Statistica Sinica **12**(2002), 337-359

DETECTING GENETIC ASSOCIATION IN CASE-CONTROL STUDIES USING SIMILARITY-BASED ASSOCIATION TESTS

Shuanglin Zhang[1,2], Kenneth K. Kidd[1] and Hongyu Zhao[1]

[1] *Yale University School of Medicine and* [2] *Michigan Technological University*

Abstract: Although traditional case-control studies 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 due to overmatching between cases and controls. Furthermore, case-control studies have the advantages of easy sample collection. Recently, several statistical methods have been proposed for association tests in structured populations using case-control designs that may be robust to population stratification. In this article, we propose a similarity-based association test (SAT) to identify association between a candidate marker and a disease of interest using case-control designs. 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 by comparing within-subpopulation allele-frequency differences between cases and controls. Simulation results show that the SAT has correct type-I error rate in the presence of population stratification. The power of the SAT is higher than that using family-based association designs and is also higher than other robust association methods when the high-risk allele is the same across all subpopulations.

Key words and phrases: Case-control studies, coalescent models, population genetics, population stratification.

1. Introduction

With the completion of the Human Genome Project, hundreds of thousands of genetic markers have been identified in humans. Such abundant information is invaluable for the efforts of mapping disease genes. Because it has become clear that traditional linkage methods may not offer enough power to localize genes conferring small to moderate risks to a trait of interest, association studies may serve as a powerful and viable alternative to mapping complex disease genes (Risch and Merikangas (1996)). One major limitation of standard case-control association studies is that *spurious association* can result from population stratification. Such spurious association can occur when the disease frequency varies across subpopulations, thereby increasing the probability that affected individuals will be sampled from certain subpopulations and increasing the chance of any marker allele with high frequency in the overrepresented subpopulations being associated with the phenotype. One well-known example is the immunoglobulin gene Gm for non-insulin-dependent-diabetes mellitus (Knowler, Williams, Pettitt and Steinberg (1988)). Among residents of the Gila River Indian Community in Arizona, diabetes was associated with the haplotype Gm. However, this association no longer exists among ethnically homogeneous subjects. The confounding by population stratification occurred because the Gm haplotype serves as a marker for European heritage, and the risk of diabetes varies with the level of this ancestry. Population genetics studies have shown that allele frequency at some loci can vary considerably among populations contributing to the U.S. white population (Kang, Palmatier and Kidd (1999)).

To avoid false positives resulting from population stratification, family-based association studies (Falk and Rubinstein (1987), Spielman, McGinnis and Ewens (1993)) have received much attention in the literature because of their robustness to population stratification and their higher power to detect genes of small to moderate effects compared to linkage studies. However, such family-based association designs are less powerful than standard case-control designs in a homogeneous population both for the analysis of qualitative traits and that of quantitative traits (Morton and Collins (1998), Risch and Teng (1998), van den Oord (1999)). Furthermore, case-control designs have practical advantages over family-based designs in that collecting DNA from unrelated cases and controls is often easier than collecting DNA from relatives of affected individuals, especially for late-onset diseases. In addition, the same genotype data from unrelated normal controls may be used for separate genetic studies. However, to fully realize the power of case-control studies, the population stratification issue still needs to be addressed.

Recently Devlin and Roeder (1999), Prichard, Stephens, Rosenberg and Donnelly (2000), Reich and Goldstein (2001), and Satten, Flanders and Yang (2001) have proposed statistical methods to use genomic markers to control population stratification in the analysis of case-control data. Such methods are more powerful than family-based association designs and robust to population stratification. In this article, we develop an alternative method, the Similarity-based Association Test (SAT), that is valid in the presence of population stratification for case-control data. To construct the test statistic, we first use the genotypes of sampled individuals at a series of independent markers to calculate similarities S_{ij} for individuals i and j. We then model the similarities using a normal mixture model to analyze the similarities as one or two clusters (within-subpopulation group and between-subpopulation group) and use the Bayesian Information Criterion (BIC) to estimate the number of clusters. The test statistic is based on a weighted average of allele frequency differences between cases and controls within each subpopulation. We have performed extensive simulations to assess whether the SAT procedure has the correct type-I error rate in the presence of population stratification and to compare its power with other test statistics, including the STRAT procedure proposed by Pritchard, Stephens, Rosenberg and Donnelly (2000) and the transmission/disequilibrium test (TDT) developed by Spielman, McGinnis and Evens (1993). The simulation results show that the SAT has correct type-I error rate in the presence of population stratification and its power compares favorable with other statistical tests that are robust to population stratification.

2. Similarity-based Association Test (SAT)

Table 1. The 2×2 contingency table for case-control studies using a biallelic marker.

	Number of allele A	Number of allele a	Total number of alleles
Cases	n_{11}	n_{12}	$2n_d$
Controls	n_{21}	n_{22}	$2n_c$
Total numbers of alleles	$2n_A$	$2n_a$	$2n$

Consider a case-control study on genetic association between a biallelic marker $\mathcal{A}$ with two alleles A and a and a trait of interest. The data can be represented in a 2×2 contingency table (Table 1). We assume that all of the individuals in the sample are unrelated. One standard test for association between the trait and marker alleles from this contingency table is the Pearson test statistic:

$$T_p = \frac{(\hat{q}_d - \hat{q}_c)^2}{\widehat{V}}, \tag{1}$$

where $\hat{q}_d = n_{11}/(2n_d)$ and $\hat{q}_c = n_{21}/(2n_c)$ are the estimated allele A frequency in the case and control groups respectively, $\widehat{V} = 2n(n_d n_c)/(n_A n_a)$ is the estimated variance of $\hat{q}_d - \hat{q}_c$ under the null hypothesis of no association. The notations n_{11}, n_{21}, n_d, n_c, n_A, and n_a are defined in Table 1. If the underlying population is homogeneous and there is no association between the trait and the marker, the test statistic T_p has a chi-square distribution with one degree of freedom. However, if the underlying population is not homogeneous, e.g., there is population stratification, the expectation of $\hat{q}_d - \hat{q}_c$ may not be zero and statistical inference based on the test statistic T_p may be biased. For example, let us assume that the sampled individuals come from two different subpopulations and there is no association between marker $\mathcal{A}$ and disease within each subpopulation. Let q_1 and q_2 denote the allele A frequency within subpopulations 1 and 2, respectively, f_1

denote the probability that a sampled affected individual is from the first subpopulation, and g_1 denote the probability that a sampled normal individual is from the first subpopulation. Then we have

$$\begin{aligned} E(\hat{q}_d - \hat{q}_c) &= q_1 f_1 + q_2(1 - f_1) - (q_1 g_1 + q_2(1 - g_1)) \\ &= (q_1 - q_2)(f_1 - g_1). \end{aligned} \tag{2}$$

Therefore, if the allele frequencies differ between the two subpopulations and $f_1 \neq g_1$ (i.e., the disease prevalence is different between the two subpopulations), then $E(\hat{q}_d - \hat{q}_c) \neq 0$. Then the simple chi-square test statistic for the 2 × 2 table does not have a chi-square distribution with one degree of freedom even when there is no disease-marker association in each subpopulation. Such statistical tests that ignore population heterogeneity may lead to erroneous conclusions.

To address this issue, we have developed a statistical procedure that is robust to population stratification in the testing of disease-marker associations. There are two steps in our procedure: (1) we use genotype data at a series of independent markers to infer whether two individuals are more likely to be within the same subpopulation or more likely to be in different subpopulations; (2) we perform an association test using the inference on individual pair relationship. The difference between our method and previous methods lies in how the information from independent markers is utilized to correct for population structures. In the rest of this section, we describe these two steps in detail.

2.1. Statistical inference on whether a pair of individuals belong to the same subpopulation

We first define similarity between two individuals using a set of independent markers in our assessment of whether two individuals are more likely to be within the same subpopulation, or more likely to be in different subpopulations. In this article, we focus on biallelic markers as they are more abundant than other types of genetic markers in the human genome and great efforts have been made to identify these markers for association studies. Suppose 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. We further suppose 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 this article, we define the similarity S_{ij} between the ith and the jth individuals as

$$S_{ij} = \frac{d_{\max} - d_{ij}}{d_{\max}}, \tag{3}$$

where $d_{\max}$ is the maximum observed value of the d_{ij} across all pairs of individuals.

For individuals within the same subpopulation, we expect similarity to be smaller than similarity between individuals from different subpopulations. Therefore, we propose to cluster 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}:

$$S_{ij} \sim \sum_{k=1}^{K} p_k N(S_{ij}, \mu_k, \sigma_k^2), \tag{4}$$

where K represents the number of components in the mixture model, p_k denotes the proportion of the kth component, and $N(s, \mu_k, \sigma_k^2)$ 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 using the Clustering Expectation-Maximization (CEM) method (Celeux and Govaert (1995)).

The choice of the number of components in the normal mixture model (4) is a difficult problem. Biernacki and Govaert (1999) and Biernacki, Celeux and Govaert (1999) discussed several criteria for choosing the number of component K, including the *Akaike information criterion* $\mathrm{AIC}(K) = -2L(K)+2M(K)$ (Akaike (1974)) and the *Bayesian information criterion* (BIC) $\mathrm{BIC}(K) = -2L(K) + M(K)\log N$, where N is the total number of observations, $L(K) = \sum_{i,j} \log(\sum_{k=1}^{K} \hat{p}_k N(S_{ij}, \hat{\mu}_k, \hat{\sigma}_k^2))$ is the maximized log likelihood for a given K, and $M(K)$ is the number of free parameters in the mixture model. Based on extensive simulations, Biernacki, Celeux and Govaert (1999) concluded that the BIC criterion behaves better in general and we use it. From our experience with simulated data sets based on both coalescent models and on empirical population genetics data, a choice for K is often made between 1 and 2. The case of $K = 1$ corresponds to a single population, no population heterogeneity, whereas $K = 2$ admits a within-population component and a between-population component. Note that $K = 2$ 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

$$t_{ijk} = \frac{\hat{p}_k N(S_{ij,}\hat{\mu}_k, \hat{\sigma}_k^2)}{\hat{p}_1 N(S_{ij}, \hat{\mu}_1, \hat{\sigma}_1^2) + \hat{p}_2 N(S_{ij}, \hat{\mu}_2, \hat{\sigma}_2^2)}$$

is the conditional probability that S_{ij} arises from the kth mixture component. Assume $\hat{\mu}_1 > \hat{\mu}_2$, if $t_{ij1} > 0.5$, we define the similarity indicator W_{ij} between the ith and the jth individuals to be 1 and assume these two individuals belong to the same subpopulation in our subsequent analysis. If $t_{ij1} < 0.5$, we define the similarity indicator W_{ij} between the ith and the jth individuals to be 0 and assume these two individuals belong to different subpopulations.

2.2. Similarity-based association test

Assume the case-control sample consists of n_d affected individuals and n_c normal individuals. Let $D_{ii'}$ denote the similarity indicator between the ith and the i'th affected individuals, B_{ij} denote the similarity indicator between the ith affected individual and the jth normal individual, and $N_{jj'}$ denote the similarity indicator between the jth and the j'th normal individuals. Let x_i denote the genotype of the ith affected individual and y_j denote the genotype of the jth normal individual at the candidate marker. We start our introduction of the Similarity-based Association Test (SAT) by considering

$$U_s = \sum_{i=1}^{n_d} x_i \sqrt{\frac{m_{di}(m_{di}+k_{di})}{k_{di}}} - \sum_{j=1}^{n_c} y_j \sqrt{\frac{m_{cj}(m_{cj}+k_{cj})}{k_{cj}}}, \qquad (5)$$

where $m_{di} = \sum_{j=1}^{n_c} B_{ij}$, $m_{cj} = \sum_{i=1}^{n_d} B_{ij}$, $k_{di} = \sum_{i'=1}^{n_d} D_{ii'}$, and $k_{cj} = \sum_{j'=1}^{n_c} N_{j'j}$. It is easy to see that m_{di} is the number of normal individuals in the same subpopulation as the ith affected individual, m_{cj} is the number of affected individuals in the same subpopulation as the jth normal individual, k_{di} is the number of affected individuals in the same subpopulation as the ith affected individual, and k_{cj} is the number of normal individuals in the same subpopulation as the jth normal individual.

To better understand the meaning of U_s, we consider an admixed population with two subpopulations. We suppose that the sample consists of n_{d_1} affected individuals and n_{c_1} normal individuals from the first subpopulation, and n_{d_2} affected individuals and n_{c_2} normal individuals from the second subpopulation. If we can correctly identify all pairwise relationships, then it is easy to show that $U_s = 2\{\sqrt{n_{d_1}n_{c_1}n_1}(\hat{q}_{d_1}-\hat{q}_{c_1})+\sqrt{n_{d_2}n_{c_2}n_2}(\hat{q}_{d_2}-\hat{q}_{c_2})\}$, where $n_1 = n_{d_1}+ n_{c_1}$ is the number of individuals from the first subpopulation, $\hat{q}_{d_1} = \sum_{i=1}^{n_{d_1}} x_i/(2n_{d_1})$ is the allele A frequency among the affected individuals in the first subpopulation, $\hat{q}_{c_1} = \sum_{j=1}^{n_{c_1}} y_j/(2n_{c_1})$ is the allele A frequency among the normal individuals within the first subpopulation, respectively, and n_2, $\hat{q}_{d_2}$ and $\hat{q}_{c_2}$ are similarly defined for the second subpopulation. Therefore, U_s is the weighted sum of the allele frequency differences between the affected individuals and the normal individuals within each subpopulation. In general, if the sampled individuals come from M subpopulations and we can correctly infer the relationship between any pair of individuals, the statistic U_s is the weighted sum of the allele frequency differences between the affected and the normal individuals within each subpopulation:

$$U_s = 2\sum_{m=1}^{M}\sqrt{n_{d_m}n_{c_m}n_m}(\hat{q}_{d_m} - \hat{q}_{c_m}), \tag{6}$$

where the notation is similarly defined as above for the mth subpopulation. Under the assumption that there is no disease-marker association within each subpopulation, $E(U_s) = 0$, and the variance of U_s is $\sigma^2 = 2\sum_{m=1}^{M} n_m^2 q_m(1 - q_m)$, where q_m is the allele A frequency in the mth subpopulation. We propose the following estimator for σ^2:

$$\hat{\sigma}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} z_i(1 - \frac{z_j}{2})W_{ij}, \tag{7}$$

where z_i ($z_i = 0, 1$ or 2) denotes the genotype score at the candidate locus for the ith individual in the sample without regard to the disease status. It is easy to show that, under the assumption that we can correctly infer the relationship between the ith and the jth individuals in the sample, $\hat{\sigma}^2 = 2\sum_{m=1}^{M} n_m^2 \hat{q}_m(1 - \hat{q}_m)$, where $\hat{q}_m$ is the observed allele A frequency in the mth subpopulation. Based on the above discussion, we define our SAT test statistic as

$$SAT = \frac{U_s}{\widehat{\sigma}}, \tag{8}$$

where U_s was defined in (5) and $\widehat{\sigma}$ was defined in (7). This test statistic asymptotically follows the standard normal distribution.

3. Simulation Models and Other Statistical Tests Considered

In this section, we discuss the simulation models used to assess whether the SAT is robust to population stratification and to compare the power of the SAT with other association tests. In our simulation studies, we either generate the data through coalescent models or through empirical population genetics data. Other parameters varied in our simulations include different modes of inheritance, different prevalences among the subpopulations, and different genetic distances between the candidate locus and the disease gene.

3.1. Coalescent models

Coalescent models introduced by Kingman (1982a,b) were used in our simulations. Recent developments of coalescent theory can be found in a review article by Fu and Li (1999). Pritchard, Stephens, Rosenberg and Donnelly (2000) considered coalescent models with constant population sizes. To be more close to reality, we consider coalescent models with variable population sizes (Griffiths and Tavaré (1994, 1997)) in our simulations, and allow subpopulations to have different sizes.

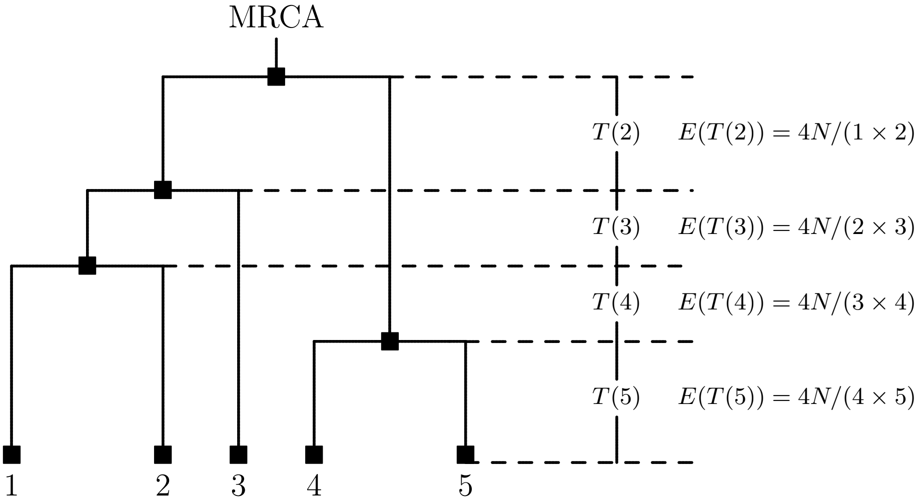
MRCA
$T(2)$ $E(T(2)) = 4N/(1 \times 2)$
$T(3)$ $E(T(3)) = 4N/(2 \times 3)$
$T(4)$ $E(T(4)) = 4N/(3 \times 4)$
$T(5)$ $E(T(5)) = 4N/(4 \times 5)$
1
2
3
4
5

Figure 1. An example of a genealogy of a sample of five haplotypes. The expected durations of the different time intervals under Wright-Fisher model are shown on the right, measured in units of generations.

We give a brief description of the coalescent methods used to generate marker genotypes for sampled individuals. Consider n haplotypes from one population, they are connected by a single phylogenetic tree or genealogy in which the root of the tree is the most recent common ancestor (MRCA) of these n haplotypes. Figure 1 shows an example of a genealogy for a sample of five haplotypes. The coalescent process we used generates the genealogy of the n sampled haplotypes and generates the alleles for every sampled haplotype (the bottom nodes of the genealogy) given the genealogy. Under the Wright-Fisher neutral model with constant population size N haplotypes, the times between the nodes of the genealogy are exponentially distributed with means $4N/[k(k-1)]$, where k is the number of lineages at the level of the bottom node of this time interval. The topology of the tree under this model can be generated by generating the time between nodes and randomly choosing lineages to combine at each node. For a given topology, if the mutation rate for each lineage in each generation is μ, the probability of no mutation on a given branch of length t is $(1-\mu)^t$. In our simulations, we assume that there are two alleles A and a and there are no recurrent mutations. Thus, if the allele of the MRCA is A, the allele at the node immediately below the MRCA is A if no mutation occurred between this node and the MRCA; the allele at the node is a if at least one mutation occurred between this node and the MRCA. In this way, we can generate the allele node by node from the MRCA to the bottom nodes. Once the allele at a node is a, all the alleles at the descendant nodes are of type a because of no recurrent mutations. When we obtain the haplotypes, we randomly pair the haplotypes in one population to form genotypes. For the case of variable population size, the distribution of the time length between nodes needs to be modified to accommodate variable population sizes, see Griffiths and Tavaré (1994, 1997) for details. When we examine the effects of recombinations between a genetic marker and the underlying gene, in addition to considering mutation events, we also consider recombination events whose probability depends on the assumed distance between the candidate marker and the disease mutation. If there is no recombination between the two sites from the MRCA to the present time, we assign the allele at the candidate marker according to the mutation process. If there is at least one recombination between the two sites from the MRCA to the present time, we assign the allele at the candidate marker according to the allele frequencies at this marker. It is straightforward to modify the above simulation methods to generate the genotype data of the structured population models to incorporate population structures.

Model A. There was an ancestral population which had evolved for a long period of time in constant population size, then this population divided into two subpopulations T generations before the present time. From the time of division, the two subpopulations experienced exponential growth without migrations. We assume that the sizes of the two subpopulations were 100 and 10000 when they divided, and the current sizes are 10^7 and 5×10^7, respectively, so the first subpopulation experienced a more rapid growth. We consider three population divergence times between the two subpopulations: (1) $T = 500$ generations, (2) $T = 1500$ generations, and (3) $T = 4500$ generations. The first two separation times can be thought of as the divergence times between non-African populations, and the third can be thought of as the divergence time between African and non-African populations (Goldstein, Linares, Cavalli-Sforza and Feldman (1995)).

Let f denote the probability that an affected individual is sampled from the first subpopulation, g denote the probability that a normal individual is sampled from the first subpopulation, and RR be the ratio of the prevalences of the disease in the two subpopulations. For a rare disease,

$$RR \approx \frac{f}{1-f}\frac{1-g}{g}. \tag{9}$$

In our simulations, we fix $g = 0.2$ and allow RR to vary from 1 to 7. The value of f may be calculated from (9). For each population divergence scenario, we independently simulate L independent markers that are not associated with the disease phenotype. For each marker, we simulate genotypes of $n_A f + n_c g$ individuals from the first subpopulation and $n_A(1 - f) + n_c(1 - g)$ individuals from the second subpopulation. We assume the mutation rate $\mu = 5 \times 10^{-7}$ per generation, and only select markers whose allele frequencies are as least 0.2 in the sample. This threshold was also used by Pritchard and Rosenberg (1999) to approximate the likely characteristic of SNP surveys (Wang et al. (1998)). Among the $n_A f + n_c g$ genotypes generated for the first subpopulation, we randomly assign $n_A f$ individuals to the case group and the others to the control group. Similarly, among the $n_A(1 - f) + n_c(1 - g)$ genotypes generated for the second subpopulation, we randomly assign $n_A(1 - f)$ individuals to the case group and the others to the control group.

To simulate genotypes at the candidate locus, we first simulate the genealogies of the chromosomes carrying the disease mutation using the same coalescent models, except that we are only concerned about the chromosomes carrying the disease mutation. We assume that the chromosomes with the disease mutation experienced the same evolutionary process as other chromosomes, and that the number of affected individuals in the two subpopulations were 10 and 50, respectively, when the two subpopulations divided. We further assume that the number of the affected individuals in both subpopulations is 5×10^5 at present. Therefore, the prevalence of the disease allele is 5% and 1% in the two subpopulations. After we simulate the genealogy, we assign the allele on the MRCA chromosome according to the allele frequencies estimated from genotypes at independent markers. At last, we get the genotypes of disease individuals according to heredity models (dominant and recessive).

Model B. In Model A, there is no migration between the two subpopulations since they divided. In Model B, we assume that the two subpopulations divided 4500 generations ago and recently merged to form an admixed population. In this admixed population, those individuals whose origin is the first subpopulation have all of their genetic materials inherited from the first subpopulation; those individuals whose origin is the second subpopulation have 80% of their genetic materials inherited from their ancestry and 20% of their genetic materials inherited from the first subpopulation. This structure could represent the genetic compositions of the European Americans and the African Americans in the United States. For Model B, we use two steps in our simulations. In the first step, we use the same procedure as that in Model A to get chromosomes' marker alleles at L independent markers and the candidate marker in the two subpopulations. In the second step, for any chromosome in the second subpopulation, we randomly choose 20% of the markers among the independent markers and replace the alleles in these 20% markers by the alleles of corresponding marker alleles of a random chosen chromosome from the first subpopulation. At the candidate marker, we randomly replace 20% of the chromosomes in the second subpopulation by a random sample of the chromosomes from the first subpopulation.

We assume that a total of 500 unlinked biallelic markers are used for our inference on the population structure, and the general population consists of two subpopulations in all coalescent models.

3.2. 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 well. Therefore, in our simulations, we also use empirical population genetics data from a population genetics database ALFRED (Cheung, Osier and Kidd (2000); 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 extracted 130 markers across four populations, including Danes, San Francisco Chinese, Biaka and Maya. We use these four populations to represent the populations from four different continents. For microsatellite markers, because we focus on the use of SNP markers in our methods, we pool the alleles to form biallelic markers with allele frequencies between 10% and 90%. However, this procedure may result in loss of information for the analysis of microsatellite markers and we are currently extending our methods to handle microsatellite markers.

If we sample from the general population, let n_{d1}, n_{d2}, n_{d3}, and n_{d4} denote the number of affected individuals sampled from the four populations (Danes, San Francisco Chinese, Biaka and Maya), respectively, and let n_{c1}, n_{c2}, n_{c3}, and n_{c4} denote the number of normal individuals sampled from the four populations, respectively. In our simulations, we vary the proportions of the affected and normal individuals from the four subpopulations as follows. We first generate the number of normal individuals from the four populations $(n_{c1,}n_{c2,}n_{c3,}n_{c4})$ from a multinomial distribution with parameters (p_1, p_2, p_3, p_4), where (p_1, p_2, p_3, p_4) are random variables generated from the Dirichlet Distribution with parameters $(n^0_{c1}, n^0_{c2}, n^0_{c3}, n^0_{c4})$, $n_c = n^0_{c1} + n^0_{c2} + n^0_{c3} + n^0_{c4}$. This means that, on average, the number of normal individuals sampled from Danes, Chinese, Biaka and Maya are n^0_{c1}, n^0_{c2}, n^0_{c3}, and n^0_{c4}, respectively. However, for each realized sample, we allow the exact number of individuals from each subpopulation to vary. In our simulations, we set $n^0_{c1} : n^0_{c2} : n^0_{c3} : n^0_{c4} = 10 : 3 : 4 : 3$ to increase the chance that every population is represented in the overall sample. We then generate the number of affected individuals from the four populations (n_{d1}, n_{d2}, n_{d3},n_{c4}) similarly, but we assume that the disease risk ratios in the four subpopulations are $1 : 1{+}R : 2{+}R : 3{+}R$, where R is a random number with uniform distribution on the interval [0,1]. Then, by using (9), we can get n_{d1}, n_{d2}, n_{d3} and n_{d4}.

In our assessment of whether SAT is robust to population stratification, we generate 100 sets of n_{ci} and n_{di}, $i = 1, \ldots, 4$. For each set of n_{ci} and n_{di}, $i = 1, \ldots, 4$, we independently generate marker data 100 times for every marker among the 130 markers mentioned above, i.e., we generate $100 \times 130 = 1.3 \times 10^4$ markers that have no association with disease phenotype. For each set of n_{ci} and n_{di}, $i = 1, \ldots, 4$, we perform statistical tests for each of the 1.3×10^4 markers.

To compare the power of SAT with other statistical tests, we systematically assign the trait locus across the 130 markers. 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. Let $f_{11} = \alpha$, $f_{22} = \beta$, $f_{12} = \gamma$ ($\gamma = \alpha$ or β corresponds to a dominant or recessive disease model), and $R_A = \alpha/\beta$. For a given R_A value and mode of inheritance, the proportions of affected individuals with genotypes AA, Aa and aa can be easily calculated. In our simulations, we vary the values of R_A and disease models. In addition, we either assume all subpopulations have the same high-risk allele or allow each individual subpopulation to have the high-risk allele chosen randomly according to its allele frequency in that subpopulation.

For every marker among the 130 markers, we independently generate the marker alleles B times according to the allele frequencies. Therefore, we obtain alleles in a set of $130 \times B$ markers, and use these $130 \times B$ markers to infer the components of similarities from a sample of individuals. For power comparisons, we fix $n_{ci} = n_{ci}^0$, $i = 1, \ldots, 4$ and $R = 1$. We independently generate the trait locus data at each of the 130 markers 1,000 times, i.e., we generate data at 1.3×10^5 trait loci.

3.3. Other association tests compared with the SAT

In our simulation studies, we compare SAT with four other statistics. The first three statistics are all applicable to case-control studies whereas the fourth test statistic requires a different design. The first statistic is the Pearson statistic at (1) which ignores potential population structure. This statistic can be quite sensitive to population heterogeneity and may lead to false-positive rates much higher than the nominal level. The second statistic, T_m, is similar to SAT except that the true pairwise relationship is used in the analysis, i.e., the W_{ij} are assigned their true values instead of being estimated from independent markers. This test represents an ideal case and the maximal power we may achieve with our procedure. The third test, STRAT, was proposed by Pritchard, Stephens, Rosenberg and Donnelly (2000).

The fourth test statistic compared is the TDT proposed by Spielman, McGinnis and Evens (1993). To apply this test statistic, we need to collect parents-child triads instead of a sample of unrelated cases and controls. For each family triad, we first generate genotypes of the affected individuals and then generate the genotypes of their parents in our simulations. The prevalence of the disease is specified under each coalescent model. When empirical population genetics data are used in our simulations, we assume the prevalence of the disease in the Danish population is 1% and the prevalence in other populations can be calculated using the relative risks among the subpopulations. In our simulations, we keep the total sample the same between the case-control design and the TDT design.

4. Results

4.1. The Accuracy of clustering for similarities

Because the SAT will only perform well if the similarities can be clustered correctly, we first assess the performance of the clustering method in our simulations. The results are summarized in Table 2 for coalescent models with different sample sizes from two subpopulations, and from the second subpopulation only.

Table 2. The performance of the similarity-based clustering method under coalescent models for different sample sizes based on 500 unlinked markers and 100 replications. T is the divergence time of the two subpopulations in generation (for the case of one subpopulation, all sampled individuals come from the second population). MK is the percentage of the 100 replications in which the number of similarity components is correctly estimated. MSS is the mean proportion of the individual pairs that are assigned the correct relationship, i.e., whether they are in the same subpopulation or in different subpopulations in 100 replications. Among all sampled individuals, there is an equal number of affected and normal individuals.

Sample size	T=500			T=1500		
	Two populations		One population	Two populations		One population
	MK	MSS	MK	MK	MSS	MK
50	100%	96.00%	100%	100%	99.76%	100%
100	100%	96.17%	100%	100%	99.82%	100%
150	100%	96.58%	100%	100%	99.82%	100%
200	100%	96.53%	100%	100%	99.83%	100%
250	100%	96.37%	100%	100%	99.82%	100%
300	100%	96.31%	100%	100%	99.82%	100%
350	100%	96.56%	100%	100%	99.83%	100%
400	100%	96.71%	100%	100%	99.83%	100%

It is easy to see that the estimation of the number of similarity groups is very reliable using the BIC criterion. In the presence of two subpopulations, more than 96% and 99.76% of all pairwise relationships are correctly inferred for the case of $T = 500$ and $T = 1500$ respectively. The results for $T = 4500$ are not shown, but it is apparent that the two subpopulations are more easily distinguished in this case. It is clear that the results are stable for varying sample sizes. Although there are some misclustered similarities, this has little effect on the type-I error rates. Table 3 summarizes the results of similarity clustering using empirical population genetics data by using different numbers of independent markers for a sample of 100 cases and 100 controls. When 520 independent markers are used to make inference on pairwise relationships, only 2.07% of the pairs are misclassified. This proportion reduces to 0.31% when 1040 markers are used for similarity inferences.

Table 3. The performance of the similarity-based clustering method using empirical population genetics data. $\hat{K}$ is the estimated number of similarity components and SS is the proportion of the individual pairs that are assigned the correct relationship, i.e., whether they are in the same subpopulation or in different subpopulations. The results are based a sample consisting of 100 affected individuals and 100 normal individuals.

Number of Loci	Four Populations		One Population			
			Biaka	Danes	Chinese	Maya
	$\hat{K}$	SS	$\hat{K}$	$\hat{K}$	$\hat{K}$	$\hat{K}$
520	2	97.93%	1	1	1	1
780	2	99.14%	1	1	1	1
1040	2	99.69%	1	1	1	1

When we apply the clustering method developed by Pritchard, Stephens and Donnelly (2000) using the same set of genotype data, the number of subpopulations estimated from their method tends to overestimate the true number of subpopulations. For example, under Model A with a population divergence time of 500 generations, the log-likelihoods of the data given the number of the populations are -237289, -237276, and -237295 for $G = 2$, 3, and 4, respectively. If we choose a uniform prior on G, the estimated number of subpopulations should be three instead of two. For the simulated data based on empirical population genetics data with 780 markers, the estimated number of populations is six by the method of Pritchard, Stephens and Donnelly (2000). Although an overestimated number of subpopulations may not induce spurious association due to population stratification, it will nonetheless reduce the power of the test.

4.2. Type-I error

Under coalescent models, we simulate a large number of independent biallelic candidate loci under the null hypothesis of no disease-marker association and compare the type-I error rates among three test statistics: the Pearson chi-square statistic T_p, SAT and T_m. The results summarized in Table 4 for three statistical significance levels (0.05, 0.01 and 0.001) are based on 10^5 replications. The standard errors for the type-I error rate estimates are $\sqrt{0.05 \times 0.95/10^5} \approx 6.9 \times 10^{-4}$, $\sqrt{0.01 \times 0.99/10^5} \approx 3.14 \times 10^{-4}$ and $\sqrt{0.001 \times 0.999/10^5} \approx 1 \times 10^{-4}$ for the true error rates 0.05, 0.01 and 0.001, respectively. It is easy to see that the estimated type-I error of SAT and T_m are not significantly different from the nominal levels. For T_p, the type-I error rate can be much higher than the nominal level in the presence of population stratification. The three significance levels considered here are appropriate if a candidate gene is studied. In practice, a more stringent criterion is needed if a genome-wide search is performed. We have also performed simulations for significance levels of 10^{-5} and 10^{-6} using 2×10^7 replications and the results show the same pattern.

Table 4. Type-I error rates in pecentage for the Pearson chi-square statistic T_p, the SAT, and the T_m under coalescent models. In the table, T is the number of generations since two subpopulations divided and RR is the relative risk between the two subpopulations. The results are based on 10^5 replications. For each replication, the sample consists of 100 affected individuals and 100 normal individuals. A total of 500 independent markers are used to make inference on the population structure.

			P=5%			P=1%			P=0.1%		
Model	T	RR	T_p	SAT	T_m	T_p	SAT	T_m	T_p	SAT	T_m
A	500	1	4.51	4.91	5.01	0.85	0.97	0.98	0.069	0.087	0.088
		4	20.4	5.05	5.01	8.77	1.03	0.98	2.421	0.101	0.094
		7	34.6	5.02	5.04	20.2	0.98	1.01	8.893	0.102	0.094
	1500	1	4.44	4.93	4.94	0.87	0.97	0.97	0.084	0.092	0.095
		4	36.5	4.92	4.95	20.9	0.97	0.96	9.073	0.102	0.092
		7	56.4	5.01	5.03	41.5	0.98	0.99	25.64	0.091	0.093
	4500	1	4.34	4.94	4.99	0.77	0.97	0.99	0.055	0.093	0.094
		4	46.1	5.02	5.05	31.3	0.98	0.98	17.24	0.095	0.093
		7	66.2	4.99	5.04	52.0	0.97	0.98	37.25	0.094	0.094
B	4500	1	4.30	5.02	5.10	0.85	1.04	1.03	0.075	0.103	0.110
		4	32.1	4.93	4.94	18.0	0.97	1.03	8.34	0.102	0.103
		7	52.2	4.99	4.98	36.6	1.02	1.02	23.12	0.109	0.108

Using empirical population genetics data the results on the type-I error rates, by using different number of independent markers to infer the similarities, are summarized in Table 5. When $4 \times 130 = 520$ independent markers are used, the type-I error rates of SAT are slightly higher than the nominal error rates, and the type-I error rates are at the nominal rates when $8 \times 130 = 1040$ independent markers are used to infer the similarities. On the other hand, the type-I error rates of T_p are much higher than the nominal error rates based on empirical population genetics data.

Table 5. Type-I error rates in percentage for the Pearson chi-square statistic T_p, the SAT, and the T_m using empirical population genetics data. Different numbers of markers are used to make inference on the population structures. Results are based a sample of 100 affected individuals and 100 normal individuals.

Number	P=5%			P=1%			P=0.1%		
of Loci	T_p	SAT	T_m	T_p	SAT	T_m	T_p	SAT	T_m
520	15.5	5.7	5.1	6.5	1.2	0.97	2.2	0.13	0.110
780	15.6	5.5	5.1	6.5	1.2	1.0	2.3	0.11	0.096
1040	15.4	5.1	5.1	6.4	1.0	0.98	2.2	0.10	0.093

4.3. Power comparisons

Under each coalescent model, we compare the power of SAT to that of TDT and T_m. The results are summarized in Table 6 and Table 7 for recessive and dominant disease models, respectively. We assume physical distances of $50 \sim 100$ kb, $20 \sim 50$ kb, and $10 \sim 20$ kb, respectively, between the candidate marker and the disease gene for population divergence times of 500, 1500 and 4500 generations between the two subpopulations. It can be seen from these two tables that the power of SAT and T_m are almost identical in all cases, which implies that there is almost no loss of information by the SAT procedure. In addition, both SAT and T_m are more powerful than TDT for all cases considered. We also compare the power of SAT with that of STRAT under a subset of simulation models because the STRAT test is computationally intensive. In addition, we assume the true number of subpopulations is known although STRAT tends to overestimate the number of populations. The results (data not shown) suggest that SAT is slightly more powerful than STRAT under Model A, and the two tests have almost identical power under Model B.

The power comparisons of between SAT, TDT, T_m and STRAT based on empirical population genetics data are summarized in Table 8. We use 8 ×

130 = 1040 markers to infer the similarities in SAT and to infer the population structures in STRAT. We assume that the number of subpopulations is known for STRAT. It can be seen from this table that SAT and T_m have almost the same power and both are more powerful than TDT in all of the cases considered. When we assume the same high-risk allele across all subpopulations, SAT is more powerful than STRAT, even when the number of subpopulations is given at the true value for STRAT. When we allow different populations to have different high-risk alleles, STRAT is more powerful because it considers one population as a unit, whereas SAT and TDT pool information across populations by assuming the same high-risk allele is shared by all subpopulations.

Table 6. Power comparisons of the three association tests (SAT, T_m and TDT) under coalescent models for recessive diseases. The results are based on 10^4 replications with each replication consisting of 100 affected individuals and 100 normal individuals. T is the number of generations since population divergence and RR is the relative risk for the two subpopulations.

			Dist	P=0.05			P=0.01		
model	T	RR	(kb)	TDT	SAT	T_m	TDT	SAT	T_m
A	500	1	50	0.777	0.884	0.881	0.701	0.814	0.811
			100	0.618	0.733	0.728	0.493	0.628	0.623
		4	50	0.756	0.879	0.879	0.652	0.797	0.799
			100	0.599	0.725	0.724	0.469	0.604	0.603
	1500	1	20	0.795	0.882	0.881	0.683	0.804	0.802
			50	0.553	0.668	0.664	0.448	0.555	0.551
		4	20	0.757	0.868	0.867	0.677	0.781	0.781
			50	0.556	0.652	0.651	0.432	0.540	0.538
	4500	1	10	0.637	0.739	0.738	0.545	0.647	0.646
			20	0.466	0.542	0.540	0.350	0.436	0.434
		4	10	0.815	0.724	0.724	0.730	0.624	0.624
			20	0.428	0.529	0.528	0.342	0.424	0.424
B	4500	1	10	0.584	0.698	0.695	0.482	0.598	0.597
			20	0.443	0.526	0.525	0.328	0.425	0.423
		4	10	0.560	0.675	0.674	0.487	0.590	0.590
			20	0.410	0.490	0.490	0.330	0.390	0.390

Table 7. Power comparisons of the three association tests (SAT, T_m and TDT) under coalescent models for dominant diseases. The results are based on 10^4 replications with each replication consisting of 100 affected individuals and 100 normal individuals. T is the number of generations since population divergence and RR is the relative risk for the two subpopulations.

			Dist	P=0.05			P=0.01		
model	T	RR	(kb)	TDT	SAT	T_m	TDT	SAT	T_m
A	500	1	50	0.517	0.577	0.576	0.374	0.448	0.444
			100	0.372	0.431	0.433	0.254	0.308	0.306
		4	50	0.514	0.573	0.575	0.365	0.440	0.442
			100	0.375	0.430	0.432	0.248	0.299	0.302
	1500	1	20	0.514	0.596	0.588	0.374	0.463	0.456
			50	0.313	0.391	0.382	0.201	0.260	0.253
		4	20	0.513	0.591	0.590	0.388	0.464	0.465
			50	0.311	0.391	0.392	0.211	0.271	0.271
	4500	1	10	0.450	0.489	0.489	0.318	0.374	0.376
			20	0.306	0.325	0.320	0.189	0.217	0.213
		4	10	0.428	0.484	0.484	0.323	0.370	0.372
			20	0.330	0.345	0.343	0.212	0.244	0.243
B	4500	1	10	0.433	0.472	0.473	0.348	0.360	0.360
			20	0.287	0.302	0.304	0.180	0.208	0.207
		4	10	0.415	0.468	0.468	0.320	0.350	0.352
			20	0.291	0.321	0.320	0.221	0.232	0.235

In our simulations, we keep the total number of subjects the same between the case-control design and the TDT design. However, we should note that a large number of genomic markers have to be typed in the application of SAT to avoid spurious association caused by population stratification, whereas only genotypes from candidate markers are required for TDT designs. In addition, there is no need to collect controls in the TDT design. In practice, the best design depends heavily on the trait being studied. For example, for diseases where parental information is relatively easy to collect, e.g., autism, the TDT design may be preferred. On the other hand, if the parental information and/or relative information is difficult to obtain, e.g., drug dependence, the case-control design may be the only feasible and powerful approach to identifying susceptibility genes.

Table 8. Power comparisons of the four association tests (SAT, T_m, TDT, and STRAT) using empirical population genetics data. The sampling of the affected individuals and normal individuals from each of the four subpopulations is discussed in the text. R_A denotes the relative attributable risk of the different genotypes. The results are based on 1000 replications on 130 markers from ALFRED and $8 \times 130 = 1040$ independent markers are used to make inference on the population structures (The fixed ancestral allele model assumes that the high-risk alleles are the same across all subpopulations; The random ancestral allele model assumes that we assign a high-risk allele according to the allele frequency in normal individuals, independently in each subpopulation).

Ancestral	Disease		P=0.05				P=0.01			
allele	model	R_A	TDT	SAT	T_m	STRAT	TDT	SAT	T_m	STRAT
fixed	recessive	4	0.313	0.443	0.444	0.380	0.15	0.254	0.259	0.200
		8	0.708	0.846	0.848	0.770	0.523	0.721	0.724	0.625
	dominant	4	0.395	0.547	0.540	0.525	0.220	0.387	0.377	0.345
		8	0.526	0.659	0.654	0.641	0.375	0.531	0.532	0.520
random	recessive	4	0.198	0.283	0.287	0.375	0.088	0.147	0.150	0.200
		8	0.379	0.530	0.532	0.766	0.242	0.390	0.392	0.618
	dominant	4	0.342	0.460	0.461	0.520	0.185	0.299	0.302	0.340
		8	0.385	0.510	0.511	0.591	0.225	0.337	0.335	0.36

5. Discussion

Although traditional case-control studies may be subject to bias caused by population stratification, alternative methods such as family-based association designs (Spielman, McGinnis and Evens (1993)) may be less powerful due to overmatching between cases and controls (Risch (2000)). Furthermore, case-control studies have the advantages of easy sample collection and possibly simple genetic analysis. Recently, Devlin and Roeder (1999), Pritchard, Stephens, Rosenberg and Donnelly (2000), Reich and Goldstein (2001), and Satten, Flanders and Yang (2001) have described statistical methods for an association test in structured populations that may be robust to population stratification.

In this article, we have developed an alternative method, the similarity-based association test (SAT), to detect association between a candidate marker and a disease of interest using case-control designs. We first infer whether two individuals are from the same subpopulation or from different subpopulations using genotype data at a series of independent markers. We then perform an association test using SAT by comparing within subpopulation allele-frequency differences between the cases and controls. We are testing for a single SNP while using all other SNPs to classify subpopulations. In practical studies, what is more realistic is that many candidate SNPs are tested and a set of independent markers are used to classify subpopulations. Simulation results show that our test has correct type-I error rate even in the presence of population stratification. In contrast with STRAT proposed by Pritchard, Stephens, Rosenberg and Donnelly (2000), SAT does not estimate the number of the subpopulations. Instead, we need only estimate the number of similarity groups. Usually a choice is made between one or two groups, resulting in a simpler problem than estimating the number of subpopulations. Furthermore, statistical inference for SAT is based on an asymptotic distribution, making statistical significance assessment less computationally demanding than for STRAT. Our simulation results show that the power of SAT is higher than STRAT when the high-risk allele is the same across all the subpopulations, even if we assume that the number of subpopulations is known when we use STRAT. On the other hand, if we allow different populations to have different high-risk alleles, STRAT is more powerful than SAT. However, it is not clear which scenario is more likely for genes responsible for complex traits. To address this problem, we may construct alternative statistics to detect different allele effects in each subpopulation. For example, by using the notation introduced previously, we may consider

$$T_d = \sum_{i,j} \sum_{i_1,j_1} (x_i - y_j) B_{ij} (x_{i_1} - y_{j_1}) B_{i_1 j_1} D_{i i_1} N_{j j_1},$$

and use simulation procedures to assess the statistical significance level of the test.

Both SAT and STRAT rely on correct inference of the population structure using a set of independent markers. Pritchard, Stephens, Rosenberg and Donnelly (2000) suggested that more than 100 microsatellite loci should be used for STRAT. Based on empirical population genetics data, our experience suggests that 500 to 1000 SNPs are needed to make accurate inference of the similarities used by SAT. If the two subpopulations are very similar, it may be the case that two clusters (the within-population cluster and the between-population cluster) of the similarities cannot be distinguished from each other with 500 to 1000 SNPs. However, the false-positives due to population stratification are likely to be small and the spurious association is not a severe problem in this case. In our simulations (results not shown), we have considered two subpopulations divided

100 generations ago. In this case, the estimated number of similarity groups by the BIC criterion is one, and the estimated type-I errors of SAT (assuming a relative risk of two) are 6% and 1.2% for nominal levels 5% and 1%, respectively, based on 10^5 replications. However, if it is affordable to type a large number of SNPs, we suggest use of more independent markers.

Although we have focused our discussion on SNPs, microsatellite markers may be more informative to identify population structures. The reason that we develop SAT for SNP markers here is that we think the abundance of SNPs in the human genome may lead more research groups to examine SNPs in genetic association studies in the future. We are currently exploring ways to extend the SAT methodology to microsatellite markers and will report our results in the future. Because the methods discussed in the paper only handle biallelic markers and we do not have access to a large number of SNP marker allele frequencies in many populations, we have to pool the alleles from microsatellite markers to form biallelic markers in our assessments of various methods using empirical population genetics data. However, if microsatellite markers are available in genetic studies, they should be directly used to increase the ability to identify population structures.

Although several methods, including ours, have been proposed to use genomic markers to control for population stratification, there has not been a published study utilizing this idea. The power and usefulness of this new approach, as well as the relative performance of various methods, need to be validated and compared through real studies.

Acknowledgements

We thank an anonymous referee for helpful comments. This work was supported in part by grants GM59507, GM57672, HD36834 from the National Institutes of Health and Research Grant FY98-0752 from the March of Dimes Birth Defects Foundation.

References

Akaike, H. (1974). A new Look at the Statistical Identification Model. *IEEE Trans. Auto. Control* **19**, 716-723.

Biernacki, C. and Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *J. Stat. Comput. Simulation* **64**, 49-71.

Biernacki, C., Celeux, G. and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern. Recogn. Lett.* **20**, 267-272.

Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Model. *Pattern Recognition* **28**, 781-793.

Cheung, K. H., Osier, M. V. and Kidd, J. R. et al. (2000). ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. *Nuclear Acids. Res.* **29**, 361-363.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. and Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**, 682-689.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997-1004.

Ewens, W. J. and Spielman, R. S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *Amer. J. Hum. Genet.* **57**, 455-464.

Falk, C. T. and Rubinstein, P. (1987). Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**, 227-233.

Fu, Y. X. and Li, W. H. (1999) Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoretical Population Biology* **56**, 1-10.

Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L. and Feldman, N. W. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* **92**, 6723-6727.

Griffiths, R. C. and Tavaré, S. (1994). Ancestral inference in population genetics. *Statist. Sci.* **9**, 307-319.

Griffiths, R. C. and Tavaré, S. (1997). Computational methods for the coalescent. In *Progress in Population Genetics and Human Evolution* (Edited by S. Tavaré and P. Donnelly), 165-182. IMA Volume 87, Springer-Verlag.

Kang, A. M., Palmatier, M. A. and Kidd, K. K. (1999). Global variation of a 40-bp VNTR in the 3'-untranslated region of the dopamine transporter gene (SLC6A3). *Biological Psychiatry* **46**, 151-160.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Process Appl.* **13**, 235-248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Probab.* **19**, 27-43.

Knowler, W. C., Williams, R. C., Pettitt, D. J. and Steinberg, A. G. (1988). Gm3-5,13,14 and type-2 diabetes mellitus: An association in American-Indians with genetic admixture. *Amer. J. Hum. Genet.* **43**, 520-526.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (Edited by D. Futuyma and J. Antonovics), 144. Oxford University Press, Oxford.

Lazzeroni, L. C. and Lange, K. (1998). A conditional inference framework for extending the transmission/disequilibrium test. *Hum. Hered.* **48**, 67-81.

Morton, N. E. and Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proc. Natl. Acad. Sci. USA* **95**, 11389-11393.

Pritchard, J. K. and Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Amer. J. Hum. Genet.* **65**, 220-228.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000). Association mapping in structured population. *Amer. J. Hum. Genet.* **67**, 170-181.

Reich, E. E. and Goldstein, D. B. (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. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**,1516-1517.

Risch, N. and 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, G. A., Flanders, W. D. and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Amer. J. Hum. Genet.* **68**, 466-477.

Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Amer. J. Hum. Genet.* **62**, 450-458.

Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Amer. J. Hum. Genet.* **52**, 506-513.

Teng, J. and 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, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N. P., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M. and Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077-1082.

Department of Mathematical Science, Michigan Technological University, Houghton, MI, U.S.A.

E-mail: shuzhang@mtu.edu

Department of Genetics, Yale University School of Medicine, New Haven, CT, U.S.A.

E-mail: kidd@biomed.med.yale.edu

Departments of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, U.S.A.

E-mail: hongyu.zhao@yale.edu

(Received March 2001; accepted October 2001)