Quantitative Similarity-Based Association Tests Using Population Samples

Shuanglin Zhang and Hongyu Zhao

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

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. Furthermore, it is often more feasible and less expensive to collect unrelated individuals. Recently, several statistical methods have been proposed for case-control association tests in a structured population; these methods may be robust to population stratification. In the present study, we propose a quantitative similarity-based association test (QSAT) to identify association between a candidate marker and a quantitative trait of interest, through use of unrelated individuals. For the QSAT, we first determine whether two individuals are from the same subpopulation or from different subpopulations, using genotype data at a set of independent markers. We then perform an association test between the candidate marker and the quantitative trait, through incorporation of such information. Simulation results based on either coalescent models or empirical population genetics data show that the QSAT has a correct type I error rate in the presence of population stratification and that the power of the QSAT is higher than that of family-based association designs.

Introduction

Population-based association studies using unrelated individuals have often been criticized for inducing spurious associations due to population stratification. As a result, family-based association designs (Spielman et al. 1993) have received great attention recently, because of their robustness to population stratification and their potentially higher power relative to linkage studies (Risch and Merikangas 1996). Population samples consisting of unrelated individuals, however, may be easier and less expensive to collect, and such designs are, in general, more powerful than family-based association designs, both for qualitative traits (Morton and Collins 1998; Risch and Teng 1998; Teng and Risch 1999; Risch 2000) and for quantitative traits (van den Oord 1999). Recently, several methods have been proposed that utilize genomic markers to control for population stratification in the analysis of unrelated individuals (Devlin and Roeder 1999; Bacanu et al. 2000; Pritchard et al. 2000*b;* Reich and Goldstein 2001; Satten et al. 2001; Zhang et al., in press). These novel approaches are promising because they may have greater power than family-based association designs and may be robust to potential population stratification. One limitation of these methods is that they are only applicable to qualitative traits, although quantitative traits may contain more information.

In the present study, we develop a quantitative similarity-based association test (QSAT) to examine associations between candidate markers and quantitative traits of interest, in a set of unrelated individuals. The QSAT controls population stratification through a set of genomic markers. To perform the QSAT, we first use the genotypes of the sampled individuals at a series of independent markers to calculate a similarity score, S_{ij}, between individuals i and j. We then model the distribution of these similarities, through use of a normal mixture model with one or two components (a within-subpopulation component and a between-subpopulation component). We then use the Bayesian information criterion to estimate the number of components and decompose each individual's genotypic score into within-subpopulation and between-subpopulation components. The QSAT is then calculated on the basis of a regression model that treats the trait value as the dependent variable and the within- and between-population genotypic scores as predictors. We evaluate the performance of the QSAT through simulations using coalescent models and empirical population genetics data. 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 (Fulker et al. 1999; Monks and Kaplan 2000; Sun et al. 2000).

Received May 16, 2001; accepted for publication June 27, 2001; electronically published July 30, 2001.

Address for correspondence and reprints: Dr. Hongyu Zhao, Department of Epidemiology and Public Health, 60 College Street, Yale University School of Medicine, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

© 2001 by The American Society of Human Genetics. All rights reserved.

0002-9297/2001/6903-0013$02.00

Methods

In this section, we first discuss the method for a homogeneous population and then discuss the QSAT for a heterogeneous population. We assume that the candidate marker is biallelic, with 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 at the candidate marker, with the value of *A* being 1, 0, and −1 for genotypes *MM, Mm,* and *mm,* respectively. We use *D* to denote the dominance genotypic score at the candidate marker, with the value of *D* being 0, 1, and 0 for genotypes *MM, Mm,* and *mm,* respectively. Let y_i denote the quantitative trait value of the *i*th individual. For a homogeneous population, genetic association between the candidate marker and the quantitative trait can be studied through the following regression model:

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

where the values of 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. In the case of a homogeneous population, the least-squares (LS) estimators of α and β, denoted by $\hat{\alpha}$ and $\hat{\beta}$, respectively, are unbiased estimators of α and β. Under the null hypothesis of no association between the candidate marker and the trait of interest, both α and β are 0, and standard statistical tests can be performed to identify deviation from the null hypothesis.

The regression method shown in equation (1) may be invalid in the presence of population stratification. To illustrate this point, let us assume that there are *k* subpopulations, with n_i individuals sampled from the *i*th subpopulation, and that each subpopulation is homogeneous. Let μ_i denote the phenotype mean in the *i*th subpopulation, let p_i and q_i denote the allele frequencies of the *M* and *m* alleles in the *i*th population, let y_{ij} denote the trait value of the *j*th individual in the *i*th subpopulation, and let A_{ij} and D_{ij} denote the additive and dominance genotypic scores of the *j*th individual in the *i*th subpopulation. We assume that the conditional expectation of the trait value of the *j*th individual in the *i*th subpopulation is

$$E(y_{ij}|A_{ij},D_{ij}) = \mu_i + \alpha_i A_{ij} + \beta_i D_{ij} \ . \qquad (2)$$

In the presence of subpopulations, the null hypothesis to be tested is that there is no association between the candidate marker and the trait value in any of the subpopulations—that is, $\alpha_1 = \ldots = \alpha_k = 0$ and $\beta_1 = \ldots = \beta_k = 0$.

If we apply the following regression model to test the null hypothesis of no association between the candidate marker and the trait,

$$y_{ij} = \mu + \alpha A_{ij} + \beta D_{ij} + e_{ij} \ , \qquad (3)$$

the conditional expectations of regression coefficients $\hat{\alpha}$ and $\hat{\beta}$, conditional on the observed values of A_{ij} and D_{ij}, are

$$E(\hat{\alpha}|A_{ij},D_{ij},\text{for all } i,j) = \mu_\alpha + \sum_{i=1}^{k} \alpha_i a_{(\alpha)i} + \sum_{i=1}^{k} \beta_i d_{(\alpha)i}$$

and

$$E(\hat{\beta}|A_{ij},D_{ij},\text{for all } i,j) = \mu_\beta + \sum_{i=1}^{k} \alpha_i a_{(\beta)i} + \sum_{i=1}^{k} \beta_i d_{(\beta)i} \ ,$$

where the notation is given in detail in Appendix A, with $\sum_{i=1}^{k} a_{(\alpha)i} = 1$, $\sum_{i=1}^{k} d_{(\alpha)i} = 0$, $\sum_{i=1}^{k} a_{(\beta)i} = 0$, and $\sum_{i=1}^{k} d_{(\beta)i} = 1$. Under the null hypothesis of no association between the candidate marker and the trait of interest, $\alpha_1 = \ldots = \alpha_k = 0$ and $\beta_1 = \ldots = \beta_k = 0$, $E(\hat{\alpha}|A_{ij},D_{ij}) = \mu_\alpha$, and $E(\hat{\beta}|A_{ij},D_{ij}) = \mu_\beta$. Therefore, $E(\hat{\alpha}) = E(\mu_\alpha)$ and $E(\hat{\beta}) = E(\mu_\beta)$, under the null hypothesis; however, $E(\mu_\alpha)$ and $E(\mu_\beta)$ may not be 0, in general, when allele frequencies and mean trait values differ among the subpopulations. Therefore, in the presence of population stratification, even under the null hypothesis of no association between the candidate marker and the trait of interest, statistical tests based on the model in equation (3) may lead to false positives due to population stratification.

In the context of analyzing sib-pair data, Fulker et al. (1999) proposed to decompose the genotypic score into two orthogonal components: the between-family (b) component and the within-family (w) component. Under this decomposition, only the between-family component is sensitive to population structure, and the within-family component is significant only when there is an association between the candidate marker and the trait. This approach has been extended to nuclear families (Abecasis et al. 2000) and general sibship data (Sham et al. 2000). To generalize this idea to population data in cases in which the exact population structure is known, we can decompose the genotypic scores into orthogonal between-population and within-population components. Specifically, we define $\overline{A}_i = \sum_{j=1}^{n_i} A_{ij}/n_i$ and $A_{\mathrm{w}ij} = A_{ij} - \overline{A}_i$ to be between-population and within-population additive genotypic scores, respectively, and define $\overline{D}_i = \sum_{j=1}^{n_i} D_{ij}/n_i$ and $D_{\mathrm{w}ij} = D_{ij} - \overline{D}_i$ to be between-population and within-population dominance genotypic scores, respectively. Having defined the notation, we consider the following regression model:

$$y_{ij} = \mu + \alpha_{\mathrm{b}}\overline{A}_i + \alpha_{\mathrm{w}} A_{\mathrm{w}ij} + \beta_{\mathrm{b}}\overline{D}_i + \beta_{\mathrm{w}} D_{\mathrm{w}ij} + e_{ij} \ . \qquad (4)$$

Denote the LS estimators of α_{b}, α_{w}, β_{b}, and β_{w} as $\hat{\alpha}_{\mathrm{b}}$, $\hat{\alpha}_{\mathrm{w}}$, $\hat{\beta}_{\mathrm{b}}$, and $\hat{\beta}_{\mathrm{w}}$, respectively. The conditional expectations of these estimators are derived in Appendix B, and it can be shown that all the spurious association between genotypic scores and trait values due to population stratification is accounted for by $\hat{\alpha}_{\mathrm{b}}$ and $\hat{\beta}_{\mathrm{b}}$. On the other hand, $\hat{\alpha}_{\mathrm{w}}$ and $\hat{\beta}_{\mathrm{w}}$ are unbiased estimates of the additive and dominance genetic values α^* and β^*, provided that all subpopulations have the same additive and dominance genetic values—that is, $\alpha_1 = \ldots = \alpha_k = \alpha^*$ and $\beta_1 = \ldots = \beta_k = \beta^*$. When the additive and dominance values are different among the subpopulations, the expectations of $\hat{\alpha}_{\mathrm{w}}$ and $\hat{\beta}_{\mathrm{w}}$ are

$$E(\hat{\alpha}_{\mathrm{w}}) = \sum_{i=1}^{k} \alpha_i a_{(\mathrm{w}\alpha)i} + \sum_{i=1}^{k} \beta_i d_{(\mathrm{w}\alpha)i}$$

and

$$E(\hat{\beta}_{\mathrm{w}}) = \sum_{i=1}^{k} \alpha_i a_{(\mathrm{w}\beta)i} + \sum_{i=1}^{k} \beta_i d_{(\mathrm{w}\beta)i} \ ,$$

where $\sum_{i=1}^{k} a_{(\mathrm{w}\alpha)i} = \sum_{i=1}^{k} d_{(\mathrm{w}\beta)i} = 1$ and $\sum_{i=1}^{k} a_{(\mathrm{w}\beta)i} = \sum_{i=1}^{k} d_{(\mathrm{w}\alpha)i} = 0$, and the details are given in Appendix B. So, under the null hypothesis of no association, $E(\hat{\alpha}_{\mathrm{w}}) = 0$ and $E(\hat{\beta}_{\mathrm{w}}) = 0$. Intuitively, when all of the subpopulations have the same additive and dominance genetic values (the mean trait values may be different among subpopulations), then $\alpha_1 = \ldots = \alpha_k = \alpha_{\mathrm{w}}$ and $\beta_1 = \ldots = \beta_k = \beta_{\mathrm{w}}$. In this case, testing the hypothesis $H_0{:}\alpha_1 = \ldots = \alpha_k = 0$ and $\beta_1 = \ldots = \beta_k = 0$ is equivalent to testing the hypothesis $H_0^1{:}\alpha_{\mathrm{w}} = \beta_{\mathrm{w}} = 0$ under the model in equation (4). When the additive and dominance genetic values vary among subpopulations, α_{w} and β_{w} are linear combinations of α_i and β_i. In this case, rejection of H_0^1 guarantees rejection of H_0. Therefore, the test for the null hypothesis H_0^1 under the model in equation (4) is still a valid test for hypothesis H_0 in a structured population.

One difficulty in the application of the above approach is that we do not know the underlying population structure. However, potential population structures can be estimated through a series of genetic markers (e.g., see the report by Pritchard et al. [2000*a*]). In the present study, instead of estimating the underlying population structure, we examine each pair of individuals and infer whether the two individuals are from the same subpopulation or from different subpopulations. Suppose that there are L independent biallelic markers $\mathcal{A}_l$, where $l = 1, \ldots, L$, and each marker $\mathcal{A}_l$ has two alleles, A_l and a_l. Further suppose that there are n individuals in our sample and let z_{il} denote the genotype of the ith individual at the lth marker, where $i = 1, \ldots, n$ and $l = 1, \ldots, L$. The value of each z_{il} can be 0, 1, or 2, corresponding to the ith individual having 0, 1, or 2 copies of allele A_l, respectively. A natural measure of the difference in genotypes between the ith and the jth individuals is $d_{ij} = \sum_{l=1}^{L} |z_{il} - z_{jl}|$. In the present study, we define the similarity, S_{ij}, between the ith and the jth individuals as $S_{ij} = d_{\max} - d_{ij}$, where $d_{\max}$ is the maximum value of the d_{ij} across all pairs of individuals.

For individuals within the same subpopulation, we expect the value of S_{ij} to be smaller than that between individuals from different subpopulations. We propose to decompose these similarity estimates into two components: a within-subpopulation component and a between-subpopulation component. To identify possible components among the S_{ij}, we assume the following normal mixture model for the similarity estimates S_{ij}:

$$S_{ij} \sim \sum_{k=1}^{K} p_k N(\mu_k, \sigma_k^2) \ ,$$

where K represents the number of components in the mixture model, p_k denotes the proportion of the kth component, and $N(\mu_k, \sigma_k^2)$ denotes the Gaussian density function with mean μ_k and variance σ_k^2. The maximum-likelihood estimates of the parameters p_k, μ_k, and σ_k, for a given K, can be obtained by means of the clustering expectation-maximization (CEM) method (Celeux and Govaert 1995). We use the Bayesian information criterion (BIC) to choose K. The BIC is defined as $\mathrm{BIC}(K) = -2L(K) + M(K)\log N$, where N is the total number of observations,

$$L(K) = \sum_{i,j} \log\left[\sum_{k=1}^{K} \hat{p}_k N(\hat{\mu}_k, \hat{\sigma}_k^2)\right]$$

is the maximized log likelihood for a given K, and $M(K)$ is the number of free parameters in the mixture model. On the basis of our experience with simulated data sets based on both coalescent models and on empirical population genetics data, a choice for K between 1 and 2 is adequate to account for population structure in the data. The case of $K = 1$ corresponds to a single population—that is, there is no population heterogeneity, whereas $K = 2$ corresponds to two components: a within-population component and a between-population component. Note that $K = 2$ implies only that there is population structure in the data, but it does not imply that there are only two subpopulations. When $K = 2$, let $\hat{p}_k$, $\hat{\mu}_k$, and $\hat{\sigma}_k$ denote the maximum-likelihood estimates of the parameters p_k, μ_k, and σ_k, respectively; then

Table 1

Type I Error Rates of the Four Test Statistics (*T,* QSAT, TDT$_{\text{MK}}$, and TDT$_{\text{VC}}$) under Coalescent Models for Different Trait-Value Distributions

Trait Distribution and No. of Generations since Population Division	Type I Error Rate (%)							
	P = .05				P = .01			
	T	QSAT	TDT$_{\text{MK}}$	TDT$_{\text{VC}}$	T	QSAT	TDT$_{\text{MK}}$	TDT$_{\text{VC}}$
Normal:								
500	38.5	4.3	4.8	4.9	25.1	.8	1.2	1.2
1,500	65.4	4.9	4.7	5.4	54.6	1.0	.95	1.2
4,500	89.3	5.4	4.6	4.3	82.6	1.1	1.0	.95
Log-normal:								
500	39.6	4.3	4.4	4.5	27.0	.85	.88	1.1
1,500	64.1	4.6	5.0	5.6	53.5	.87	1.1	.9
4,500	87.4	4.5	5.6	5.3	81.5	.85	1.2	1.0

$$t_{ijk}=\frac{\hat{p}_kN(\hat{\mu}_k,\hat{\sigma}_k^2)}{\hat{p}_1N(\hat{\mu}_1,\hat{\sigma}_1^2)+\hat{p}_2N(\hat{\mu}_2,\hat{\sigma}_2^2)}$$

is the conditional probability that S_{ij} arises from the kth mixture component. Assuming that $\hat{\mu}_1>\hat{\mu}_2$ if $t_{ij1}>.5$, we define the similarity indicator W_{ij} between the ith and the jth individuals to be 1 and assume that these two individuals belong to the same subpopulation in our subsequent analysis. If $t_{ij1}<.5$, we define the similarity indicator W_{ij} between the ith and the jth individuals to be 0 and assume that these two individuals belong to different subpopulations.

Let y_i, A_i, and D_i denote the trait value, additive genotypic score, and dominance genotypic score, respectively, of the ith individual. Let $n_i=\sum_{j=1}^{n}W_{ij}$, with n_i defined as the number of individuals estimated to be in the same subpopulation as the ith individual. Using A_i and W_{ij}, we can decompose the additive genotypic score, A_i, into two components: a between-subpopulation component, $\overline{A}_i=(\sum_{j=1}^{n}A_jW_{ij})/n_i$, and a within- subpopulation component, $A_{\text{w}i}=A_i-\overline{A}_i$. Similarly, we can decompose the dominance genotypic score, D_i, into two components: a between-subpopulation component $\overline{D}_i=(\sum_{j=1}^{n}D_jW_{ij})/n_i$ and a within-subpopulation component $D_{\text{w}i}=D_i-\overline{D}_i$. On the basis of these definitions, we fit the following regression model:

$$y_i=\mu+\alpha_{\text{b}}\overline{A}_i+\alpha_{\text{w}}A_{\text{w}i}+\beta_{\text{b}}\overline{D}_i+\beta_{\text{w}}D_{\text{w}i}+e_i\ . \quad (5)$$

When there are k subpopulations, and under the assumption that we can make correct inference about whether two individuals are from the same or different subpopulations, the between-subpopulation components and the within-subpopulation components are orthogonal. The LS estimates of α_{w} and β_{w} are

$$\hat{\alpha}_{\text{w}}=\frac{V_{D_{\text{w}}}C_{A_{\text{w}}y}-C_{A_{\text{w}}D_{\text{w}}}C_{D_{\text{w}}y}}{V_{A_{\text{w}}}V_{D_{\text{w}}}-C_{A_{\text{w}}D_{\text{w}}}^2}$$

and

$$\hat{\beta}_{\text{w}}=\frac{V_{A_{\text{w}}}C_{D_{\text{w}}y}-C_{A_{\text{w}}D_{\text{w}}}C_{A_{\text{w}}y}}{V_{A_{\text{w}}}V_{D_{\text{w}}}-C_{A_{\text{w}}D_{\text{w}}}^2}\ ,$$

where

$$V_{A_{\text{w}}}=\sum_{i=1}^{n}A_{\text{w}i}^2\ ,$$

$$V_{D_{\text{w}}}=\sum_{i=1}^{n}D_{\text{w}i}^2\ ,$$

$$C_{A_{\text{w}}y}=\sum_{i=1}^{n}A_{\text{w}i}(y_i-\overline{y})\ ,$$

$$C_{D_{\text{w}}y}=\sum_{i=1}^{n}D_{\text{w}i}(y_i-\overline{y})\ ,$$

and

$$C_{A_{\text{w}}D_{\text{w}}}=\sum_{i=1}^{n}D_{\text{w}i}A_{\text{w}i}\ .$$

To test the null hypothesis that there is no association between the candidate marker and the trait of interest in all subpopulations, we may test the null hypothesis $H_0{:}\alpha_w=\beta_w=0$, through use of the regression model in equation (5). If we assume that e_i are independent normal variables with the same variance, the usual test statistic is the F test statistic, $F=T/\hat{\sigma}^2$, where $T=\hat{\boldsymbol{\eta}}^TV\hat{\boldsymbol{\eta}}/2$, $\hat{\boldsymbol{\eta}}^T=(\hat{\alpha}_{\text{w}},\hat{\beta}_{\text{w}})$,

$$V=\begin{pmatrix}V_{A_{\text{w}}} & C_{A_{\text{w}}D_{\text{w}}}\\ C_{A_{\text{w}}D_{\text{w}}} & V_{D_{\text{w}}}\end{pmatrix}\ ,$$

and $\hat{\sigma}^2$ is an estimate of the variance of the e_i. However, the e_i may not follow the normal distribution and may not have the same variance, especially for different genotypes and in different subpopulations. Therefore, sta-

Table 2

Type I Error Rates of the Four Tests (T, QSAT, $\mathrm{TDT_{MK}}$, and $\mathrm{TDT_{VC}}$) in Simulations based on Empirical Population Genetics Data, under the Random Sampling Scheme

	Type I Error Rate (%)							
No. of Independent Markers, Status of High-Risk Allele, and Trait Distribution	P = .05				P = .01			
	T	QSAT	$\mathrm{TDT_{MK}}$	$\mathrm{TDT_{VC}}$	T	QSAT	$\mathrm{TDT_{MK}}$	$\mathrm{TDT_{VC}}$
520:								
Fixed:								
Normal	13.5	4.8	4.6	4.6	5.1	1.1	1.0	.9
Log-normal	14.6	4.4	4.5	4.4	6.0	.8	1.0	1.0
Random:								
Normal	14.1	4.9	4.7	5.5	5.7	1.1	1.1	1.2
Log-normal	13.6	4.5	4.7	4.6	5.8	.8	.9	1.0
1,040:								
Fixed:								
Normal	13.4	5.1	5.3	5.7	5.3	1.0	1.3	1.2
Log-normal	14.5	4.9	5.1	5.2	6.2	.8	1.1	1.0
Random:								
Normal	13.2	5.1	4.4	4.9	5.0	1.3	.9	1.1
Log-normal	14.5	4.5	5.9	5.0	5.3	.9	1.4	1.2

tistical inferences using the F statistic may not lead to correct statistical significance levels.

In the present study, we propose to use T as our QSAT and to use simulations to evaluate statistical significance for the test statistic. The basic idea of the simulation method is to permute the trait values of the individuals within the same subpopulation, in order to derive an empirical distribution for the test statistic; however, one practical difficulty in implementing this method directly is that we do not know exactly how many subpopulations there are or which individuals belong to the same subpopulation. As a result, we propose the following simulation method to approximate the distribution of the QSAT:

1. Randomly choose one individual—say, the i_1th individual—in the sample. Then randomly choose one individual from the set $\{i\colon W_{i_1 i} = 1\}$—say, the i_1^*th individual. Denote the trait value of the i_1^*th individual as $y_{i_1}^*$;
2. Randomly choose one individual from all sampled individuals except the i_1th individual—say, the i_2th individual. Then randomly choose one individual from the set $I_{i_2} = \{i\colon W_{i_2 i} = 1\}/\{i_1^*\}$—say, the i_2^*th individual. Denote the trait value of the i_2^*th individual by $y_{i_2}^*$. If I_{i_2} is an empty set, define $y_{i_2}^* = y_{i_2}$;
3. Randomly choose one from all the sampled individuals except individuals $i_1, i_2, \ldots, i_{(j-1)}$—say, individual i_j—and randomly choose one individual from the set $I_{i_j} = \{i\colon W_{i_j i} = 1\}/\{i_1^*, i_2^*, \ldots, i_{(j-1)}^*\}$—individual i_j^*, for example. Denote the trait value of i_j^*th individual as $y_{i_j}^*$. If I_{i_j} is an empty set, define $y_{i_j}^* = y_{i_j}$.

In the end, we generate a set of new trait values: $y_1^*, y_2^* \ldots$, and y_n^* for the n individuals in the sample. For this simulated sample, we calculate the test statistic. We repeatedly generate m sets of simulated data sets, and we can then estimate the level of statistical significance from these test statistics.

Simulation Models

In this section, we discuss the simulation models used to assess whether the QSAT is robust to population stratification and to compare the power of the QSAT with other association tests. In our simulation studies, we generate the data either through coalescent models or through empirical population genetics data.

Coalescent Models

In this set of simulations, we use coalescent models to generate genotypes of the sampled individuals in a structured population. Pritchard et al. (2000*b*) considered coalescent models with constant population sizes. We consider coalescent models with variable population sizes (Griffiths and Tavaré 1994, 1997) in our simulations and allow subpopulations to have different population sizes. We assume that there was an ancestral population that had evolved for a long period of time with a constant population size; this population was then divided into two subpopulations, T generations before the present time. From the time of division, the two subpopulations have experienced exponential growth independently, without migrations. We assume that, at the time of division, the population sizes of the two subpopulations were 100 and 10^4, respectively, and that the population sizes at the present time are 10^7 and 5×10^7, respectively. Therefore, the first subpopulation has

Table 3

Type I Error Rates of the Four Tests (T, QSAT, $\mathrm{TDT_{MK}}$, and $\mathrm{TDT_{VC}}$) in Simulations based on Empirical Population Genetics Data, under the Selective Sampling Scheme

	Type I Error Rate (%)							
No. of Independent Markers, Status of High-Risk Allele, and Trait Distribution	$P = .05$				$P = .01$			
	T	QSAT	$\mathrm{TDT_{MK}}$	$\mathrm{TDT_{VC}}$	T	QSAT	$\mathrm{TDT_{MK}}$	$\mathrm{TDT_{VC}}$
520:								
Fixed:								
Normal	22.5	4.5	4.9	5.2	12.1	.80	1.42	1.12
Log-normal	18.9	4.3	4.3	4.2	9.0	.74	.88	.68
Random:								
Normal	22.3	4.5	4.8	5.5	11.7	.81	.92	1.20
Log-normal	19.6	4.5	5.4	5.4	8.40	.98	1.30	1.25
1,040:								
Fixed:								
Normal	22.6	4.6	4.8	5.1	11.7	.84	.91	.98
Log-normal	19.4	4.5	4.6	4.9	9.2	.84	.88	.85
Random:								
Normal	22.0	4.9	5.5	5.3	11.7	.95	1.36	1.22
Log-normal	18.6	4.3	5.8	5.7	9.5	.81	1.25	1.05

experienced more rapid growth than has the second subpopulation. 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 first two separation times probably correspond to the divergence time between non-African populations, and the third separation time probably corresponds to the divergence time between African and non-African populations (Goldstein et al. 1995).

We assume that a total of 500 independent biallelic markers are used for our inference on the population structure. The sample consists of 25 individuals from the first subpopulation and 125 individuals from the second subpopulation. We assume that the mutation rate is $\mu = 5 \times 10^{-7}$ per generation and only select markers with allele frequencies of $\geqslant .2$ in the sample. This threshold was also used by Pritchard and Rosenberg (1999) to approximate the likely characteristics of single-nucleotide polymorphism (SNP) surveys (Wang et al. 1998). We use the same procedure to simulate genotypes at the candidate locus. On the basis of the genotype at the candidate locus, the trait values are generated according to the following model:

$$y_{ij} = \mu_i + \alpha_i A_{ij} + \beta_i D_{ij} + e_{ij} \ , \qquad (6)$$

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 the type I error examination and power comparison, respectively. We also vary genetic models and trait distributions (either normally distributed or log-normally distributed) in our simulations. In the determination of the allele that increases the quantitative trait values, we fix the same allele in the two subpopulations.

Empirical Population Genetics Data

One limitation of the simulations based on coalescent models is that these models may not represent the human population evolutionary histories accurately. Therefore, in our simulations, we also use empirical population genetics data from the population genetics database ALFRED (Osier et al. 2001; ALFRED Web site), which provides allele frequencies for SNPs and for microsatellite markers in different populations. For our simulations, we extracted 130 markers across four populations, including Danes, San Francisco Chinese, Biaka, and Maya. We use these four populations to represent populations from four different continents. For microsatellite markers, we pool the alleles to form biallelic markers with allele frequencies of 10%–90%.

For simulations based on empirical population genetics data, we consider different numbers of markers used to infer pairwise relationships, different trait-value distributions, and different schemes to assign alleles conferring high trait values. We generate 20 replications, with each replication consisting of a total of n individuals. Among these n individuals, there are $.5n$ individuals sampled from the Danes, $.2n$ individuals from the Chinese, $.2n$ individuals from the Biaka, and $.1n$ individuals from the Maya. In the determination of the allele that increases the quantitative trait values, we either fix the same allele in the two subpopulations (denoted as the

Table 4

Power Comparisons of the Three Tests (QSAT, TDT_{MK}, and TDT_{VC}), under Coalescent Models for Different Trait-Value Distributions

Trait Distribution, No. of Generations since Population Division, and Model	Power													
	P = .05							P = .01						
		TDT_{MK}			TDT_{VC}				TDT_{MK}			TDT_{VC}		
	QSAT	n/3	2n/3	n	n/3	2n/3	n	QSAT	n/3	2n/3	n	n/3	2n/3	n
Normal:														
500:														
Dominant	.99	.55	.81	.88	.48	.68	.76	.97	.35	.66	.80	.29	.53	.65
Additive	.99	.55	.86	.95	.50	.70	.85	.99	.35	.75	.90	.25	.55	.74
Recessive	.99	.47	.80	.87	.38	.63	.76	.97	.24	.62	.77	.20	.50	.65
1,500:														
Dominant	.99	.46	.72	.82	.34	.50	.56	.97	.22	.55	.69	.18	.36	.49
Additive	.97	.46	.80	.90	.32	.57	.69	.94	.22	.61	.78	.16	.48	.59
Recessive	.97	.40	.74	.82	.28	.50	.67	.95	.19	.54	.67	.15	.35	.55
4,500:														
Dominant	.97	.35	.66	.76	.23	.45	.53	.92	.16	.44	.59	.12	.35	.43
Additive	.91	.34	.63	.74	.18	.40	.55	.81	.12	.38	.53	.08	.27	.38
Recessive	.90	.24	.52	.64	.16	.34	.46	.82	.10	.30	.46	.07	.22	.33
Log-normal:														
500:														
Dominant	.99	.53	.81	.88	.39	.67	.76	.97	.36	.67	.79	.24	.54	.67
Additive	.99	.63	.89	.97	.47	.74	.86	.98	.40	.77	.91	.31	.61	.76
Recessive	.99	.55	.84	.92	.43	.70	.80	.97	.34	.68	.81	.25	.53	.70
1,500:														
Dominant	.97	.48	.75	.83	.29	.54	.61	.94	.29	.58	.69	.16	.37	.49
Additive	.97	.54	.83	.89	.35	.56	.65	.92	.32	.67	.78	.21	.44	.54
Recessive	.97	.46	.76	.84	.29	.52	.60	.93	.26	.60	.72	.17	.40	.50
4,500:														
Doninant	.96	.38	.67	.79	.19	.42	.48	.88	.22	.48	.63	.09	.26	.38
Additive	.89	.39	.65	.76	.19	.40	.52	.77	.22	.48	.60	.10	.30	.39
Recessive	.87	.33	.57	.68	.17	.35	.45	.76	.16	.40	.52	.08	.24	.35

Note.—Sample size is n = 150 for the QSAT and 50, 100, and 150 for the TDT tests.

"fixed" simulation design in the following discussion) or randomly choose one of the alleles with probability according to allele frequency in each subpopulation (denoted as the "random" simulation design in our following discussion). The trait values are generated according to the model in equation (6) above, with the only difference being that there are four population trait means, $\mu_1, \ldots, \mu_4$, considered in the simulation, where $\mu_i = \mu_{00}$, $\alpha_i = \beta_i = \mu_0 \times R_i$, and the e_{ij} are random variables from a normal distribution or random variables from a log-normal distribution. In the type I error examination, we set $\mu_{00} = 2$ and $\mu_0 = 0$. For power comparisons, for each replication we systematically assign the trait locus to be one of the markers. Therefore, for each replication sample, we generate 130 samples with trait values determined from different markers. We set $\mu_0 = \mu_{00} = 2$, $R_1 = 1/4$ for Danes, $R_2 = 1/3$ for Chinese, $R_3 = 1$ for Biaka, and $R_4 = 1/2$ for Maya. In both type I error assessments and power comparisons, we use 2,000 simulated samples to estimate the P value for each simulated sample.

We choose individuals by two sampling schemes. In the random sampling scheme, we select $n = 150$ individuals from the overall population. In the selective sampling scheme, we first randomly sample 500 individuals from the overall population and then select individuals in the top 10% and bottom 10% of the trait distribution, resulting in a sample size of 100 individuals.

Other Association Tests Considered

In addition to the QSAT, we also consider three other association tests in our simulations. The first test is the test that ignores potential population stratification, and this test statistic is denoted by T in the following discussion. The difference between this test and the QSAT is that, in the T test, we always treat the sampled individuals as if they were from a homogeneous population.

Through use of either coalescent models or empirical population genetics data, we also simulate a set of family triads and apply two family-based association tests, to determine whether there is an association between the marker and the trait. The first test is the test pro-

Table 5

Power Comparisons of the Three Tests (QSAT, TDT_{MK}, and TDT_{VC}) in Simulations based on Empirical Population Genetics Data, under the Random Sampling Scheme

	Power													
	P = .05							P = .01						
Status of High-Risk Allele, Trait Distribution, and Model		TDT_{MK}			TDT_{VC}				TDT_{MK}			TDT_{VC}		
	QSAT	n/3	2n/3	n	n/3	2n/3	n	QSAT	n/3	2n/3	n	n/3	2n/3	n
Fixed:														
Normal:														
Dominant	.98	.53	.77	.86	.31	.55	.64	.97	.31	.61	.77	.16	.37	.50
Additive	.98	.44	.74	.87	.30	.54	.65	.96	.20	.55	.74	.13	.39	.50
Recessive	.91	.32	.55	.65	.23	.43	.53	.85	.14	.35	.50	.10	.29	.37
Log-normal:														
Dominant	.97	.60	.78	.86	.32	.54	.64	.94	.43	.67	.79	.18	.38	.48
Additive	.96	.61	.79	.90	.36	.57	.66	.90	.39	.64	.80	.21	.40	.53
Recessive	.90	.40	.57	.66	.26	.47	.54	.83	.21	.42	.53	.13	.32	.43
Random:														
Normal:														
Dominant	.90	.31	.43	.54	.20	.38	.48	.81	.17	.27	.39	.09	.22	.34
Additive	.91	.48	.65	.80	.29	.53	.64	.82	.26	.44	.64	.16	.34	.50
Recessive	.96	.60	.75	.84	.33	.55	.64	.93	.41	.61	.75	.19	.38	.48
Log-normal:														
Dominant	.86	.30	.47	.55	.19	.43	.48	.78	.20	.32	.42	.14	.28	.35
Additive	.88	.46	.72	.80	.25	.50	.63	.80	.24	.54	.69	.15	.37	.48
Recessive	.93	.58	.77	.85	.30	.55	.64	.90	.39	.64	.77	.17	.40	.50

NOTE.—Sample size is n = 150 for the QSAT and 50, 100, and 150 for TDT tests.

posed by Monks and Kaplan (2000), and we denote this test the "TDT_{MK}." Similar tests have been proposed by Sun et al. (2000). The second test is based on variance-components models proposed by Fulker et al. (1999), and we denote this test the "TDT_{VC}." In the power comparisons, we simulate n/3, 2n/3, and n trios in the family-based association design, where n is the total number of 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, as parents, 2n/3, 4n/3, and 2n individuals in the total population, and generate the children's genotypes according to their parents' genotypes. For the selective sampling scheme, we choose individuals according to the children's trait values, and the trait values are generated according to the same model as above. The P values of these two tests are evaluated by the simulations.

Results

Population-Structure Inference

The first step in the QSAT procedure is to estimate whether the number of components in the mixture model is one, corresponding to one homogenous population, or two, which implies that there are subpopulations in the sample. When the number of components is estimated to be two, we infer whether two individuals are more likely to be from the same subpopulation or from different subpopulations. In our simulations, when 500 independent biallelic markers are used for the coalescent models, and when 4 × 130 and 8 × 130 markers are used for empirical population genetics data, the number of components can be correctly estimated under all situations, and the relationship between two individuals (whether they are from the same or from different subpopulations) can be correctly inferred >97% of the time (Zhang et al., in press).

Type I Error Rates

Table 1 summarizes type I error rates for the four test statistics under the coalescent models. The results are based on 2,000 replications, with each replication consisting of $n = 150$ randomly sampled individuals for all four tests (n/3 trios for TDT-type tests). A total of 2,000 simulated data sets are used for each sample in the estimation of the P values. Therefore, for the two levels of statistical significance considered, .05 and .01, the standard errors for the type I error rate estimate are $\sqrt{.05 \times .95/2{,}000} \approx 4.87 \times 10^{-3}$ and $\sqrt{.01 \times .99/2{,}000} \approx 2.22 \times 10^{-3}$, respectively. It is apparent from table 1 that the esti-

Table 6

Power Comparisons of the Three Tests (QSAT, TDT_{MK}, and TDT_{VC}) in Simulations based on Empirical Population Genetics Data, under the Selective Sampling Scheme

Status of High-Risk Allele, Trait Distribution, and Model	Power													
	P = .05							P = .01						
		TDT_{MK}			TDT_{VC}				TDT_{MK}			TDT_{VC}		
	QSAT	*n*/3	2*n*/3	*n*	*n*/3	2*n*/3	*n*	QSAT	*n*/3	2*n*/3	n	*n*/3	2*n*/3	*n*
Fixed:														
Normal:														
Dominant	.97	.76	.91	.95	.48	.70	.78	.96	.60	.84	.90	.34	.55	.67
Additive	.98	.76	.95	.98	.39	.66	.75	.97	.54	.88	.96	.30	.45	.65
Recessive	.96	.55	.76	.85	.33	.50	.68	.93	.35	.61	.75	.29	.40	.52
Log-normal:														
Dominant	.98	.80	.93	.97	.49	.66	.80	.96	.63	.86	.94	.35	.50	.69
Additive	.98	.79	.91	.95	.46	.55	.72	.94	.64	.86	.93	.38	.46	.58
Recessive	.82	.49	.65	.78	.39	.51	.65	.73	.35	.56	.66	.29	.39	.50
Random:														
Normal:														
Dominant	.92	.49	.65	.76	.38	.48	.65	.88	.32	.52	.65	.26	.39	.52
Additive	.76	.40	.55	.67	.29	.39	.52	.66	.26	.43	.55	.19	.32	.44
Recessive	.94	.41	.51	.65	.28	.37	.49	.91	.34	.45	.50	.24	.35	.40
Log-normal:														
Dominant	.94	.55	.68	.77	.43	.51	.57	.90	.40	.57	.69	.33	.40	.45
Additive	.91	.50	.62	.71	.35	.41	.49	.82	.36	.50	.61	.29	.33	.38
Recessive	.79	.46	.54	.66	.35	.40	.48	.66	.33	.49	.58	.27	.30	.37

Note.—Sample size is n = 100 for the QSAT and 33, 67, and 100 for TDT tests.

mated type I error rates of the QSAT, TDT_{MK}, and TDT_{VC} are not statistically significantly different from the nominal levels. In contrast, the test statistic T, 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.

The type I error results of simulations using empirical population genetics data are summarized in tables 2 and 3, for random sampling and selective sampling, respectively. The standard errors for the type I error rate estimate are $\sim\sqrt{.05 \times .95/2,600} \approx 4.27 \times 10^{-3}$ and $\sqrt{.01 \times .99/2,600} \approx 1.95 \times 10^{-3}$ for the true error rates of .05 and .01, respectively. It can be seen from tables 2 and 3 that the type I error rates of the QSAT, TDT_{MK}, and TDT_{VC} are not statistically significant from the nominal levels, whereas the type I error rate for the test statistic T is substantially higher than the nominal level in the presence of population stratification.

Power Comparisons

The results of our power comparisons under coalescent models and random sampling are summarized in table 4. The results are based on 2,000 replications, with each replication consisting of n = 150 individuals for the QSAT and $n/3$, $2n/3$, and n trios for TDT-type tests. The QSAT is more powerful than TDT-type tests with three different sample sizes ($n/3$ $2n/3$, and n), and the TDT_{MK} is more powerful than the TDT_{VC}. We also observe that when the population divergence increases, the power of the statistical tests decreases. In addition, the trait distribution and the genetic models affect the power of the tests.

Through use of empirical population genetics data and random sampling, power comparisons are performed under several conditions, including different schemes for the assignment of alleles conferring high trait values, different genetic models, different distributions of trait values, and different sample sizes for TDT-type tests. We use 8 × 130 = 1,040 markers to infer the relationship between each pair of individuals. The results are summarized in table 5. Similar to the simulation results based on coalescent models, the QSAT has the highest power and the TDT_{VC} has the lowest power among the three test statistics compared.

The results of power comparisons under empirical population genetics data and selective sampling are summarized in table 6. The pattern is the same as that under the random sampling scheme. However, the difference between the power of the QSAT and TDT-type tests is not as great as that under the random sampling scheme.

Discussion

It is well known that one major limitation of the traditional association test based on population-based samples is that it is susceptible to population stratification. As a result, recent studies have produced many developments in family-based association designs that are robust to population stratification. However, the traditional association test is, in general, more powerful than family-based association designs, and the sample collection is also easier and less expensive (Risch 2000). Recently, several studies have appeared to use genomic markers to control for population stratification in the analysis of population-based data for qualitative traits (Devlin and Roeder 1999; Pritchard et al. 2000*b;* Reich and Goldstein 2001; Satten et al. 2001; Zhang et al., in press). These studies have demonstrated that this general approach is more efficient than family-based association designs and that it is also robust to population stratification. To extend this general approach to quantitative traits, we have developed a statistical procedure, the QSAT, to identify association between candidate markers and quantitative traits, using population-based data. Our simulation results show that the QSAT has a correct type I error rate in the presence of population structure and that it is more powerful than family-based association designs. The computer program for the QSAT will be made available at the Hongyu Zhao Lab of Statistical Genetics Web site.

Although we have compared the power of the QSAT with that of the $\mathrm{TDT}_{\mathrm{MK}}$ and $\mathrm{TDT}_{\mathrm{VC}}$, using three different sample sizes, the comparisons are based on the assumption that a set of independent markers are available for population-structure inferences. If there is only one candidate locus, the QSAT may require substantially greater genotyping efforts; however, given the low prior probability of a specific gene producing a given trait and the ever-decreasing genotyping cost, it may be more cost-effective to perform a population-based study.

In the present study, we have used a simple statistical procedure to infer whether two individuals are likely to be from the same subpopulation. In our simulations, we have used ⩾500 markers to make such inferences. Because SNPs are less informative than microsatellite markers, fewer markers may be needed for studies involving microsatellite markers; for example, Pritchard et al. (2000*b*) have suggested that >100 microsatellite loci should be used for inferring population structure. In general, it is not easy to give a general statement about the number of markers needed to identify population structure in a sample. On the basis of our simulation studies, we feel that 500–1,000 SNPs will allow us to make relatively accurate inferences. If two subpopulations are very similar, >1,000 SNPs may be required to distinguish them from one another; however, in this case, spurious association would not pose a severe problem, since the two subpopulations are sufficiently similar to each other. In addition, with the rapid progress in the identification of polymorphic markers in the human genome and many ongoing population genetics studies, some genetic markers may be found to have better power for distinguishing subpopulations. Progress in this area will likely lead to a set of markers that are more informative for population-structure inferences. In addition, genotyping cost will definitely decrease.

In the case that multiallelic markers are used in a genetic association study, here we outline one approach to extending the QSAT method to a multiallelic trait locus. Suppose that there are m alleles $A_1, \ldots, A_m$ at the trait locus; hence, there are $m(m+1)/2$ genotypes A_iA_j $(1 \geqslant i \geqslant j \geqslant m)$. If we denote the $m(m+1)/2$ genotypes as G_j, where $j = 1, 2, \ldots, m(m+1)/2$, and denote the genotypic score of the ith individual and the jth genotype as

$$X_{ij} = \begin{cases} 1 & \text{if the genotype of the } i\text{th individual is } G_j \\ 0 & \text{otherwise} \end{cases} \; .$$

Following the definitions of $\overline{A}_i$, $\overline{D}_i$, A_{wi}, and D_{wi} in the Methods section, we may similarly define $\overline{X}_{ij}$ and X_{wij}. We can then test the null hypothesis $H_0\colon \beta_1 = \cdots = \beta_{m(m+1)/2-1} = 0$ through use of the following regression model:

$$y_i = \alpha_0 + \sum_{j=1}^{m(m+1)/2-1} (\alpha_j \overline{X}_{ij} + \beta_j X_{wij}) + e_{ij} \; .$$

In the present study, we have introduced a similarity indicator, W_{ij}, between the ith and the jth individuals from the t_{ijk} to characterize whether these two individuals are more likely to be from the same or from different subpopulations. An alternative approach to using the t_{ijk} values is to directly apply these estimated probabilities in the QSAT method; however, we have found that this approach is less powerful than that using the W_{ij} values (data not shown).

The QSAT proposed in this article involves the pooling of information from all subpopulations. If there are two subpopulations, allele A_1 increases trait values in one subpopulation, and another allele, A_2 increases trait values in another subpopulation, the QSAT may lose power. An alternative method is to directly test the hypothesis $H_0\colon \alpha_1 = \ldots = \alpha_k = 0$ and $\beta_1 = \ldots = \beta_k = 0$ under the model in equation (3). To apply this procedure, we need to infer population structure through use of genomic markers—for example, by means of the procedure proposed by Pritchard et al. (2000*b*). There are two potential problems with this alternative approach: (1) the estimation procedure proposed by Pritchard et al. (2000*b*) tends to overestimate the number of subpopulations in a sample and (2) the degrees of freedom for the test statistic is $2k$, where k is the number of estimated subpopulations, and a test statistic with many degrees of freedom may lose power. If the same allele increases trait values in all subpopulations, the QSAT is likely to be more powerful than this alternative testing procedure. If different alleles increase trait values in different subpopulations, the relative performance of the statistical tests needs further investigation.

Acknowledgments

We thank two referees for their constructive comments, and we thank Dr. Kenneth K. Kidd for access to the ALFRED population genetics database. This work was supported, in part, by National Institutes of Health grant GM59507.

Appendix A

The Expectation of $\hat{\alpha}$ and $\hat{\beta}$ under the Model in Equation (3)

Suppose that there are k subpopulations, with n_i individuals sampled from the ith subpopulation. Let $n = \sum_{i=1}^{k} n_i$ denote the total sample size, μ_i denote the phenotype mean in the ith subpopulation, and let p_i and q_i denote the allele frequencies in the ith subpopulation.

Under the model in equation (3), the LS estimators of α and β are

$$\hat{\alpha} = \frac{V_D C_{Ay} - C_{AD} C_{Dy}}{V_A V_D - C_{AD}^2} ,$$

$$\hat{\beta} = \frac{V_A C_{Dy} - C_{AD} C_{Ay}}{V_A V_D - C_{AD}^2} , \tag{A1}$$

where

$$V_A = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (A_{ij} - \overline{A})^2 ,$$

$$C_{Ay} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (A_{ij} - \overline{A})(y_{ij} - \overline{y}) ,$$

$$V_D = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (D_{ij} - \overline{D})^2 ,$$

$$C_{Dy} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (D_{ij} - \overline{D})(y_{ij} - \overline{y}) ,$$

and

$$C_{AD} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (A_{ij} - \overline{A})(D_{ij} - \overline{D}) .$$

From equation (2), we have

$$E(\hat{\alpha} | A_{ij}, D_{ij}) = \mu_\alpha + \sum_{i=1}^{k} \alpha_i a_{(\alpha)i} + \sum_{i=1}^{k} \beta_i d_{(\alpha)i} ,$$

$$E(\hat{\beta} | A_{ij}, D_{ij}) = \mu_\beta + \sum_{i=1}^{k} \alpha_{\mathrm{i}} a_{(\beta)i} + \sum_{i=1}^{k} \beta_{\mathrm{i}} d_{(\beta)i} , \tag{A2}$$

where

$$\mu_\alpha = \frac{V_D C_{Au} - C_{AD} C_{Du}}{V_A V_D - C_{AD}^2} ,$$

$$\mu_\beta = \frac{V_A C_{Du} - C_{AD} C_{Au}}{V_A V_D - C_{AD}^2} ,$$

$$C_{A\mu} = \sum_{i=1}^{k} n_i (\overline{A}_i - \overline{A}) \mu_i ,$$

and

$$C_{D\mu} = \sum_{i=1}^{k} n_i (\overline{D}_i - \overline{D}) \mu_i .$$

In the equation, $\alpha_{(\alpha)i}$, $d_{(\alpha)i}$, $a_{(\beta)i}$, and $d_{(\beta)i}$ are functions of A_{ij} and D_{ij} and satisfy the following conditions:

$$\sum_{i=1}^{k} a_{(\alpha)i} = \sum_{i=1}^{k} d_{(\beta)i} = 1 ,$$

$$\sum_{i=1}^{k} a_{(\beta)i} = \sum_{i=1}^{k} d_{(\alpha)i} = 0 . \tag{A3}$$

If $\alpha_1 = \ldots = \alpha_k = \alpha^*$ and $\beta_1 = \ldots = \beta_k = \beta^*$, it follows from equations (A1), (A2), and (A3) that $E(\hat{\alpha} | A_{ij}, D_{ij}) = \mu_\alpha + \alpha^*$ and $E(\hat{\beta} | A_{ij}, D_{ij}) = \mu_\beta + \beta^*$. Furthermore, under the null hypothesis $H_0 : \alpha_1 = \ldots = \alpha_k = 0$ and $\beta_1 = \ldots = \beta_k = 0$, $E(\hat{\alpha}) = E(\mu_\alpha)$ and $E(\hat{\beta}) = E(\mu_\beta)$.

Let $V = V_A V_D - C_{AD}^2$. For large values of n and n_i $(i = 1, \ldots, k)$,

$$E(\mu_\alpha) \approx \frac{1}{E(V)} [E(V_D) E(C_{Au}) - E(C_{AD}) E(C_{D\mu})]$$

and

$$E(\mu_\beta) \approx \frac{1}{E(V)} [E(V_A) E(C_{Du}) - E(C_{AD}) E(C_{A\mu})] .$$

Through some calculations, we have

$$E(V_A) = \sum_{i=1}^{k} n_i(p_i^2 + q_i^2) - \frac{1}{n}\left\{\sum_{i=1}^{k} 2n_ip_iq_i + \left[\sum_{i=1}^{k} n_i(p_i - q_i)\right]^2\right\} ,$$

$$E(V_D) = \frac{n-1}{n}\sum_{i=1}^{k} 2n_ip_iq_i - \frac{1}{n}\left[\left(\sum_{i=1}^{k} 2n_ip_iq_i\right)^2 - \sum_{i=1}^{k} 4n_ip_i^2q_i^2\right] ,$$

$$E(C_{AD}) = -\frac{1}{n}\left[\sum_{i=1}^{k} n_i(p_i - q_i) \cdot \sum_{i=1}^{k} 2n_ip_iq_i - \sum_{i=1}^{k} 2n_i(p_i - q_i)p_iq_i\right] ,$$

$$E(C_{A\mu}) = \sum_{i=1}^{k} 2n_i(p_i - \overline{p})\mu_i = \sum_{i=1}^{k} 2\mu_i(p_i - \overline{p})(\mu_i - \overline{\mu}) ,$$

and

$$E(C_{D\mu}) = \sum_{i=1}^{k} 2n_i[p_i - \overline{p} - (p_i^2 - \overline{p}^2)](\mu_i - \overline{\mu}) ,$$

where $\overline{p} = \frac{1}{n}\Sigma_{i=1}^{k} n_i$ and $\overline{p}^2 = \frac{1}{n}\Sigma_{i=1}^{k} n_ip_i^2$. In the case of two subpopulations—that is, when $k = 2$—and with an equal number of individuals from each subpopulation—that is, when $n_1 = n_2$—we have

$$E(\mu_a) = (p_1 - p_2)(\mu_1 - \mu_2)\frac{p_1 + p_2 - 2p_1p_2}{\Omega} ,$$

$$\mathrm{E}(\mu_\beta) = (\mathrm{p}_1 - \mathrm{p}_2)^3(\mu_1 - \mu_2)\frac{1 - (\mathrm{p}_1 + \mathrm{p}_2)}{\Omega} , \qquad (A4)$$

where $\Omega = 2(p_1q_1 + p_2q_2)[(p_1q_1 + p_2q_2)^2 + (p_1 - p_2)^2]$. From equation (A4), we can see that if the phenotypic means and allele frequencies vary between subpopulations—that is, if $\mu_1 \neq \mu_2$ and $p_1 \neq p_2$—then $E(\mu_\alpha) \neq 0$. Furthermore, if $p_1 + p_2 \neq 1$, then $E(\mu_\beta) \neq 0$.

Appendix B

The Expectations of $\hat{\alpha}_w$, $\hat{\beta}_w$, $\hat{\alpha}_b$, and $\hat{\beta}_b$ under the Model in Equation (4)

Under the model in equation (4), the LS estimates of $\hat{\alpha}_w$ and $\hat{\beta}_w$ are

$$\hat{\alpha}_\mathrm{w} = \frac{V_{D_\mathrm{w}}C_{A_\mathrm{w}y} - C_{A_\mathrm{w}D_\mathrm{w}}C_{D_\mathrm{w}y}}{V_{A_\mathrm{w}}V_{D_\mathrm{w}} - C_{A_\mathrm{w}D_\mathrm{w}}^2}$$

and

$$\hat{\beta}_\mathrm{w} = \frac{V_{A_\mathrm{w}}C_{D_\mathrm{w}y} - C_{A_\mathrm{w}D_\mathrm{w}}C_{A_\mathrm{w}y}}{V_{A_\mathrm{w}}V_{D_\mathrm{w}} - C_{A_\mathrm{w}D_\mathrm{w}}^2} ,$$

where

$$V_{A_\mathrm{w}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} A_{\mathrm{w}ij}^2 ,$$

$$C_{A_\mathrm{w}y} = \sum_{i=1}^{k}\sum_{i=1}^{n_i} A_{\mathrm{w}ij}(y_{ij} - \overline{y}) ,$$

$$V_{D_\mathrm{w}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} D_{\mathrm{w}ij}^2 ,$$

$$C_{D_\mathrm{w}y} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} D_{\mathrm{w}ij}(y_{ij} - \overline{y}) ,$$

and

$$C_{A_\mathrm{w}D_\mathrm{w}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} A_{\mathrm{w}ij}D_{\mathrm{w}ij} .$$

Note that $\sum_{j=1}^{n_i} A_{\mathrm{w}ij} = \sum_{j=1}^{n_i} D_{\mathrm{w}ij} = 0$. Under the model in equation (2), we have

$$E(C_{A_\mathrm{w}y}) = \sum_{i=1}^{k}\left(\alpha_i\sum_{j=1}^{n_i} A_{\mathrm{w}ij}^2 + \beta_i\sum_{j=1}^{n_i} A_{\mathrm{w}ij}D_{\mathrm{w}ij}\right)$$

and

$$E(C_{D_\mathrm{w}y}) = \sum_{i=1}^{k}\left(\alpha_i\sum_{j=1}^{n_i} A_{\mathrm{w}ij}D_{\mathrm{w}ij} + \beta_i\sum_{j=1}^{n_i} D_{\mathrm{w}ij}^2\right) .$$

After some algebraic calculations, we obtain

$$E(\hat{\alpha}_\mathrm{w}|A_{ij}, D_{ij}) = \sum_{i=1}^{k} \alpha_i a_{(\mathrm{w}\alpha)i} + \sum_{i=1}^{k} \beta_i d_{(\mathrm{w}\alpha)i}$$

and

$$E(\hat{\beta}_\mathrm{w}|A_{ij}, D_{ij}) = \sum_{i=1}^{k} \alpha_i a_{(\mathrm{w}\beta)i} + \sum_{i=1}^{k} \beta_i d_{(\mathrm{w}\beta)i} ,$$

where $a_{(\alpha)i}$, $d_{(\alpha)i}$, $a_{(\beta)i}$, and $d_{(\beta)i}$ are functions of A_{ij} and D_{ij} and satisfy

$$\sum_{i=1}^{k} a_{(\mathrm{w}\alpha)i} = \sum_{i=1}^{k} d_{(\mathrm{w}\beta)i} = 1$$

and

$$\sum_{i=1}^{k} a_{(\mathrm{w}\beta)i} = \sum_{i=1}^{k} d_{(\mathrm{w}\alpha)i} = 0 .$$

If $\alpha_1 = \dots = \alpha_k = \alpha^*$ and $\beta_1 = \dots = \beta_k = \beta^*$, it follows that $E(\hat{\alpha}|A_{ij}, D_{ij}) = \alpha^*$ and $E(\hat{\beta}_{ij}|A_{ij}, D_{ij}) = \beta^*$. In this case, $\hat{\alpha}_\mathrm{w}$ and $\hat{\beta}_\mathrm{w}$ are both unbiased estimators of the additive and dominance genetic values α^* and β^*, respectively. Even if the additive and dominance genetic values vary among subpopulations, we still have $E(\hat{\alpha}_{\text{w}}) = E(\hat{\beta}_{\text{w}}) = 0$ under the null hypothesis H_0.

Under the model in equation (4), the LS estimates of $\hat{\alpha}_{\text{b}}$ and $\hat{\beta}_{\text{b}}$ are given by

$$\hat{\alpha}_{\text{b}} = \frac{V_{D_{\text{b}}}C_{A_{\text{b}}y} - C_{A_{\text{b}}D_{\text{b}}}C_{D_{\text{b}}y}}{V_{A_{\text{b}}}V_{D_{\text{b}}} - C^2_{A_{\text{b}}D_{\text{b}}}}$$

and

$$\hat{\beta}_{\text{b}} = \frac{V_{A_{\text{b}}}C_{D_{\text{b}}y} - C_{A_{\text{b}}D_{\text{b}}}C_{A_{\text{b}}y}}{V_{A_{\text{b}}}V_{D_{\text{b}}} - C^2_{A_{\text{b}}D_{\text{b}}}} \ ,$$

where

$$V_{A_{\text{b}}} = \sum_{i=1}^{k} n_i(\overline{A}_i - \overline{A})^2 \ ,$$

$$C_{A_{\text{b}}y} = \sum_{i=1}^{k} n_i(\overline{A}_i - \overline{A})(\overline{y}_i - \overline{y}) \ ,$$

$$V_{D_{\text{b}}} = \sum_{i=1}^{k} n_i(\overline{D}_i - \overline{D})^2 \ ,$$

$$C_{D_{\text{b}}y} = \sum_{i=1}^{k} n_i(\overline{D}_i - \overline{D})(\overline{y}_i - \overline{y}) \ ,$$

and

$$C_{A_{\text{b}}D_{\text{b}}} = \sum_{i=1}^{k} n_i(\overline{A}_i - \overline{A})(\overline{D}_i - \overline{D}) \ .$$

It follows from the model in equation (2) that, after some algebraic calculations,

$$E(\hat{\alpha}_{\text{b}}|A_{ij}, D_{ij}) = \mu_{\text{b}\alpha} + \sum_{i=1}^{k} \alpha_i a_{(\text{b}\alpha)i} + \sum_{i=1}^{k} \beta_i d_{(\text{b}\alpha)i}$$

and

$$E(\hat{\beta}_{\text{b}}|A_{ij}, D_{ij}) = \mu_{\text{b}\beta} + \sum_{i=1}^{k} \alpha_i a_{(\text{b}\beta)i} + \sum_{i=1}^{k} \beta_i d_{(\text{b}\beta)i} \ ,$$

where

$$\mu_{\text{b}\alpha} = \frac{V_{D_{\text{b}}}C_{A_{\text{b}}\mu} - C_{A_{\text{b}}D_{\text{b}}}C_{D_{\text{b}}\mu}}{V_{D_{\text{b}}}V_{A_{\text{b}}} - C^2_{A_{\text{b}}D_{\text{b}}}} \ ,$$

$$\mu_{\text{b}\beta} = \frac{V_{A_{\text{b}}}C_{D_{\text{b}}\mu} - C_{A_{\text{b}}D_{\text{b}}}C_{A_{\text{b}}\mu}}{V_{D_{\text{b}}}V_{A_{\text{b}}} - C^2_{A_{\text{b}}D_{\text{b}}}} \ ,$$

$$C_{A_{\text{b}}\mu} = \sum_{i=1}^{k} n_i(\overline{A}_i - \overline{A})\mu_i \ ,$$

and

$$C_{D\mu} = \sum_{i=1}^{k} n_i(\overline{D}_i - \overline{D})\mu_i \ .$$

The variables $a_{(\text{b}\alpha)i}$, $d_{(\text{b}\alpha)i}$, $a_{(\text{b}\beta)i}$, and $d_{(\text{b}\beta)i}$ are functions of A_{ij} and D_{ij} and satisfy

$$\sum_{i=1}^{k} a_{(\text{b}\alpha)i} = \sum_{i=1}^{k} d_{(\text{b}\beta)i} = 1$$

and

$$\sum_{i=1}^{k} a_{(\text{b}\beta)i} = \sum_{i=1}^{k} d_{(\text{b}\alpha)i} = 0 \ .$$

If $\alpha_1 = \ldots = \alpha_k = \alpha^*$ and $\beta_1 = \ldots = \beta_k = \beta^*$, it follows that $E(\hat{\alpha}_{\text{b}}|A_{ij}, D_{ij}) = \mu_{\text{b}\alpha} + \alpha^*$ and $E(\hat{\beta}_{\text{b}}|A_{ij}, D_{ij}) = \mu_{\text{b}\beta} + \beta^*$. Furthermore, under the null hypothesis $\alpha_1 = \ldots = \alpha_k = 0$ and $\beta_1 = \ldots = \beta_k = 0$, $E(\hat{\alpha}) = E(\mu_\alpha)$ and $E(\hat{\beta}) = E(\mu_\beta)$.

Let $V_{\text{b}} = V_{A_{\text{b}}}V_{D_{\text{b}}} - C^2_{A_{\text{b}}D_{\text{b}}}$. For large values of n and n_i ($i = 1, \ldots, k$),

$$E(\mu_{\text{b}\alpha}) \approx \frac{1}{E(V)}[E(V_D)E(C_{Au}) - E(C_{AD})E(C_{D\mu})]$$

and

$$E(\mu_{\text{b}\beta}) \approx \frac{1}{E(V)}[E(V_A)E(C_{Du}) - E(C_{AD})E(C_{A\mu})] \ .$$

Therefore, we have

$$E(V_{A_{\text{b}}}) = \sum_{i=1}^{k}\left[2p_iq_i + n_i(p_i - q_i)^2 - \frac{2n_i}{n}p_iq_i - n(2\overline{p} - 1)^2\right] ,$$

$$E(V_{D_{\text{b}}}) = 2\sum_{i=1}^{k}\left[\frac{n - n_i}{n}p_iq_i(1 - 2p_iq_i) + 2n_ip_i^2q_i^2\right] - \frac{1}{n}\left[\sum_{i=1}^{k} 2n_ip_iq_i\right]^2 ,$$

$$E(C_{A_{\text{b}}D_{\text{b}}}) = 2\sum_{i=1}^{k}\left[n_i\left(1 + \frac{1}{n}\right) - 1\right](p_i - q_i)p_iq_i - \frac{2}{n}\sum_{i=1}^{k} n_i(p_i - q_i) \cdot \sum_{i=1}^{k} n_ip_iq_i \ ,$$

$$E(C_{A_{\text{b}}\mu}) = \sum_{i=1}^{k} 2\mu_i(p_i - \overline{p})\mu_i \ ,$$

and

$$E(C_{D\mu}) = \sum_{i=1}^{k} 2n_i[p_i - \overline{p} - (p_i^2 - \overline{p}^2)]\mu_i \ .$$

For $k = 2$ and $n_1 = n_2$, we have

$$E(\mu_{\text{b}a}) = (p_1 - p_2)(\mu_1 - \mu_2)\frac{(p_1q_1 + p_2q_2)(p_1 + p_2 - 2p_1p_2)}{\triangle} \ ,$$

$$E(\mu_{\text{b}\beta}) = (p_1 - p_2)^3(\mu_1 - \mu_2)\frac{1 - (p_1 + p_2)}{\triangle} \ , \tag{A5}$$

where

$$\triangle = 2[(p_1 - p_2)^2[2p_1^2q_1^2 + 2p_2^2q_2^2 + n(p_1 - p_2)^2(p_1q_1 + p_2q_2)] + \frac{2}{n}[p_1^3q_1^3 + p_2^3q_2^3 + 2p_1q_1p_2q_2(p_1 - p_2)^2] \; .$$

From equation (A5), we can see that if the phenotypic means and allele frequencies vary between the two subpopulations—that is, if $\mu_1 \neq \mu_2$ and $p_1 \neq p_2$—then $E(\mu_{b\alpha}) \neq 0$. Furthermore, if $p_1 + p_2 \neq 1$, then $E(\mu_{b\beta}) \neq 0$.

Electronic-Database Information

The URLs for data in this article are as follows:

ALFRED, http://alfred.med.yale.edu/alfred/index.asp (for empirical population genetics data)

Hongyu Zhao Lab of Statistical Genetics, http://bioinformatics.med.yale.edu/ (for QSAT computer program)

References

Abecasis GR, Cardon LR, Cookson OC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292

Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944

Celeux G, Govaert G (1995) Gaussian parsimonious clustering model. Pattern Recognition 28:781–793

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

Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman NW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92: 6723–6727

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. Stat Sci 9:307–319

——— (1997) Computational methods for the coalescent. In: Tavaré S, Donnelly P (eds) Progress in population genetics and human evolution. IMA Vol 87. Springer-Verlag, pp 165–182

Monks SA, Kaplan NL (2000) Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. Am J Hum Genet 66:576–592

Morton NE, Collins A (1998) Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci USA 95:11389–11393

Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK (2001) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms: an update. Nucleic Acids Res 29:317–319

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Pritchard JK, Stephens M, Donnelly P (2000*a*) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000*b*) Association mapping in structured population. Am J Hum Genet 67:170–181

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

Risch N (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. Genome Res 8:1273–1288

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 68:466–477

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 66:1616–1630

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–513

Sun F, Flanders WD, Yang Q, Zhao HY (2000) Transmission/disequilibrium tests for quantitative traits. Ann Hum Genet 64:555–565

Teng J, Risch N (1999) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. Genome Res 9:234–241

van den Oord EJCG (1999) A comparison between different designs and tests to detect QTLs in association studies. Behav Genet 29:245–256

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Zhang SL, Kidd KK, Zhao HY. Detecting genetic association in case-control studies using similarity-based association tests. Statistica Sinica (in press)