

# Qualitative Semi-Parametric Test for Genetic Associations in Case-Control Designs Under Structured Populations

H.-S. Chen<sup>1</sup>, X. Zhu<sup>2</sup>, H. Zhao<sup>3</sup> and S. Zhang<sup>1,4,\*</sup>

<sup>1</sup>Department of Mathematical Sciences, Michigan Technological University, Houghton, MI

<sup>2</sup>Department of Preventive Medicine and Epidemiology, Loyola Medical School, Maywood, IL

<sup>3</sup>Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT

<sup>4</sup>Department of Mathematics, Heilongjiang University, Harbin, China

---

## Summary

Recently, statistical methods have been proposed using genomic markers to control for population stratification in genetic association studies. However, these methods either have unacceptable low power when population stratification becomes strong or cannot control for population stratification well under admixture population models. In this paper, we propose a semiparametric association test to detect genetic association between a candidate marker and a qualitative trait of interest in case-control designs. The performance of the test is compared to other existing methods through simulations. The results show that our method gives correct type I error rate both under discrete population models and admixture population models, and our method is robust to the extent of the population stratification. In most of the cases we considered, our method has higher power and, in some cases, substantially higher power than that of existing methods.

---

Keywords: population stratification, case-control study, semi-parametric model, smoothing method

## Introduction

Recently, several methods have been proposed to utilize a set of independent markers, referred as genomic markers, to control for population stratification (Devlin & Roeder, 1999; Bacanu *et al.* 2000; Pritchard *et al.* 2000b; Devlin *et al.* 2001; Reich & Goldstein, 2001; Satten *et al.* 2001; Zhang & Zhao, 2001; Bacanu *et al.* 2002; Zhang *et al.* 2002; Zhu *et al.* 2002). These methods may have greater power and be easier to collect DNA sample than that of family-based association designs, and they may be also robust against potential population stratification. These methods can be broadly divided into two classes. The first

class, called the GC (Genomic Control) method, consists of the methods proposed by Devlin & Roeder (1999), Bacanu *et al.* (2000), Devlin *et al.* (2001), and Reich & Goldstein (2001). They adjust the ordinary chi-square test statistic  $X^2$  to  $X^2/\lambda$ , where  $\lambda$  can be estimated using genomic markers and it is assumed that  $X^2/\lambda$  still has a  $\chi^2$  distribution. If the sample comes from the population which consists of two subpopulations with different disease prevalences and different allele frequencies, as pointed out by Pritchard & Rosenberg (1999) and Zhang *et al.* (2003),  $E(X) \neq 0$  and, up to a constant,  $X^2$  will asymptotically follow a noncentral  $\chi^2$  distribution  $\chi^2(\delta)$ . If the noncentral parameter  $\delta$  is small, the  $\chi^2$  distribution adjusted by a suitable constant can be a good approximation of  $\chi^2(\delta)$ . However, if  $\delta$  is large, using a  $\chi^2$  adjusted by a constant as an approximation of a noncentral chi-square  $\chi^2(\delta)$  will either lead to false positives or lose power, though we do not know yet how large  $\delta$  may be in practice. The second

\* Corresponding author: Shuanglin Zhang, Ph.D., Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931. Phone: (906) 487-2175; Fax: (906) 487-3133 (Fax to Shuanglin Zhang). E-mail: shuzhang@mtu.edu

class of methods, including those proposed by Pritchard *et al.* (2000, 2001), Satten *et al.* (2001), Zhang *et al.* (2002) and Zhu *et al.* (2002) for qualitative traits and Zhang & Zhao (2001) for quantitative traits, essentially use clustering methods to infer the population structure and incorporating this information into the association test. Although simulation results have found that these methods generally perform well under discrete subpopulation models, they may not be effective when the population under study is a mixture of ancestral populations. In this situation, inference of population structure (the number of ancestral populations and/or the probabilities that each individual belongs to every subpopulation) is very difficult, because virtually everybody in the sample is admixed, and there is little information about the ancestral populations (Pritchard *et al.* 2000b). Inaccurate estimate of the population structure will affect the validity of the test.

For a quantitative trait of interest, Zhang *et al.* (2003) proposed a Semi-Parametric Test of Association to avoid such aforementioned drawback. The test is derived by first estimating the genetic background variable value for each sampled individual using the principal components of many independent marker genotypes, and then modeling the relation of trait values, genotypic scores of candidate marker and the genetic background variable through a semi-parametric model. In this model, the trait value is treated as the dependent variable, genotypic scores as linear predictors and genetic background variable also as predictor but the relationship between the trait value and the genetic background variable is an unknown nonparametric function (possibly nonlinear). The simulation results indicate that this method possesses correct type-I error rate, and is more powerful than a TDT-like test in all cases considered.

In this article, we develop an analogous Qualitative Semi-Parameter Test (QualSPT) for the qualitative trait of interest. The method also hinges on finding the background variable of each sampled individual via the principal component procedure. We model the relationship between a disease status and genotypic scores and the genetic background variable through a semi-parametric logistic regression model. The model is semiparametric in the sense that the relation between the disease probability and the genetic background variable has an unknown form of function, that is, a nonparametric func-

tion, while the genotypic scores contribute to the logit disease probability in a linear function. The method is easy to include other variables. The estimation of the parameters in the linear function and the estimation of the nonparametric function can be carried out by the profile likelihood estimation method. To test if the given set of the independent markers can well control for population stratification, we perform the test to every independent marker and get a series of p-values, and then test if the p-values are uniformly distributed using the Kolmogorov test statistic. Uniformly distributed p-values imply that the population stratification is well controlled.

We evaluate the performance of the QualSPT through simulations both with discrete population models and admixture population models. The simulation results show that our procedure has correct type-I error rate in the presence of population stratification using a sufficient number of independent markers and is more powerful than other statistical association tests for family-based association designs (Spielman *et al.* 1993) using the same number of individuals. Furthermore, QualSPT is more powerful than STRAT (Pritchard *et al.* 2000b) when STRAT overestimates the number of subpopulations (i.e., ancestral populations). In addition, QualSPT is much more powerful than the GC method (Devlin & Roeder, 1999; Bacanu *et al.* 2000) when population stratification is strong.

## Method

### Notation and Statistical Model

Let  $X$  denote a vector of the numerical codes for the candidate locus genotype,  $g$  (see section "genotype scoring" for detail). Let  $y$  denote the trait value. For the case-control design considered in this article,  $y = 1$  denotes disease and  $y = 0$  denotes normal. Let  $(y_1, X_1), \dots, (y_n, X_n)$  denote the trait values and genotypic codes of  $n$  individuals. For a homogeneous population, genetic association between the qualitative trait and a candidate marker can be studied through the following logistic regression model:

$$\log \frac{P(y_i = 1 | X_i)}{1 - P(y_i = 1 | X_i)} = \mu + X_i' \beta, \quad (1)$$

where  $\beta$  represents the effect of the genotype on the trait and  $\mu$  is the intercept. We call  $\mu$  the logistic phenotypic

mean in the following discussion. Under this model, the log-likelihood function for given  $X_i$  is given by

$$\sum_{i=1}^n [y_i(\mu + X_i'\beta) - \log(1 + e^{\mu + X_i'\beta})].$$

The maximum likelihood estimate of  $\beta$  is asymptotically unbiased and the likelihood ratio test can be used to test the null hypothesis of no association,  $H_0: \beta = 0$ . The advantage of this model is that it is straightforward to include environmental variables and to extend to multiple loci and gene-gene interactions.

In the presence of population stratification, however, model (1) may be invalid. Specifically, if the individuals come from different subpopulations with different  $\mu$  in model (1) and different allele frequencies, then  $\hat{\beta}$  will not be an unbiased estimator of  $\beta$  (see Appendix). The test mentioned above will lead to false-positives due to population stratification. Similar problem occurs if the sampled individuals come from an admixture population. Consequently, model (1) will not be valid to model the relationship between the trait and the marker genotype. In a non-homogeneous population,  $\mu$  may vary across different genetic background. Zhang *et al.* (2003) introduced a genetic background variable  $t$  (may be multi-dimensional) that can characterize the difference among the individuals with different genetic backgrounds and estimate the genetic background variable  $t$  using the principal components of the genotypes of many independent markers across genome. In this article, we use the same method as that proposed by Zhang *et al.* (2003) to estimate the genetic background variable  $t$ . Briefly, let  $x_{ij}$  denote a column vector of the numerical codes of the  $i$ th individual at the  $j$ th maker and  $x_i = (x'_{i1}, x'_{i2}, \dots, x'_{iL})'$  be the numerical code vector of the  $L$ -marker genotype of the  $i$ th individuals, where  $x'_{ij}$  denotes the transpose of vector  $x_{ij}$ . If the  $j$ th marker is a biallelic marker with allele  $A_1$  and  $A_2$ , we can code  $x_{ij} = 0, 1$  and  $2$  for genotype  $A_1A_1, A_1A_2$  and  $A_2A_2$ , respectively. For the marker with more than two alleles, the detail of coding  $x_{ij}$  can be found in Zhang *et al.* (2003). The principal component technique is used to reduce the dimension of the data. Let  $\Sigma$  denote the sample variance-covariance matrix of the sample  $x_1, x_2, \dots, x_n$  and  $e_j$  denote the eigenvector corresponding to the  $j$ th largest eigenvalue of  $\Sigma$ . We can then calculate the  $j$ th principal component of the  $i$ th individual as  $t_{ij} = x'_{ij}e_j$ .

Let  $t_i = (t_{i1}, t_{i2}, \dots, t_{iM})$  denote the first  $M$  principal components of the  $i$ th individual. We use  $t_i$  as the estimated value of the genetic background variable of the  $i$ th individual. We will discuss how to choose  $M$  later.

To take the genetic background effect into account, we assume the following semi-parametric model or logistic partial linear model:

$$\log \frac{P(y = 1|X, t)}{1 - P(y = 1|X, t)} = X'\beta + \mu(t), \tag{2}$$

where  $\mu(t)$  is a unknown smooth function of the genetic background  $t$  and is not parametrized. In this article, a smooth function means that the function has continuous derivative. The assumption that the function  $\mu(\cdot)$  is smooth is based on the consideration that similar genetic backgrounds should lead to similar phenotypic means. For the sake of simplicity, we do not include any covariates in model (2), but this is not a necessary restriction.

The log-likelihood score of the  $i$ th individual for a given genotype code  $X_i$  and genetic background variable  $t_i$  is given by

$$\begin{aligned} l(\beta, \mu(t_i), X_i, y_i) &= \log(\Pr(y_i|X_i, t_i)) \\ &= y_i[X_i'\beta + \mu(t_i)] \\ &\quad - \log[1 + \exp(X_i'\beta + \mu(t_i))]. \end{aligned}$$

The log-likelihood function for the data of  $n$  individuals is given by

$$l(\beta, \mu) = \sum_{i=1}^n l(\beta, \mu(t_i), X_i, y_i).$$

### Estimation and Test of Association

The estimation of parameter  $\beta$  and non-parametric function  $\mu(t)$  under logistic partial linear models has been developed recently. Several methods have been proposed in the statistics literature, for example, see Severini & Wong (1992), Severini & Staniswalis (1994), and Carroll *et al.* (1997). Here we follow the approach of Severini & Staniswalis (1994), which uses two different likelihood functions, likelihood function  $l(\beta, \mu)$  and a local likelihood function. The local likelihood function around point  $t$  is defined as

$$\sum_{i=1}^n K\left(\frac{t_i - t}{h}\right) l(\beta, \eta, X_i, y_i), \tag{3}$$

where  $t$  is an  $M$  dimensional real vector;  $h$  is a parameter called smoothing parameter; the function  $K(\cdot)$  is

a real function on  $R^M$  called kernel function with the properties that  $K(\cdot)$  reaches its maximum value at origin. Although different kernels have little effect on the estimation, the effects of the smoothing parameter  $h$  can be strong (Hart, 1997; Simonoff, 1996). In this paper, we use the quadratic kernel  $K(z) = \prod_{i=1}^M k(z_i)$  with

$$k(z_i) = \begin{cases} (1 - z_i^2)^2 & \text{if } |z_i| \leq 1, \\ 0 & \text{if } |z_i| > 1 \end{cases}$$

and where  $z = (z_1, z_2, \dots, z_M)'$ . At present stage, we assume the value of smoothing parameter  $h$  is known. The method of choosing smoothing parameter  $h$  will be discussed in later section.

The estimation procedure has several steps. First, for  $t = t_j$ , maximize the local log-likelihood function (3) with respect to  $\eta$ , assuming a fixed  $\beta$ . The maximizer  $\hat{\mu}_\beta(t_j) = \hat{\eta}$  is an estimator of  $\mu(t_j)$  for  $j = 1, 2, \dots, n$ . The values  $\hat{\mu}_\beta(t_1), \hat{\mu}_\beta(t_2), \dots, \hat{\mu}_\beta(t_n)$  are then used to estimate  $\beta$  by maximizing the log-likelihood function with respect to  $\beta$ ,

$$l(\beta, \hat{\mu}_\beta) = \sum_{i=1}^n l(\beta, \hat{\mu}_\beta(t_i), X_i, \gamma_i). \quad (4)$$

Although there are no explicit solutions from (3) and (4), the estimation problem can be solved iteratively as follows:

Step 1. Solve the equation of  $\eta$

$$\sum_{i=1}^n K\left(\frac{t_i - t}{h}\right) \frac{\partial}{\partial \eta} l(\beta_m, \eta, X_i, \gamma_i) = 0. \quad (5)$$

Denote  $\hat{\mu}_m(t_1), \hat{\mu}_m(t_2), \dots, \hat{\mu}_m(t_n)$  be the solutions of  $\eta$  for  $t = t_1, t = t_2, \dots, t = t_n$ , respectively. Here  $\beta_m$  is the current estimated value of  $\beta$ .

Step 2. Solve the equation of  $\beta$

$$\sum_{i=1}^n \frac{\partial}{\partial \beta} l(\beta, \hat{\mu}_m(t_i), X_i, \gamma_i) = 0. \quad (6)$$

This yields the new parameter estimate  $\beta_{m+1}$ .

We repeat this two-step process until convergence occurs.

In the appendix, we show that  $\hat{\beta}$ , the estimator of  $\beta$ , is an asymptotically unbiased estimate of  $\beta$  under model (2). Therefore, under the null hypothesis of no association,  $E(\hat{\beta}) \approx 0$ . Our QualSPT test can be constructed

based on the likelihood ratio test using the estimates calculated from the aforementioned procedure, and the test statistic has a  $\chi^2$  distribution with degrees of freedom equal to the dimension of the numerical vector of the candidate locus genotype.

### Choosing Smoothing Parameter and the Number of Principal Components

The test procedure mentioned above assumes a given smoothing parameter  $h$ . However, the effect of the smoothing parameter  $h$  can be strong. We follow the method proposed in Zhang *et al.* (2003) to choose the smoothing parameter  $h$ . Briefly, for any given value of  $h$ , let  $p_1(h), \dots, p_L(h)$  be the p-value when we perform QualSPT for all the independent markers using the given value of  $h$ . We denote the empirical distribution of the p-value based on the sample  $p_1(h), \dots, p_L(h)$  by  $F_n(x, h)$ . The Kolmogorov test statistic is defined as  $M(h) = \max_x |F_n(x, h) - F(x)|$ , with  $F$  being the uniform distribution function. The method is to choose the smoothing parameter  $h^*$  such that it minimizes the Kolmogorov test statistic

$$M(h^*) = \min_{0 < h < 1} M(h).$$

The rationale of the method is that for independent markers, the p-values should follow a uniform distribution if population stratification is well controlled for. The purpose of using independent markers is to control population stratification. Therefore, the smoothing parameter  $h$  is chosen to minimize the difference between the empirical distribution and the uniform distribution. This procedure also provides a statistical test to assess if population stratification can be well controlled for by 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 i.e. the population stratification is reasonably controlled, if  $\sqrt{n}M(h^*) \leq 1.36$  (see Nguyen & Roger, 1989, p. 373). Pritchard *et al.* (2000b) also suggested this idea to estimate the number of subpopulations used in their test STRAT.

For a given data set, another question is that how many principal components we should use. Our

suggestion is that, first we use only the first principal component. If the Kolmogorov test shows that the population stratification can not be well controlled for, we will use more principal components. Further discussion about how many principal components we should use will be given in the Discussion section.

### Genotype Scoring

The candidate locus genotype can be scored in a number of ways. A simple scheme is to count the number of alleles that an individual possesses at the candidate locus. If there are  $m$  alleles,  $A_1, A_2, \dots, A_m$ , at the candidate locus, this scheme is to create a numerical vector  $X = (x_1, \dots, x_{m-1})$  for the genotype, where  $x_i$  is the number of allele  $A_i$  in the genotype ( $i = 1, 2, \dots, m - 1$ ). This scheme only accounts for the additive effect of the alleles and will lead to a  $\chi^2$  distribution with the degrees of freedom  $m - 1$  of the statistic of QualSPT. Another scheme is as follows: Let  $G_1, G_2, \dots, G_{m(m+1)/2}$  denote the total possible genotypes of the  $m$  alleles. Define  $X = (x_1, \dots, x_{m(m+1)/2-1})$  as the numerical vector of genotype  $G$ , where

$$x_i = \begin{cases} 1 & \text{if the genotype } G = G_i \\ 0 & \text{otherwise.} \end{cases}$$

This scheme will account for both the additive and dominant effect of the alleles. However, it makes the degrees of freedom much larger, i.e.  $m(m + 1)/2 - 1$ .

For a biallelic locus,  $m = 2$ , we usually use the second scheme to score the genotype and this is what we use in our simulations. If  $m$  is large, the first scheme may be more powerful. However, if we know some prior information of the heredity model, it will be helpful for the scoring scheme. For example, for a recessive disease, we may score  $x_i$  to be 1 for genotype  $A_i A_i$  and 0 for all other genotypes.

### Simulation Models

In this section, we discuss the simulation models used to assess whether QualSPT is robust to population stratification and to compare the power of QualSPT with other association tests. In order to compare the performance of QualSPT with that of the Genome Control method (GC) developed by Devin *et al.* (1999) and Bacanu *et al.* (2000), we only consider biallelic markers in our sim-

ulations, because the GC method is only applicable to biallelic markers. In our simulation studies, we either generate the data through discrete subpopulation models or continuous admixture population models. Other parameters varied in our simulations include different modes of inheritance and different prevalences among the subpopulations.

### Discrete Subpopulation Models

We use empirical population genetics data from a population genetics database ALFRED (Osier *et al.* (2001); <http://info.med.yale.edu/genetics/kkidd>) that provides allele frequencies for both SNPs and microsatellite markers in different populations. For our simulation purposes, we extract 100 markers across four populations, including Danes, San Francisco Chinese, Maya and Biaka. For microsatellite markers, because we focus on the use of SNP markers in our simulation, we pool the alleles to form biallelic markers with allele frequencies between 10% and 90%. We consider different numbers of markers, 100, 200,  $\dots$ , 500 by using the 100 markers multiple times to infer the genetic background variable.

Let  $f_i$  denote the probability that an affected individual is sampled from the  $i$ th subpopulation,  $g_i$  denote the probability that a normal individual is sampled from the  $i$ th subpopulation, and  $P_i$  be the prevalence of the disease in the  $i$ th subpopulation. For a rare disease,  $\frac{f_i}{f_j} \approx \frac{P_i g_i}{P_j g_j}$  (Pritchard & Rosenberg, 1999; Zhang *et al.* 2002). We sample 50, 15, 15 and 20 normal individuals from Danes, San Francisco Chinese, Maya and Biaka, respectively. We consider two cases of relative prevalences  $P_1:P_2:P_3:P_4 = 1:2:3:4$  and  $P_1:P_2:P_3:P_4 = 1:4:6:8$  which correspond to sample 24, 14, 23, 39 and 13, 16, 27, 44 diseased individuals from the four subpopulations, respectively.

In our assessment of whether QualSPT is robust to population stratification, we independently generate marker data 10 times for every marker among the 100 markers mentioned above, i.e. we generate  $10 \times 100 = 1000$  markers that have no association with disease phenotype. For each case of relative prevalences, we perform statistical tests for each of the 1000 markers.

To compare the power of QualSPT with other statistical tests, we generate 1000 data set. For each data set,

the genotypes for the trait locus are resimulated using the marker allele frequencies of each of the 100 loci in turn. Let  $A$  and  $a$  denote the two alleles and  $f_{11}$ ,  $f_{12}$ , and  $f_{22}$  denote the penetrances for genotypes  $AA$ ,  $Aa$ , and  $aa$ , respectively ( $f_{12} = f_{11}$  or  $f_{22}$  corresponds to a dominant or recessive disease model). Let the relative risk  $R_A = f_{11}/f_{22}$ . For a given  $R_A$  value and the mode of inheritance, the proportions of affected individuals with genotypes  $AA$ ,  $Aa$  and  $aa$  can be easily calculated. Let  $R_{A_1}$ ,  $R_{A_2}$ ,  $R_{A_3}$ , and  $R_{A_4}$  denote the relative risk in Danes, San Francisco Chinese, Maya, and Biaka, respectively. In our simulations, we vary the values of  $R_{A_i}$  ( $i = 1, 2, 3, 4$ ) and disease models.

### Admixture Populations

We assume that the population under study is the admixture of two ancestral populations. In our simulations, we use Danes and Biaka as the two ancestral populations to represent Europeans and Africans, and extract the allele frequencies of 100 biallelic markers as mentioned above from ALFRED and repeat these 100 markers if we need more marker data. We simulate data sets using similar model to model C in Pritchard *et al.* (2000b). For each individual, we first simulate  $q$ , where  $q$  is the fraction of European ancestry and  $1 - q$  the fraction of African ancestry. Then, at each locus, two alleles are drawn independently with probability  $q$  from Danes, and probability  $1 - q$  from Biaka allele-frequency distributions.

*Model A.* we assume that  $q$  is uniformly distributed in interval  $(0, 1)$ . Normal individuals are sampled from this distribution. In our assessment of type-I errors, 100 normal individuals and 100 diseased individuals are sampled. In order to simulate diseased individuals, we assume that the prevalence of the disease is eight-fold higher in Danes than in Biaka. The rejection sampling as described in Pritchard *et al.* (2000b) is used to simulate the diseased individuals.

To compare the power, let  $R_{g_i}$  denote the relative risk of individual  $i$  with genotype  $g_i$ . We consider two same type alleles as different alleles if they come from different ancestral subpopulations. For example, we use  $A_1$  and  $A_2$  to denote the  $A$  allele which is transmitted from Danes and Biaka, respectively. So, the relative risk  $R_{g_i}$  depends on both the genetic background  $q_i$

and the genotype at the candidate locus. We then simulate the genotypes of the diseased individuals using the probability

$$\Pr[g_i | q_i, \text{diseased}] = \frac{R_{g_i} \Pr[g_i | q_i]}{\sum_g R_g \Pr[g | q_i]}.$$

For the details, see Pritchard *et al.* (2000b). In our simulations, we let  $R_{a_1a_1} = R_{a_1a_2} = R_{a_2a_2} = 1$  and varied  $R_{A_1A_1}$  and  $R_{A_2A_2}$  and set  $R_{A_1A_2} = (R_{A_1A_1} + R_{A_2A_2})/2$ . The relative risk of other genotypes is calculated according to different disease models. For example,  $R_{A_1a_2} = (R_{A_1A_1} + R_{a_2a_2})/2$  under the additive disease model.

*Model B.* Instead of using the uniform distribution to generate  $q$  in model A, we generate  $q$  from beta distribution  $B(2, 6)$  for normal individuals and from  $B(\alpha, 2)$  for diseased individuals. This means that, on average, 1/4 of the genetic materials are from Danes and 3/4 are from Biaka for normal individuals. For diseased individuals, on average,  $\alpha/(\alpha + 2)$  of the genetic materials are from Danes and  $2/(\alpha + 2)$  are from Biaka. We use either  $\alpha = 2$  or  $\alpha = 4$  in our simulations.

For the above simulations, we assume that all the subpopulations have the same high risk allele. We also conduct another set of simulations for power comparison under admixture model B and  $\alpha = 2$ , where we randomly assign high risk allele independently in each of the ancestral population according to the allele frequency. For example, consider a candidate locus with two alleles  $A$  and  $a$ . The allele frequencies of allele  $A$  in two ancestral populations are  $p_1$  and  $p_2$ , respectively. Then, for each of the 1000 replicated data sets,  $A$  is assigned as the high risk allele with probabilities  $p_1$  and  $p_2$  in the two ancestral populations, respectively.

### Other Association Tests Considered

In addition to QualSPT, we also consider several other association tests in our simulations. The first test is the  $\chi^2$  test that ignores potential population stratification. The second test is the GC (Genomic Control) method developed by Devlin & Roeder (1999) and Bacanu *et al.* (2000), which uses the test statistic  $\chi^2/\lambda$  and the parameter  $\lambda$  is estimated by  $\hat{\lambda} = \text{median}(\chi_1^2, \chi_2^2, \dots, \chi_L^2)/0.456$ , where  $\chi_i^2$  is the value of  $\chi^2$  test statistic for the  $i$ th independent marker. The third test is STRAT

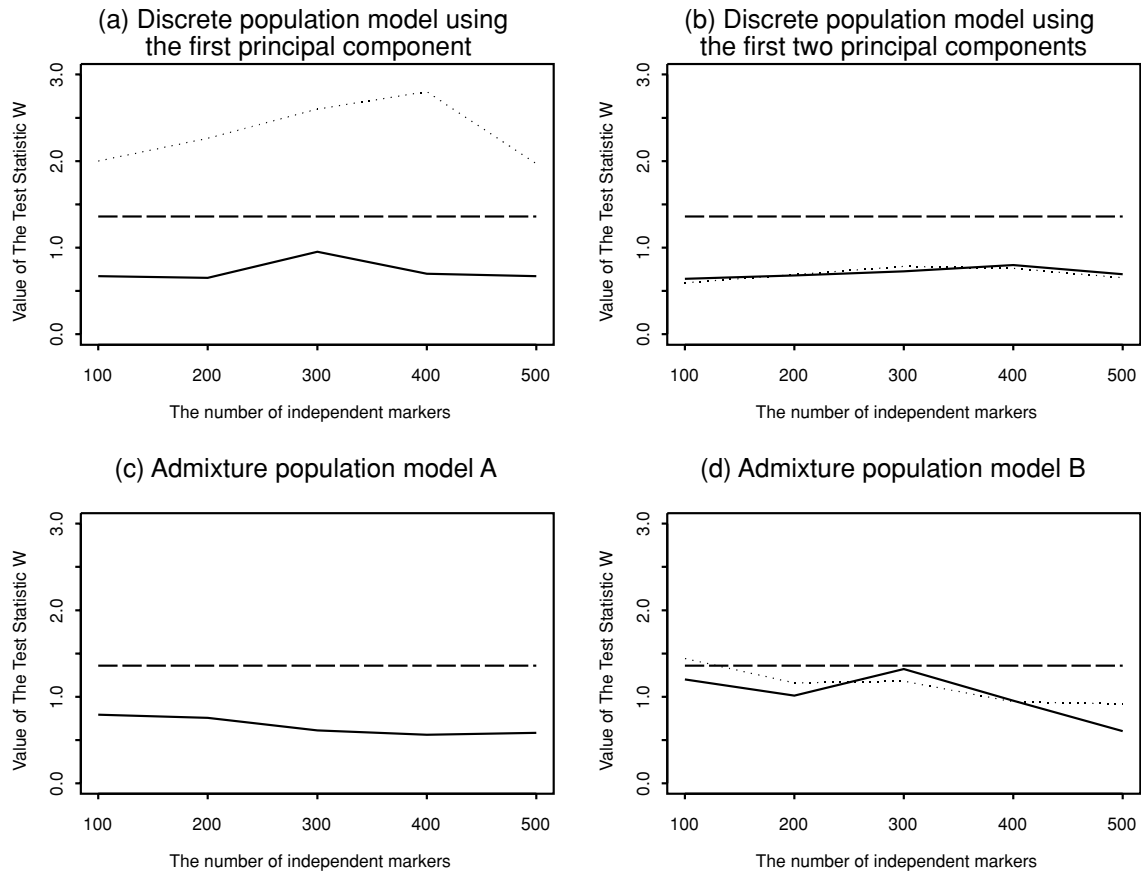
proposed by Pritchard *et al.* (2000b). To perform the test of STRAT, we first use *Structure* (Pritchard *et al.* 2000a) to estimate the probabilities that each individual belongs to each subpopulation, under the assumption that the number of subpopulations is known. Using either discrete subpopulation models or admixture population models, we also simulate a set of family triads and apply a TDT test proposed by Spielman *et al.* (1993) to determine whether there is an association between the marker and the trait. We denote this test by TDT. In power comparisons, we simulate  $2n/3$  and  $n$  triads, respectively, in the family-based association design, where  $n$  is the total number of diseased individuals in the sample of unrelated individuals. The reason that we cover a range of sample sizes in the power comparisons is that 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  $2n/3$  and  $n$  diseased individuals, respectively, in the total population as children, and then generated their parents' genotypes. The p-values of the TDT are also evaluated by the simulations.

## Results

*Test whether population stratification is reasonably controlled for:* The first step of QualSPT is to evaluate whether the population stratification can be well controlled by a given set of genetic markers. We begin with the use of the first principal component. If the Kolmogorov test indicates that the population stratification can not be well controlled for, we will use the first two or three principal components (see how to choose the number of principal components in Discussion). Figure 1 summarizes the test statistic  $W = \sqrt{n}M(h)$  corresponding to different numbers of genetic markers under three population models. Under discrete population model and when population stratification is strong ( $P_1:P_2:P_3:P_4 = 1:4:6:8$ ), the Kolmogorov test shows that the population stratification can not be well controlled only using the first principal component and can be well controlled by using the first two principal components (Figure 1(a) and (b)). For almost of all the cases under admixture model A and B, Kolmogorov test shows that the population stratification can be well controlled us-

ing the first principal component (Figure 1(c) and (d)). This observation is consistent with the results given in Figure 2, 3 and 4. where we compare the type-I error rates of various statistical tests using different numbers of independent markers. These results suggest that the Kolmogorov test described above has good utility in the determination of whether a set of genomic markers can control for population stratification and in choosing the number of principal components for a given set of independent genetic markers. Under discrete population models, we include the type-I error results using the first principal component and the first two principal components in order to evaluate the performance of the Kolmogorov test. The final results of QualSPT are obtained using the first principal component for the case of  $P_1:P_2:P_3:P_4 = 1:2:3:4$  and the first two principal components for the case of  $P_1:P_2:P_3:P_4 = 1:4:6:8$ . In the following discussion, we only compare the final results of QualSPT with the results of other tests.

*Type-I error rates:* It is well known that TDT is robust to population stratification. For our type-I error evaluations, we only consider the four tests,  $\chi^2$ , QualSPT, STRAT and GC, which are based on unrelated samples. Figures 2, 3 and 4 summarize the type-I error rates for the four test statistics by using different numbers of markers in simulations through the discrete population models, admixture population model (A) and (B), respectively. The results are based on 1000 replications (10 replications for each of 100 markers, as if there were  $10 \times 100 = 1000$  replications) with each replication consisting of 100 diseased individuals and 100 normal individuals for all four tests. A total of 1000 simulated data sets are used for each sample in the estimation of the p-values. Therefore, for the statistical significance level of 0.05, the standard error for the type-I error rate estimate is  $\sqrt{0.05 \times 0.95/1000} \approx 0.007$  and the 95% confidence interval of the type-I error is (0.036, 0.064). It is apparent from the figures that the  $\chi^2$  test, which ignores potential population stratification, may have type-I error rate that is substantially higher than the nominal level in the presence of population stratification (in all the cases considered here). Under the discrete population models (Figure 2), the type-I errors of both QualSPT (using the first principal component for the case of  $P_1:P_2:P_3:P_4 = 1:2:3:4$  and using the first two principal components for the case  $P_1:P_2:P_3:$

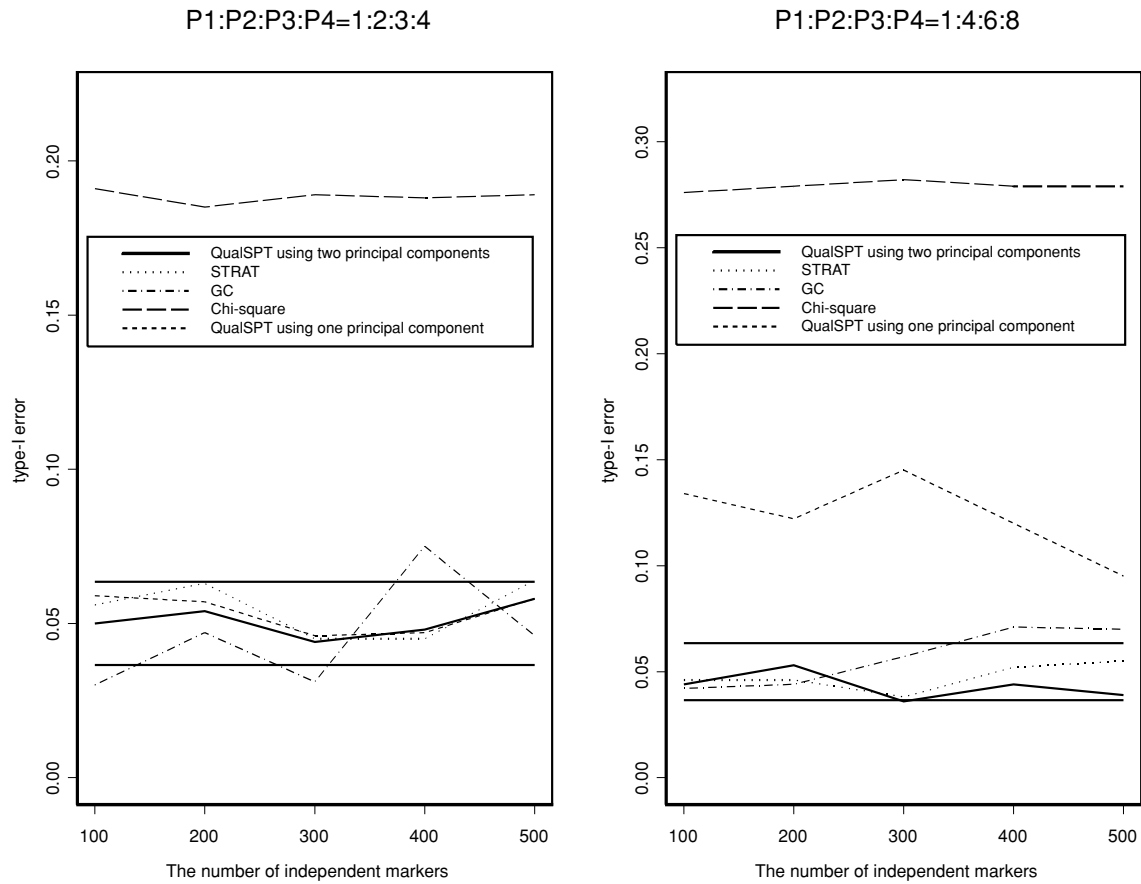


**Figure 1** The number of independent markers *versus* the value of the test statistic  $W = \sqrt{n}M(h)$  under the null hypothesis of no association. The dashed line is the critical value for the significance level of 5%. In figures (a) and (b), the solid line and the dotted line denote the results of the cases  $P_1:P_2:P_3:P_4 = 1:2:3:4$  and  $P_1:P_2:P_3:P_4 = 1:4:6:8$ , respectively. In figure (d), the solid line and the dotted line denote the results of  $\alpha = 2$  and  $\alpha = 4$ , respectively, under admixture population model B.

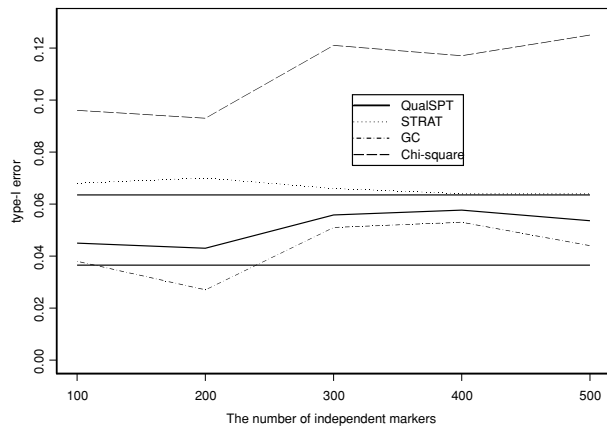
$P_4 = 1:4:6:8$ ) and STRAT are within the 95% confidence interval of the nominal type-I error rate using independent markers from 100 to 500. Although the type-I error rates of the GC are much smaller than that of  $\chi^2$  test, there are several cases in which the type-I error rates of GC test are beyond the boundaries of the 95% confidence interval and these cases seem independent of the number of the markers used to control for population stratification. Population stratification under the admixture model A (Figure 3) is not as strong as other models: the type-I error rates of  $\chi^2$  test are from 10% to 12% for the nominal level 5%. Under this model, the type-I error rate of QualSPT and GC tests (except using 200 markers) are within the 95% confidence interval. The type-I error rate of STRAT is slightly higher than the upper boundary of 95% confidence interval even as

many as 500 markers are used. The population stratification under admixture model B is much stronger than that under admixture population model A. The type-I error rates of  $\chi^2$  test are around 24% and 40% for  $\alpha = 2$  and  $\alpha = 4$ , respectively. Under this model, the type-I errors of QualSPT are within the 95% confidence interval of the nominal type-I error rate using 200 independent markers or more; except a few cases, the type-I errors of GC test are also within the 95% confidence interval of the nominal type-I error rate; though the type-I error rate of STRAT decreases with the increasing of the number of markers used to control for population stratification, this error rate seems to be stable at 7 ~ 8% using 400 independent markers or more.

In summary, the type-I errors of QualSPT are within the 95% confidence interval of the nominal type-I error



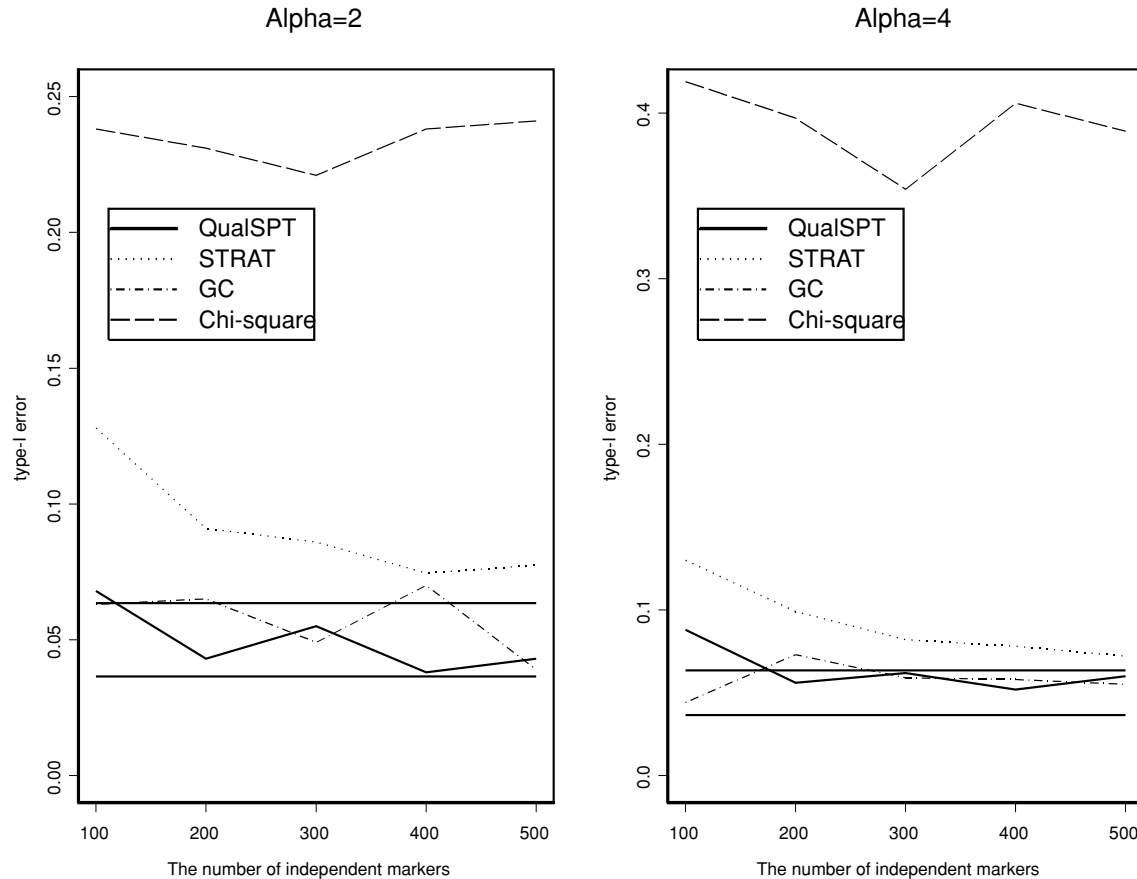
**Figure 2** Type-I error comparisons of the four tests at the nominal value of 5% under discrete population models. The sample consists of 100 diseased individuals and 100 normal individuals. The two straight solid lines around 0.05 (type-I error) form the 95% confidence interval of the type-I error rate under the null hypothesis.



**Figure 3** Type-I error comparisons of the four tests at the nominal value of 5% under admixture population model A. The sample size is 100 diseased individuals and 100 normal individuals. The two straight solid lines around 0.05 (type-I error) form the 95% confidence interval of the type-I error rate under the null hypothesis.

rate using 200 markers or more for all the cases we considered. The type-I errors of GC are around the nominal level, though there are several cases in which the type-I error rates of GC test are beyond the boundaries of the 95% confidence interval. Under discrete population models, the type-I errors of STRAT are within the 95% confidence interval of the nominal type-I error rate using 100 independent markers or more. Under admixture models, the type-I error rate of STRAT decrease with the number of independent markers used to control for population stratification. However, when the number of independent markers is more than 300, type-I error rates of STRAT decrease very slowly and are still beyond the upper boundary of the 95% confidence interval even using 500 markers.

*Power Comparisons:* In this set of simulations, we compare the power of the four tests, QualSPT, STRAT, GC



**Figure 4** Type-I error comparisons of the four tests at the nominal value of 5% under admixture population model B. The sample size is 100 diseased individuals and 100 normal individuals. The two straight solid lines around 0.05 (type-I error) form the 95% confidence interval of the type-I error rate under the null hypothesis.

and TDT. The results are based on 1000 replications with each replication consisting of  $n = 100$  diseased individuals and 100 normal individuals for QualSPT, STRAT and GC, and  $2n/3$  and  $n$  triads for TDT test. A total of 300 markers are used by QualSPT, STRAT and GC to control for population stratification. For almost all the cases including different population models and different disease models we considered, QualSPT is more powerful than TDT when TDT uses  $2n/3$  triads, and QualSPT is less powerful than TDT when TDT uses  $n$  triads. The power comparisons of the three tests, QualSPT, STRAT and GC, are more complicated.

The results of our power comparisons under discrete population model and the assumption of the same high risk allele in all the subpopulations are summarized in Table 1. For almost all the cases in this set of simulations, QualSPT is more powerful than STRAT. The power of

GC is substantially lower than that of both QualSPT and STRAT, especially when population stratification becomes stronger ( $P_1:P_2:P_3:P_4 = 1:4:6:8$ ).

Under admixture population model A and the assumption of the same high risk allele in two ancestral populations, the results of power comparisons are given in Table 2. Under this population model, the population stratification is not strong. The power of the three tests, QualSPT, STRAT and GC, are similar.

The results of power comparisons under admixture population model B and the assumption of the same high risk allele in two ancestral populations are summarized in Table 3. The results show that if the relative risk is high ( $R_{A_1A_1} = 2$  and  $R_{A_2A_2} = 4$ ), QualSPT is more powerful than STRAT. If the relative risk is low ( $R_{A_1A_1} = 1$  and  $R_{A_2A_2} = 2$ ), QualSPT and STRAT have similar power. The power of GC is much less than that of both QualSPT and STRAT, especially for the

**Table 1** Power comparisons of the four tests under discrete subpopulation models and the assumption of same high risk allele in all the subpopulations. The sample size is  $n = 100$  diseased and 100 normal individuals for QualSPT, SRAT, and GC. The sample size is  $2n/3$  and  $n$  triads for TDT.  $P_i$  is the relative disease prevalence of the  $i$ th subpopulation and  $R_{A_i}$  is the relative risk of genotype  $AA$  in the  $i$ th subpopulation ( $i = 1, 2, 3, 4$ )

$P_1:P_2:P_3:P_4$ $R_{A_1}, R_{A_2}, R_{A_3}, R_{A_4}$	Disease Model	$P = 0.05$				$P = 0.01$			
		QualSPT	STRAT	GC	$\frac{TDT}{\frac{2n}{3}}$ $n$	QualSPT	STRAT	GC	$\frac{TDT}{\frac{2n}{3}}$ $n$
1:2:3:4 1,2,3,4	Domi.	0.42	0.33	0.21	0.34 0.48	0.24	0.15	0.06	0.15 0.27
	Add.	0.28	0.30	0.28	0.32 0.46	0.10	0.11	0.10	0.13 0.23
	Rec.	0.35	0.30	0.23	0.34 0.43	0.18	0.16	0.10	0.17 0.26
2,4,6,8	Domi.	0.86	0.76	0.58	0.73 0.84	0.75	0.60	0.30	0.52 0.72
	Add.	0.79	0.65	0.59	0.74 0.87	0.59	0.45	0.31	0.52 0.73
	Rec.	0.78	0.71	0.59	0.71 0.80	0.65	0.57	0.42	0.53 0.70
1:4:6:8 1,4,6,8	Domi.	0.88	0.64	0.52	0.70 0.83	0.77	0.46	0.27	0.50 0.70
	Add.	0.77	0.70	0.47	0.72 0.87	0.61	0.53	0.24	0.52 0.75
	Rec.	0.80	0.73	0.51	0.72 0.81	0.68	0.60	0.34	0.56 0.69
2,8,12,16	Domi.	0.97	0.80	0.78	0.84 0.90	0.94	0.65	0.61	0.71 0.84
	Add.	0.95	0.90	0.65	0.90 0.96	0.89	0.79	0.36	0.80 0.91
	Rec.	0.96	0.90	0.78	0.90 0.94	0.90	0.83	0.68	0.84 0.90

**Table 2** Power comparisons of four tests under admixture population model A and the assumption of the same high risk allele in the two ancestral populations. The sample size is  $n = 100$  diseased and 100 normal individuals for QualSPT, SRAT, and GC. The sample size is  $2n/3$  and  $n$  triads for TDT.  $R_{A_i}, A_i$  is the relative risk of genotype  $AA$  in the  $i$ th ancestral population ( $i = 1, 2$ )

$R_{A_1}, A_1, R_{A_2}, A_2$	Disease Model	$P = 0.05$				$P = 0.01$			
		QualSPT	STRAT	GC	$\frac{TDT}{\frac{2n}{3}}$ $n$	QualSPT	STRAT	GC	$\frac{TDT}{\frac{2n}{3}}$ $n$
2,2	Domi.	0.66	0.67	0.63	0.57 0.72	0.44	0.43	0.40	0.36 0.51
	Add.	0.60	0.65	0.64	0.55 0.70	0.39	0.44	0.41	0.35 0.48
	Rec.	0.58	0.60	0.61	0.52 0.63	0.34	0.39	0.37	0.33 0.44
4,4	Domi.	0.93	0.92	0.89	0.88 0.95	0.88	0.83	0.74	0.77 0.89
	Add.	0.93	0.92	0.79	0.90 0.96	0.85	0.84	0.59	0.76 0.90
	Rec.	0.93	0.92	0.80	0.89 0.94	0.83	0.82	0.64	0.76 0.89
2,4	Domi.	0.87	0.86	0.81	0.82 0.91	0.75	0.73	0.69	0.64 0.80
	Add.	0.87	0.89	0.83	0.82 0.90	0.74	0.75	0.71	0.67 0.81
	Rec.	0.84	0.84	0.74	0.80 0.88	0.71	0.72	0.59	0.63 0.79
4,8	Domi.	0.96	0.94	0.95	0.94 0.97	0.94	0.90	0.90	0.87 0.94
	Add.	0.97	0.95	0.93	0.95 0.98	0.94	0.91	0.87	0.89 0.95
	Rec.	0.96	0.95	0.94	0.94 0.98	0.92	0.92	0.91	0.88 0.95

case of strong population stratification i.e.  $\alpha = 4$ . For significant level  $P = 0.01$ , the power of QualSPT is 2 to 3 times as that of GC for  $\alpha = 2$ , and the power of QualSPT is 4 to 5 times as that of GC for  $\alpha = 4$ .

The results of power comparisons under assumption of random high risk allele are summarized in Table 4. In this set of simulations, STRAT is slightly more powerful than QualSPT and both STRAT and QualSPT are much more powerful than GC.

## Discussion

Recently, several statistical methods have been proposed to use genomic markers to control for population stratification in the analysis of population-based data. These approaches are promising because they may both have greater power than family-based association designs and they may be robust against potential population stratification. As described in the introduction, these general

**Table 3** Power comparisons of four tests under admixture population model B and the assumption of the same high risk allele in the two ancestral populations. The sample size is  $n = 100$  diseased and 100 normal individuals for QualSPT, SRAT, and GC. The sample size is  $2n/3$  and  $n$  triads for TDT.  $R_{A_i, A_i}$  is the relative risk of genotype  $AA$  in the  $i$ th ancestral population ( $i = 1, 2$ ).  $\alpha$  is the parameter in the Gamma distribution as described in the text

$\alpha$	Disease Model	$P = 0.05$				$P = 0.01$			
		QualSPT	STRAT	GC	$\frac{2n}{3}$ TDT $n$	QualSPT	STRAT	GC	$\frac{2n}{3}$ TDT $n$
2									
1,2	Dom.	0.55	0.57	0.29	0.55 0.70	0.33	0.33	0.11	0.33 0.50
	Add.	0.56	0.55	0.23	0.50 0.68	0.33	0.33	0.07	0.27 0.44
	Rec.	0.56	0.58	0.34	0.49 0.64	0.33	0.37	0.13	0.29 0.45
2,4	Dom.	0.94	0.89	0.71	0.87 0.95	0.85	0.76	0.45	0.74 0.89
	Add.	0.92	0.88	0.50	0.86 0.94	0.79	0.75	0.20	0.70 0.88
	Rec.	0.89	0.87	0.62	0.85 0.92	0.76	0.74	0.31	0.70 0.85
4									
1,2	Dom.	0.38	0.37	0.11	0.40 0.60	0.16	0.17	0.02	0.17 0.30
	Add.	0.30	0.31	0.10	0.37 0.54	0.14	0.14	0.02	0.16 0.32
	Rec.	0.28	0.30	0.10	0.32 0.48	0.14	0.15	0.03	0.17 0.30
2,4	Dom.	0.82	0.74	0.36	0.81 0.90	0.63	0.53	0.15	0.60 0.80
	Add.	0.80	0.75	0.36	0.76 0.90	0.57	0.54	0.14	0.59 0.76
	Rec.	0.74	0.70	0.27	0.75 0.90	0.48	0.45	0.12	0.53 0.68

**Table 4** Power comparisons of four tests under admixture population model B and the assumption of random high risk alleles in the two ancestral populations. The sample size is  $n = 100$  diseased and 100 normal individuals for QualSPT, SRAT, and GC. The sample size is  $2n/3$  and  $n$  triads for TDT.  $R_{A_i, A_i}$  is the relative risk of genotype  $AA$  in the  $i$ th ancestral population ( $i = 1, 2$ ). Here, allele  $A$  denotes the high risk allele which may be different in different ancestral populations

$R_{A_1, A_1}, R_{A_2, A_2}$	Disease Model	$P = 0.05$				$P = 0.01$			
		QualSPT	STRAT	GC	$\frac{2n}{3}$ TDT $n$	QualSPT	STRAT	GC	$\frac{2n}{3}$ TDT $n$
1,2	Dom.	0.55	0.71	0.19	0.31 0.41	0.33	0.58	0.05	0.14 0.25
	Add.	0.56	0.51	0.10	0.30 0.41	0.33	0.33	0.04	0.13 0.21
	Rec.	0.56	0.42	0.12	0.26 0.34	0.33	0.23	0.03	0.13 0.18
2,4	Dom.	0.94	0.74	0.30	0.40 0.60	0.85	0.63	0.14	0.23 0.41
	Add.	0.92	0.66	0.23	0.50 0.63	0.79	0.48	0.10	0.31 0.43
	Rec.	0.89	0.89	0.33	0.59 0.69	0.76	0.76	0.16	0.39 0.52

methods can be broadly divided into two classes. The first class, the GC method, may lose power when population stratification is strong. The second class, the SA method, may have difficulty in estimating the population structure under admixture population and the inaccurate estimate of the population structure will affect the validity of the test.

More recently, Zhu *et al.* (2002) proposed a mixture model using principal components to test association between a marker and a qualitative trait. Zhang *et al.* (2003) further developed a semi-parametric test of association based on partial linear model to detect associations between candidate markers and quantitative traits using population-based data. They have shown through

simulations that the test has correct type-I error rate under both discrete subpopulation models and admixture population models.

In this article, using the idea similar to that of Zhang *et al.* (2003), we have developed a semi-parametric test of association, QualSPT, based on a logistic partial linear model to detect the association between a candidate marker and a qualitative trait using the case-control design. We have compared the performance of our test with that of the family-based TDT method, the GC method and STRAT, a SA method proposed by Pritchard *et al.* (2002b). Our simulation results show that QualSPT has correct type-I error rate under both discrete subpopulation models and admixture

population models. QualSPT is more powerful than family-based TDT test when the two tests use the same sample size (sample number of individuals). QualSPT is much more powerful than GC when population stratification is strong. Compared with STRAT, QualSPT has correct type-I error rate using 200 or more independent marker to control for population stratification; whereas, under admixture population models and the assumption that the number of ancestral populations is known, STRAT may still show an excess of false-positive results using 500 independent markers to control for population stratification. For most cases, QualSPT is also more powerful than STRAT when all subpopulations have the same high risk allele. Moreover, it is straightforward to include covariates in QualSPT. If different subpopulations may have different high risk alleles, STRAT is more powerful than QualSPT. However, as argued by Bourgain *et al.* (2000), there may be different high risk alleles for different individuals even in the same subpopulation. If this is the case, all the tests considered in this article will lose power. In this case, the methods based on multi-marker haplotypes may be more appropriate and powerful. The methods based on multi-marker haplotypes need further investigations.

Although we have compared the power of QualSPT with that of TDT with two different sample sizes (the same number of total individuals used and same number of diseased individuals used), the comparisons are based on the assumption that a set of independent markers are available to estimate the genetic background variables. If there is only one candidate locus, QualSPT may require substantially more genotyping efforts. However, given the low prior probability for a specific gene to be involved for a complex trait and the ever-decreasing genotyping cost, it may be more cost effective to perform a population-based study.

Another question is how many principal components we should use for a given data set. Generally speaking, more principal components will have more information. However, the first principal component will summarize most of the data information followed by the second, the third, and other principal component. After the first few principal components, additional principal components have little information but make the model more complex and bring more uncertainty (need to estimate more parameters). Our suggestion is that we first use the first

principal component. If the Kolmogorov test indicates that it can not control for the population stratification, we will use the first two or three principal components. In general, for most of the population genetic data we have analyzed, the first three principal components account for the majority of the genetic variations observed in the data.

## Appendix

### A.1. The expectation of $\beta$ under Model 1

For simplicity, we consider the case that the population under study consists of two subpopulations and the candidate locus is biallelic with allele  $A$  and  $a$ . The numerical codes  $x = 1, 0$  and  $-1$  represent genotypes  $AA, Aa$  and  $aa$ , respectively. This coding scheme is equivalent to the first scheme as described in the section "Genotype Scoring". Let  $y_{ij}$  and  $x_{ij}$  denote the disease status and numerical code of the genotype of the  $j$ th individuals in the  $i$ th subpopulation, respectively. If different subpopulations have different logistic phenotype means, the model is given by

$$\log \frac{p(y_{ij} = 1)}{1 - p(y_{ij} = 1)} = \mu_i + x_{ij}\beta \quad (\text{A1})$$

for  $j = 1, \dots, n_i, i = 1, 2$ . Suppose we ignore the population structure, by assuming the following model

$$\log \frac{p(y_{ij} = 1)}{1 - p(y_{ij} = 1)} = \mu + x_{ij}\beta \quad (\text{A2})$$

for  $j = 1, \dots, n_i, i = 1, 2$ . If we estimate the parameters under Model (A2), the maximum likelihood estimator of  $\mu$  and  $\beta$  are obtained by maximizing the log-likelihood function

$$\begin{aligned} \log L(\mu, \beta) &= \sum_{i=1}^2 \sum_{j=1}^{n_i} [y_{ij} \log P(y_{ij} = 1) \\ &\quad + (1 - y_{ij}) \log(1 - P(y_{ij} = 1))] \\ &= \sum_{i=1}^2 \sum_{j=1}^{n_i} [y_{ij}(\mu + x_{ij}\beta) \\ &\quad - \log(1 + e^{\mu + x_{ij}\beta})]. \end{aligned}$$

Therefore,  $\hat{\mu}$  and  $\hat{\beta}$  satisfy

$$\frac{\partial \log L(\hat{\mu}, \hat{\beta})}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \log L(\hat{\mu}, \hat{\beta})}{\partial \beta} = 0.$$

We can prove that, through some calculation, if  $E(\hat{\beta}) \rightarrow \beta$  as  $n = n_1 + n_2 \rightarrow \infty$ , then either  $\mu_1 = \mu_2$  or  $p_1 = p_2$ , where  $p_1$  and  $p_2$  are the frequencies of allele  $A$  in first subpopulation and second subpopulation, respectively. Thus, we have the following:

**Proposition 1** *Under structured population, i.e. model (A1), if  $\mu_1 \neq \mu_2$  and  $p_1 - p_2 \neq 0$ , then  $E(\hat{\beta}) \neq \beta$  as  $n = n_1 + n_2 \rightarrow \infty$ .*

The proposition implies that  $E(\hat{\beta}) \neq \beta$  even when the sample size becomes large unless the two populations have the same logistic phenotype mean  $\mu$  or the same allele frequency.

## A.2. Asymptotic unbiasedness of $\hat{\beta}$

The estimate  $\hat{\beta}$  for model (2) is asymptotically unbiased if the bandwidth  $h$  in the kernel is of order  $O(n^{-\alpha})$  with  $0 < \alpha < 1/4$ . The sketch of the proof is as follows.

Using the notation in the text and denotes the log-likelihood function by

$$L(\beta, \mu) = \sum_{i=1}^n \{ y_i(x_i\beta + \mu(t_i)) - \log[1 + \exp(x_i\beta + \mu(t_i))] \}.$$

Based on the local log-likelihood (3), if  $\mu(t)$  is a smooth function, i.e., the derivative  $\mu'(t)$  is continuous, and the bandwidth  $h$  satisfies  $h = O(n^{-\alpha})$  with  $0 < \alpha < 1/4$ , then there exists a function  $\mu_\beta(t)$  such that  $\hat{\mu}_\beta(t_i) \rightarrow \mu_\beta(t_i)$  in probability (Severini & Wong, 1992). Through some standard arguments from profile likelihood estimation, we can show that  $\sqrt{n}(\hat{\beta} - \beta_0)$  has a limiting normal distribution  $N(0, [i(\beta)]^{-1})$ , where,

$$i(\beta) = - \frac{1}{\sqrt{n}} \frac{d^2 L(\beta, \mu_\beta)}{d\beta^2} \Big|_{\beta=\beta_0},$$

and that  $\hat{\beta}$  is asymptotically unbiased.

## Acknowledgments

We thank Dr. Kenneth K. Kidd for our access to the ALFRED population genetics database. This work was

supported in part by grant GM59507 from the National Institutes of Health and NNSF of China No 10071011.

## References

- Bacanu, S. A., Devlin, B. & Roeder, K. (2000) The power of genomic control. *Am J Hum Genet* **66**, 1933–1944.
- Bacanu, S. A., Devlin, B. & Roeder, K. (2002) Association studies for quantitative traits in structured populations. *Genet Epidemiol* **22**, 78–93.
- Bourgain, C., Genin, E. & Quesneville, H. (2000) Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet* **64**, 255–265.
- 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 Pop Biol* **60**, 155–166.
- Hart JD (1997) *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.
- Nguyen, H. T. & Rogers, G. S. (1989) *Fundamentals of Mathematical Statistics: vol. II: Statistical Inference*. Springer-Verlag, New York.
- Osier, M. V., Cheung, K.-H., Kidd, J. R., Pakstis, A. J., Miller, P. L. & Kidd, K. K. (2001) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms: an update. *Nucl Acids Res* **29**, 317–319.
- Pritchard, J. K. & Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**, 220–228.
- Pritchard, J. K. & Donnelly, P. (2001) Case-control studies of association in structured or admixed populations. *Theor Pop Biol* **60**, 227–237.
- Pritchard, J. K., Stephens, M. & Donnelly, P. (2000a) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.
- Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. (2000b) Association mapping in structured population. *Am J Hum Genet* **67**, 170–181.
- Reich, D. E. & Goldstein, D. B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* **20**, 4–16.
- Satten, G. A., Flanders, W. D. & 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**, 466–477.
- Severini, T. A. & Staniswalis, J. G. (1994) Quasi-likelihood estimation in semiparametric models. *J Amer Statist Assoc* **89**, 501–511.
- Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. Springer, New York.
- Severini, T. A. & Wong, W. (1992) Profile likelihood and conditionally parametric models. *Ann Statist* **20**, 1768–1802.

- Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506–513.
- Zhang, S. L. & Zhao, H. Y. (2001) Quantitative similarity-based association tests using population samples. *Am J Hum Genet* **69**, 601–614.
- Zhang, S. L., Kidd, K. K. & Zhao, H. Y. (2002) Detecting genetic association in case-control studies using similarity-based association tests. *Statistica Sinica* **12**, 337–359.
- Zhang, S. L., Zhu, X. & Zhao, H. Y. (2003) On a semi-parametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol*. In press.
- Zhu, X., Zhang, S. L., Zhao, H. Y. & Cooper, R. S. (2002) Association mapping using a mixture model for complex traits. *Genet Epidemiol* **23**, 181–196.

*Received:* 7 August 2002

*Accepted:* 6 December 2002