

relevant information from the genotypes at a set of markers than the method that is used in QSAT.

Power comparisons. The results of our power comparisons under coalescent models are summarized in Table I. The results are based on 5,000 replications, with each replication consisting of $n=150$ individuals for SPTA and QSAT and $n/3$, $2n/3$, and n triads for TDT_{MK} . Totals of 200 markers and 1,000 markers are used by SPTA and QSAT, respectively. We use different numbers of markers, because the false-positives can be controlled for by these numbers of markers by SPTA and QSAT, respectively. Both SPTA and QSAT are more powerful than TDT_{MK} for all three different sample sizes ($n/3$, $2n/3$, and n triads), and the powers of SPTA and QSAT are almost the same. We can also see that when the population divergence increases, the power of statistical tests decreases. In addition, the trait distribution and genetic models also affect the power of these tests.

Using empirical population genetics data, power comparisons are performed under several conditions, including schemes in the assignment of alleles conferring high trait values, different genetic models, different distributions of trait values, and different sample sizes for TDT_{MK} . We use $10 \times 20=200$ markers to infer the genetic background variable for SPTA and $50 \times 20=1,000$

markers to infer the relationship between each pair of individuals for QSAT. The results are summarized in Table II. Similar to the simulation results based on coalescent models, both SPTA and QSAT are more powerful than TDT_{MK} for all three different sample sizes, and the powers of SPTA and QSAT are almost the same.

The results of power comparisons under continuous admixture population models are summarized in Table III. The sample and the number of markers used are the same as those used in the power comparison, using empirical population genetics data. In this case, SPTA is more powerful than QSAT and TDT_{MK} for all three different sample sizes. For power comparisons between QSAT and TDT_{MK} , for dominant and additive models, the power of QSAT with sample size 150 is approximately the same as that of TDT_{MK} with 100 triads. However, for recessive models, the power of QSAT with sample size 150 is approximately the same as that of TDT_{MK} with 150 triads. QSAT is less powerful than SPTA, because the admixture population may be divided into too many subpopulations in QSAT under continuous admixture population models.

Although the simulation results given in this article are based on random samples of the individuals, we have also investigated selective

TABLE I. Power comparisons of three tests (SPTA, QSAT, and TDT_{MK}) under coalescent models for different trait value distributions (normal or log-normal)^a

Trait pdf	T	Model	Significance level $P=0.05$						Significance level $P=0.01$					
			SPTA	QSAT	TDT_{MK}			SPTA	QSAT	TDT_{MK}				
					$\frac{n}{3}$	$\frac{2n}{3}$	n			$\frac{n}{3}$	$\frac{2n}{3}$	n		
Normal	500	Dom.	0.99	0.99	0.60	0.81	0.85	0.97	0.97	0.39	0.67	0.78		
		Add.	0.99	0.99	0.63	0.84	0.95	0.99	0.99	0.38	0.75	0.90		
		Rec.	0.99	0.99	0.55	0.80	0.87	0.97	0.97	0.30	0.60	0.75		
	1,500	Dom.	0.99	0.99	0.50	0.72	0.82	0.98	0.97	0.30	0.55	0.69		
		Add.	0.98	0.97	0.47	0.80	0.90	0.93	0.94	0.30	0.58	0.79		
		Rec.	0.98	0.97	0.42	0.72	0.82	0.95	0.95	0.25	0.55	0.65		
	4,500	Dom.	0.98	0.97	0.43	0.66	0.76	0.93	0.92	0.25	0.45	0.60		
		Add.	0.93	0.91	0.39	0.60	0.75	0.80	0.81	0.20	0.37	0.54		
		Rec.	0.90	0.91	0.33	0.54	0.66	0.80	0.82	0.18	0.32	0.43		
Log-N	500	Dom.	0.99	0.99	0.63	0.81	0.86	0.97	0.97	0.43	0.66	0.77		
		Add.	0.99	0.99	0.72	0.89	0.96	0.98	0.98	0.53	0.77	0.91		
		Rec.	0.99	0.99	0.60	0.84	0.92	0.97	0.96	0.40	0.68	0.80		
	1,500	Dom.	0.97	0.98	0.58	0.73	0.84	0.94	0.94	0.31	0.56	0.69		
		Add.	0.97	0.96	0.64	0.80	0.89	0.93	0.92	0.41	0.68	0.78		
		Rec.	0.97	0.97	0.56	0.74	0.86	0.93	0.93	0.32	0.60	0.72		
	4,500	Dom.	0.96	0.94	0.37	0.63	0.78	0.89	0.88	0.24	0.48	0.63		
		Add.	0.89	0.89	0.51	0.63	0.76	0.75	0.77	0.31	0.48	0.63		
		Rec.	0.87	0.87	0.41	0.56	0.69	0.76	0.76	0.23	0.42	0.55		

^aSample size is $n=150$ for the SPTA and QSAT, and 50, 100, and 150 triads for the TDT_{MK} . Dom., dominant models; Add., additive models; Rec., receiving models.

TABLE II. Power comparisons of three tests (SPTA, QSAT, and TDT_{MK}) through empirical population genetics data-based simulations under the random sampling scheme^a

High-risk allele	Trait pdf	Trait model	Significance level $P=0.05$						Significance level $P=0.01$					
			SPTA	QSAT	TDT_{MK}			SPTA	QSAT	TDT_{MK}				
					$\frac{n}{3}$	$\frac{2n}{3}$	n			$\frac{n}{3}$	$\frac{2n}{3}$	n		
Fixed	Normal	Dom.	0.98	0.98	0.61	0.77	0.84	0.96	0.97	0.38	0.65	0.75		
		Add.	0.99	0.98	0.55	0.79	0.91	0.96	0.96	0.30	0.57	0.76		
		Rec.	0.91	0.90	0.40	0.56	0.64	0.86	0.84	0.19	0.37	0.49		
	Log-N	Dom.	0.97	0.98	0.64	0.79	0.85	0.94	0.95	0.45	0.67	0.77		
		Add.	0.98	0.98	0.66	0.82	0.90	0.90	0.92	0.42	0.67	0.79		
		Rec.	0.90	0.87	0.43	0.59	0.69	0.83	0.82	0.25	0.41	0.53		
Random	Normal	Dom.	0.90	0.89	0.29	0.43	0.55	0.81	0.82	0.17	0.27	0.39		
		Add.	0.91	0.91	0.31	0.61	0.80	0.82	0.83	0.20	0.44	0.64		
		Rec.	0.96	0.94	0.41	0.71	0.84	0.91	0.93	0.29	0.61	0.75		
	Log-N	Dom.	0.86	0.87	0.33	0.47	0.56	0.78	0.80	0.19	0.33	0.42		
		Add.	0.85	0.84	0.41	0.68	0.80	0.80	0.84	0.22	0.54	0.69		
		Rec.	0.95	0.94	0.50	0.70	0.85	0.90	0.89	0.30	0.60	0.77		

^aSample size is $n=150$ for the SPTA and QSAT, and 50, 100, and 150 triads for the TDT_{MK} . Dom., dominant models; Add., additive models; Rec., receiving models.

samplings, e.g., only sample top 10% and bottom 10% of the total population with respect to the trait of interest. The simulation results show a similar pattern to that of the random sampling approach.

DISCUSSION

In this article, we have developed a semiparametric test of association, SPTA, to detect associations between candidate markers and quantitative

TABLE III. Power comparisons of three tests (SPTA, QSAT, and TDT_{MK}) through empirical population genetics data based on continuous admixture models^a

Trait pdf	Disease model	Significance level $P=0.05$					Significance level $P=0.01$				
		SPTA	QSAT	TDT_{MK}			SPTA	QSAT	TDT_{MK}		
				$\frac{n}{3}$	$\frac{2n}{3}$	n			$\frac{n}{3}$	$\frac{2n}{3}$	n
Normal	Dom.	0.96	0.70	0.53	0.75	0.81	0.94	0.60	0.31	0.66	0.73
	Add.	0.99	0.83	0.48	0.82	0.97	0.98	0.80	0.34	0.68	0.90
	Rec.	0.91	0.67	0.40	0.64	0.68	0.85	0.56	0.32	0.51	0.59
Log-N	Dom.	0.98	0.70	0.57	0.72	0.80	0.96	0.60	0.41	0.67	0.72
	Add.	0.99	0.87	0.60	0.88	0.95	0.94	0.83	0.46	0.78	0.91
	Rec.	0.89	0.79	0.43	0.64	0.70	0.81	0.71	0.34	0.54	0.60

^aSample size is $n=150$ for the SPTA and QSAT, and 50, 100, and 150 triads for the TDT_{MK} . Dom., dominant models; Add., additive models; Rec., receiving models.

traits using population-based data. Our simulation results show that SPTA has a correct type I error rate under both discrete subpopulation models and continuous admixture population models. It is more powerful than family-based association tests in all the cases considered, and it is more powerful than QSAT under continuous admixture population models and requires far fewer genomic markers to control for population stratification. Furthermore, SPTA allows us to assess whether we can control for population stratification using a set of genomic markers. Moreover, it is straightforward to include covariates, and it is computationally much more efficient than structured association tests. In contrast, other several statistical methods recently proposed may have lower power to detect a true association for an admixture population, due to a potentially large number of inferred subpopulations. To evaluate the number of subpopulations inferred by these methods in the continuous admixture model (two ancestral populations) described in this article, we chose four data sets with 160 independent SNPs each and four data sets with 300 independent SNPs each from our simulations. Note that for the SPTA, the number of independent SNPs needed to control for population stratification is about 120. We ran the clustering program *Structure* developed by Pritchard et al. [2000a] with $burnin=10,000$ and $numreps=10,000$ for each of the eight data sets. For the four data sets with 160 SNPs, the estimated number of subpopulations is five for three data sets, and four for the other data set. For the four data sets with 300 SNPs, the estimated number of subpopulations is five for

two data sets and four for the other two data sets. More simulations conducted by Zhu et al. [2002] led to consistent results. These results suggest that the method developed by Pritchard et al. [2000a] may be suboptimal for admixed populations.

Although we have compared the power of SPTA with that of TDT_{MK} with three different sample sizes, the comparisons are based on the assumption that a set of independent markers is available for a genetic background variable value. If there is only one candidate locus, SPTA may require substantially more genotyping efforts. However, given the low prior probability for a specific gene to be involved for a given trait and the ever-decreasing genotyping cost, it may be more cost-effective to perform a population-based study.

Although we only use the first principal component in our simulations, multiple components (corresponding to multiple genetic background variables) can be used in our method. This leads to the question of how many principal components we should use for a given data set. Our suggestion is that we first use the first principal component. If the Kolmogorov test indicates that it cannot control for the population stratification, we will use the first two or three principal components. In general, for most of the population genetics data we have analyzed, the first three principal components account for the majority of genetic variations observed in the data.

In the case that multiallelic markers are used in an association study, we outline one approach to extending SPTA to a multiallelic trait locus.

Suppose that there are m alleles at the trait locus; hence, there are $m(m+1)/2$ genotypes. We can use $m(m+1)/2 - 1$ numerical variables to denote all the possible genotypic scores. Thus, the similar model can apply (more independent variables) to the multiallelic marker.

The SPTA proposed in this article is also applicable to qualitative traits. For example, we can let $y = 1$ for cases and $y = 0$ for controls in a case-control study. However, in this case, other models (e.g., semiparametric logistic models) may be more appropriate. Estimation methods and the performance of the test based on semiparametric logistic models need further investigation.

ACKNOWLEDGMENTS

We thank Dr. Kenneth K. Kidd for our access to the ALFRED population genetics database. We thank two anonymous reviewers and Dr. Schaid for their constructive comments on the manuscript.

APPENDIX: EXPECTATIONS OF $\hat{\alpha}$ AND $\hat{\beta}$ UNDER MODEL (5)

Let $X_1 = (\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n)^T$, $X_2 = (\tilde{D}_1, \tilde{D}_2, \dots, \tilde{D}_n)^T$, $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n)$, $X = (X_1, X_2)$, and $\eta = (\alpha, \beta)^T$. Then, $\hat{\eta} = (X^T X)^{-1} X^T \tilde{y}$. Furthermore, we make the following assumptions:

$$(S1) \quad \lim_{n \rightarrow \infty} \frac{1}{n} X^T X = \lim_{n \rightarrow \infty} \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n \tilde{A}_i^2 & \sum_{i=1}^n \tilde{A}_i \tilde{D}_i \\ \sum_{i=1}^n \tilde{A}_i \tilde{D}_i & \sum_{i=1}^n \tilde{D}_i^2 \end{pmatrix} = V > 0,$$

where $V > 0$ means that V is a positive definite matrix.

(S2) $\mu(t)$ is a continuous function.

Note that the adjusted additive and dominant genotypic scores \tilde{A}_i and \tilde{D}_i are the values of the additive and dominant genotypic scores minus their local means, respectively. Therefore, we may expect that $E(\tilde{A}_i) \approx 0$ and $E(\tilde{D}_i) \approx 0$. So, $X^T X/n$ is approximately the sample variance-covariance matrix. Intuitively, the assumption $V > 0$ means that both additive and dominant genotypic scores vary among individuals, i.e., they are not constant.

PROPOSITION

Under model (5) and assumptions (S1) and (S2), $\hat{\eta}$ is an asymptotically unbiased estimator of η , i.e.,

$$\lim_{n \rightarrow \infty, h \rightarrow 0} E(\hat{\eta}) = \eta.$$

PROOF

Let $\tilde{\mu} = (\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_n)$, where $\tilde{\mu}_i = \mu(t_i) - \sum_{j=1}^n w_j(i) \mu(t_j)$. We first prove $\lim_{h \rightarrow 0} \tilde{\mu}_i = 0$ ($i = 1, \dots, n$). From assumption (S2), we know that for any $\epsilon > 0$, there exists a constant h_0 such that $|\mu(t_i) - \mu(t_j)| \leq \epsilon$ when $|t_i - t_j| \leq h_0$. Note that if $w_j(i) \neq 0$, i.e., $K((t_i - t_j)/h) \neq 0$, then $|t_i - t_j| \leq h$ (see definition of the kernel function $K(\cdot)$ in Methods). So, for any $\epsilon > 0$, when $h \leq h_0$, we have

$$\begin{aligned} \tilde{\mu}_i &= \sum_{j=1}^n w_j(j) |\mu(t_i) - \mu(t_j)| \\ &\leq \sum_{j=1}^n w_j(i) |\mu(t_i) - \mu(t_j)| \end{aligned} \quad (A1)$$

$$\leq \epsilon \sum_{j=1}^n w_j(i) = \epsilon,$$

$$\text{i.e., } \lim_{h \rightarrow 0} \tilde{\mu}_i = 0 \text{ for } i = 1, 2, \dots, n.$$

Further, we have $\lim_{h \rightarrow 0} \frac{1}{n} \sum_{i=1}^n \tilde{A}_i \tilde{\mu}_i = 0$ and $\lim_{h \rightarrow 0} \frac{1}{n} \sum_{i=1}^n \tilde{D}_i \tilde{\mu}_i = 0$, and so

$$\lim_{h \rightarrow 0} \frac{1}{n} X^T \tilde{\mu} = 0 \quad (A2)$$

by noting that $|A_i| \leq 1$ and $|D_i| \leq 1$, and so $|\tilde{A}_i| \leq 2$ and $|\tilde{D}_i| \leq 2$.

Next, note that

$$\tilde{y}_i = y_i - \sum_{j=1}^n w_j(i) y_j = \tilde{A}_i \alpha + \tilde{D}_i \beta + \tilde{\mu}_i.$$

Therefore, $\tilde{y} = X\eta + \tilde{\mu}$, and we have

$$\begin{aligned} (X^T X)^{-1} X^T \tilde{y} &= (X^T X)^{-1} X^T (X\eta + \tilde{\mu}) \\ &= \eta + \left(\frac{1}{n} X^T X \right)^{-1} \frac{1}{n} X^T \tilde{\mu}, \end{aligned} \quad (A3)$$

and the proposition follows from (A3), (A2), and assumption (S1).

REFERENCES

- Abecasis GR, Cardon LR, Cookson OC. 2000. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279–92.

- Bacanu SA, Devlin B, Roeder K. 2000. The power of genomic control. *Am J Hum Genet* 66:1933–44.
- Bacanu SA, Devlin B, Roeder K. 2002. Association studies for quantitative traits in structured populations. *Genet Epidemiol* 22:78–93.
- 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.
- Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55:997–1004.
- Devlin B, Roeder K, Wasserman L. 2001. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 60:155–66.
- Engle R, Granger C, Rice J, Weiss A. 1986. Nonparametric estimates of the relation between weather and electricity sales. *J Am Stat Assoc* 81:310–320.
- Fulker DW, Cherny SS, Sham PC, Hewitt JK. 1999. Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* 64:259–67.
- Green P, Jennison C, Seheult A. 1985. Analysis of field experiments by least squares smoothing. *J R Stat Soc B* 47:299–315.
- Hamilton SA, Truong YK. 1997. Local linear estimation in partly linear model. *J Multi Anal* 60:1–19.
- Hart JD. 1997. *Nonparametric smoothing and lack-of-fit tests*. New York: Springer.
- Monks SA, Kaplan NL. 2000. Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. *Am J Hum Genet* 66:576–92.
- Nguyen HT, Rogers GS. 1989. *Fundamentals of mathematical statistics: volume II: statistical inference*. New York: Springer-Verlag.
- Pritchard JK, Donnelly P. 2001. Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 60:227–37.
- Pritchard JK, Stephens M, Donnelly P. 2000a. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000b. Association mapping in structured population. *Am J Hum Genet* 67:170–81.
- Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 20:4–16.
- Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 79:516–24.
- 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.
- Sham PC, Cherny SS, Purcell S, Hewitt JK. 2000. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 66:1616–30.
- Shiau J, Wahba G, Johnson DR. 1986. Partial spline models for the inclusion of tropopause and frontal boundary information in otherwise smooth two and three dimensional objective analysis. *J Atmos Ocean Technol* 3:714–25.
- Simonoff JS. 1996. *Smoothing methods in statistics*. New York: Springer.
- Speckman P. 1988. Kernel smoothing in partial linear models. *J R Stat Soc B* 50:413–36.
- Sun F, Flanders WD, Yang Q, Zhao HY. 2000. Transmission/disequilibrium tests for quantitative traits. *Ann Hum Genet* 64:555–65.
- Wahba G. 1984. Partial spline models for semiparametric estimation of functions of several variables. In *Analyses for Time Series, Japan-U.S. Joint Sem*, 319–29. Tokyo: Institute of Statistical Mathematics.
- Zhang SL, Zhao HY. 2001. Quantitative similarity-based association tests using population samples. *Am J Hum Genet* 69:601–14.
- Zhang SL, Kidd KK, Zhao HY. 2002. Detecting genetic association in case-control studies using similarity-based association tests. *Stat Sin* 12:337–59.
- Zhu X, Zhang SL, Zhao HY, Cooper RS. 2002. Association mapping using a mixture model for complex traits. *Genetic Epidemiol* 23:181–196.