Am. J. Hum. Genet. 62:703–712, 1998

Marker Selection for the Transmission/Disequilibrium Test, in Recently Admixed Populations

N. L. Kaplan,[1] E. R. Martin,[1,2] R. W. Morris,[3,4] and B. S. Weir[2]

[1]Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC; Programs in [2]Statistical Genetics and [3]Biomathematics, Department of Statistics, North Carolina State University, Raleigh; and [4]Analytical Sciences, Inc., Durham

Summary

Recent admixture between genetically differentiated populations can result in high levels of association between alleles at loci that are $\leqslant 10$ cM apart. The transmission/disequilibrium test (TDT) proposed by Spielman et al. (1993) can be a powerful test of linkage between disease and marker loci in the presence of association and therefore could be a useful test of linkage in admixed populations. The degree of association between alleles at two loci depends on the differences in allele frequencies, at the two loci, in the founding populations; therefore, the choice of marker is important. For a multiallelic marker, one strategy that may improve the power of the TDT is to group marker alleles within a locus, on the basis of information about the founding populations and the admixed population, thereby collapsing the marker into one with fewer alleles. We have examined the consequences of collapsing a microsatellite into a two-allele marker, when two founding populations are assumed for the admixed population, and have found that if there is random mating in the admixed population, then typically there is a collapsing for which the power of the TDT is greater than that for the original microsatellite marker. A method is presented for finding the optimal collapsing that has minimal dependence on the disease and that uses estimates either of marker allele frequencies in the two founding populations or of marker allele frequencies in the current, admixed population and in one of the founding populations. Furthermore, this optimal collapsing is not always the collapsing with the largest difference in allele frequencies in the founding populations. To demonstrate this strategy, we considered a recent data set, published previously, that provides frequency estimates for 30 microsatellites in 13 populations.

Received October 3, 1997; accepted for publication January 15, 1998; electronically published February 25, 1998.

Address for correspondence and reprints: Dr. Norman Kaplan, Biostatistics Branch, NIEHS, P.O. Box 12233, Research Triangle Park, NC 27709. E-mail: norm@seth.niehs.nih.gov

0002-9297/98/6203-0026$02.00

Introduction

Evidence of association does not always imply that two loci are linked, since association can occur between alleles at unlinked loci in the presence of forces such as selection or population admixture. To support the conclusion that a marker associated with a disease is physically close to a disease-susceptibility locus, evidence of linkage also is needed. For complex diseases, parametric linkage analysis using pedigree data may not be reliable, because the genetic model for the disease is unknown. Affected-sib-pair tests often are used to detect linkage, but, without large sample sizes, these methods have little power to identify susceptibility loci having small effects (Cox and Spielman 1989). Spielman et al. (1993) proposed an alternative approach that uses family data and that examines marker allele-transmission patterns from parents to affected children. Their transmission/disequilibrium test (TDT) has the power to detect linkage if there is association and therefore is ideally suited, as a test of linkage, to follow up a positive test for association, for those diseases for which family data are available. For late-onset diseases such as Alzheimer, the TDT is not useful, because parental data typically are missing.

Populations with a history of recent admixture between genetically differentiated groups can be enriched for association; for example, alleles at loci that are $\leqslant 10$ cM apart can have appreciable association in the initial generations following admixture (Chakraborty and Weiss 1988). This raises the possibility that, for these special populations, the TDT might be a powerful test for detection of linkage to susceptibility genes for diseases for which parental data are available. A number of authors have studied recently admixed populations (Chakraborty and Weiss 1988; Risch 1992; Stephens et al. 1994), but their focus was on detecting association, not linkage. McKeigue (1997) explored the use of the TDT in admixed populations, to identify linkage to susceptibility loci. He assumed that suitable markers that have alleles for which descent can be assigned uniquely to one of the founding populations can be identified throughout the genome. A more probable scenario is

that investigators will use published sets of markers that have been chosen for optimal screening of the human genome (Sheffield et al. 1995; Yuan et al. 1997) and that these markers are likely to have alleles that are in both founding populations. Therefore, answers to questions of descent are likely to be ambiguous.

Increased association between loci in an admixed population results only when the allele frequencies for each locus in the founding populations are different (Chakraborty and Weiss 1988). Since disease genes are unknown, disease penetrances usually are assumed to be the same in the founding populations, and, consequently, the frequencies of alleles at the disease locus are different between high- and low-risk populations. Hence, admixed populations are most useful for the study of those diseases with a large relative risk (i.e., ratio of disease prevalences) between the founding populations (Risch 1992; McKeigue 1997). Marker selection also is critical if one hopes to find increased association in admixed populations. Chakraborty and Weiss (1988) and Stephens et al. (1994) showed that two-allele markers with large differences in allele frequencies, in the founding populations, lead to powerful χ^2 tests of association. McKeigue's (1997) recommendation for the TDT is an extreme case in which a marker allele has a frequency of 1.0 in one founding population and of 0 in the other founding population.

To date, most sets of markers used for screening of the human genome are multiallelic microsatellites (Sheffield et al. 1995; Yuan et al. 1997), and investigators are beginning to use a genome scan to identify markers that may be linked to susceptibility genes for complex diseases (Copeman et al. 1995; Sawcer et al. 1996). Recently, Spielman and Ewens (1996) proposed a multiallelic version of the TDT, and its properties were studied by Kaplan et al. (1997). Additional transmission-based test statistics for multiallelic markers have been proposed by Sham and Curtis (1995) and Terwilliger (1995). In general populations, alleles at two linked loci will seldom be associated, unless their genetic map distance is low, typically <1 cM (Bodmer 1986). Hence, the TDT is expected to have little statistical power, unless the markers in the scan are very dense. For these populations, the TDT is not a recommended test of linkage, for 5–10-cM genome scans.

Since the increased association in admixed populations depends on the marker, the question of which markers to choose for a genome scan is important. We first consider a multiallelic marker and argue that it may be advantageous to group alleles on the basis of information about the founding populations, thereby collapsing the marker into one with fewer alleles. There are many ways to group alleles, and, therefore, a method is needed for the identification of a good collapsing. The criterion we propose is based on a result from the study by Kaplan et al. (1997), which shows that, for randomly mating populations, the power of the TDT for an m-allele marker can be approximated from a noncentral χ^2 distribution with $m - 1$ df and a noncentrality parameter that can be estimated from case-control data.

The results of the study by Kaplan et al. (1997) are based on the assumption of random mating, which may or may not hold for the admixed population under study. If the departure from random mating is not large, then it is reasonable to assume that results based on random mating should not lead one too far astray. In particular, if the amount of admixture in each generation is low, such as in the African American population, then it may be reasonable to assume that random mating holds in the population. Examination of marker data from African American samples supports the random-mating assumption (Weir 1992; Maiste and Weir 1995).

A good collapsing strategy lowers the df as much as possible while reducing the noncentrality parameter as little as possible. The maximum collapsing is to two alleles, and, therefore, we consider this case only. In this article, we consider an admixed population with two founding populations and develop a general method to identify the collapsing, to two allelic classes, that minimizes the reduction of the noncentrality parameter. With this method, which is minimally dependent on the disease, estimates of marker allele frequencies in the founding populations can be used. When there is limited information about one of the founding populations, such as for the African American population, the method can be modified for use of estimates of marker allele frequencies in the admixed population, instead of in the founding populations.

Jorde et al. (1995) published allele-frequency data for 30 microsatellite markers in 13 different populations. To demonstrate the potential benefit of the collapsing strategy, the power of the TDT for the optimal collapsed biallelic marker was calculated for each of the 30 microsatellites, assuming that the power of the TDT was 50% for each noncollapsed microsatellite. When association was tested, by use of case-control data, the optimal collapsed two-allele marker was the one with the largest difference in allele frequencies in the founding populations (Stephens et al. 1994). For comparison, we also calculated power by using this collapsing.

Even for the optimal collapsed two-allele marker, the sample size required for a desired power may be too large for practical purposes. McKeigue (1997) calculated the number of informative case-parent pairs for a specified power, for various genetic models, under a simple admixture model and very idealized assumptions about the markers. His results would not apply to the African American population; thus, for illustrative purposes, we have performed sample-size calculations for such a population.

Methods

We begin by reviewing properties of $T_{m\text{het}}$, the multiallelic version of the TDT statistic (Spielman and Ewens 1996; Kaplan et al. 1997). Suppose a marker locus has m alleles, $M_1, M_2, \ldots, M_m$, with frequencies $q_1, q_2, \ldots, q_m$, and the disease locus has two alleles, D_1 (disease allele) and D_2 (normal allele), with frequencies p_1 and p_2. Let θ denote the recombination fraction between the marker and disease loci, and let $\Pr(M_i|D_r)$ be the probability that a gamete carries allele M_i, given that it carries allele D_r. Penetrance f_{rs} is the probability that individuals with genotype D_rD_s are affected with the disease. We assume that $f_{11} \geqslant f_{12} = f_{21} \geqslant f_{22}$, with at least one inequality strict. Finally, assuming that random mating holds in the population, $A = \sum_{rs} p_r p_s f_{rs}$ is the prevalence of the disease—that is, the probability that a random individual is affected.

Suppose that there are N_T families with one affected child, and the mother, father, and affected child are all typed at the marker. The data can be arranged in an $m \times m$ contingency table in which the rows and columns indicate transmitted and nontransmitted parental alleles, respectively. Let n_{ij} denote the number of observations in the cell for row i and column j (i.e., the number of M_iM_j parents who transmitted allele M_i to the affected child), and let $n_{i.}$ and $n_{.j}$ denote the row and column marginal sums, respectively. Spielman and Ewens (1996) proposed the following statistic as a generalization of their two-allele TDT statistic:

$$T_{m\text{het}} = \frac{m-1}{m} \sum_{j=1}^{m} \frac{(n_{j.} - n_{.j})^2}{n_{j.} + n_{.j} - 2n_{jj}} .$$

Note that the value of $T_{m\text{het}}$ only depends on transmissions from heterozygous parents. Kaplan et al. (1997) studied the statistical properties of $T_{m\text{het}}$ and showed that, for a randomly mating population, it is approximately a χ^2 statistic with $m-1$ df, under the null hypothesis. Furthermore, if A is small, then the power of the test for the alternative hypothesis of linkage can be approximated from a noncentral χ^2 distribution with $m-1$ df and the noncentrality parameter

$$2N_T(1-2\theta)^2 I^* , \qquad (1)$$

where

$$I^* = \sum_{j=1}^{m} \frac{[\Pr(M_j \mid \text{affected}) - \Pr(M_j \mid \text{unaffected})]^2}{\Pr(M_j \mid \text{affected}) + \Pr(M_j \mid \text{unaffected})} . \qquad (2)$$

The value of I^* depends on the amount of association between alleles at the marker and disease loci. In particular, if there is no association, then $I^* = 0$, and the test has no power.

Kaplan et al. (1997) showed that if there is random mating, then I^* can be written in the following form:

$$I^* = \frac{X^2}{2} \sum_j \frac{[\Pr(M_j \mid D_1) - q_j]^2}{q_j + \frac{X}{2}(1-2A)[\Pr(M_j \mid D_1) - q_j]} , \qquad (3)$$

where $X = (B - p_1 A)/p_2 A(1-A)$ and $B = f_{11}p_1^2 + f_{12}p_1p_2$. This form of I^* is easier to work with for analysis of admixed populations, since only the quantities in the sum depend on the marker. In particular, for two markers with the same θ with regard to the disease locus, the ratio of the noncentrality parameters is the ratio of these sums, since X depends only on the disease locus and is, therefore, the same for each marker. For the following analysis, it is convenient to make one additional algebraic manipulation of the sum in equation (3). It is not difficult to show that if $f_{11} \geqslant f_{12} = f_{21} \geqslant f_{22}$, then the value $\nu = X(1-2A)/2$ is between 0 and 1, and, thus, I^* can be written as

$$I^* = \frac{X^2}{2} \sum_j \frac{[\Pr(M_j \mid D_1) - q_j]^2}{q_j(1-\nu) + \nu \Pr(M_j \mid D_1)} . \qquad (4)$$

The importance of equation (4) is that the value of ν does not depend on the marker.

We assume that there are two founding populations (Chakraborty and Weiss 1988; Stephens et al. 1994) and that the dynamics of the admixed population follow model 2 in the study by Ewens and Spielman (1995). In generation 0, individuals belong to either of the founding populations, and random mating is allowed only within each population. Allele frequencies in generation 0, at the marker and disease loci in population i, are denoted as q_{ij} and p_{ir}, respectively ($j = 1, \ldots, m$; $r = 1, 2$; $i = 1, 2$). We denote the frequencies of gametes carrying M_jD_r, in population i in generation 0, as g_{ijr}. In founding populations, it is unlikely that most markers in a scan with a density of $\geqslant 5$ cM will be associated with the disease locus, so the marker and disease loci are assumed to be in linkage equilibrium in the founding populations; that is, $g_{ijr} = q_{ij}p_{ir}$, for all i, j, and r.

Generation 1 consists of the offspring of generation 0. If α is the proportion of generation 1 from population 1, then allele frequencies for the marker and disease loci in this generation are

$$q_{(1)j} = \alpha q_{1j} + (1-\alpha) q_{2j} \text{ and}$$
$$p_{(1)r} = \alpha p_{1r} + (1-\alpha) p_{2r} , \qquad (5)$$

where $1 \leqslant j \leqslant m$ and $r = 1, 2$ (the subscript integer in parentheses indicates the number of generations of admixture). Analogous expressions hold for gamete frequencies in generation 1, $\Pr_{(1)}(M_jD_1)$. Therefore,

Table 1

Parameters of the Admixture Model, Assuming Two Founding Populations

Parameter	Value(s)
Genetic model	Recessive; dominant; additive
(p_{11}, p_{21})	(.001, .02); (.001, .02); (.1, .3)
Genotype risk ratio f_{11}/f_{22}[a]	10; 100; 1,000; 10,000
θ	.05
Initial proportions of founding populations (α)	.5; .98
Migration rates (γ_1, γ_2)	(0, .02); (.01, .01)
No. of generations since admixture	10

[a] The only values permitted are those that result in a relative risk of 2–10.

$$\Pr_{(1)}(\mathrm{M}_j \mid \mathrm{D}_1) = \frac{\alpha g_{1j1} + (1-\alpha) g_{2j1}}{p_{(1)1}} = \eta q_{1j} + (1-\eta) q_{2j} , \quad (6)$$

where $\eta = \alpha p_{11}/p_{(1)1}$. Individuals in generation 1 mate at random, and their offspring are joined by the offspring of recent immigrants who mated in one of the founding populations, to form generation 2. Future generations are formed in the same way. We assume that, in each generation, the proportion of individuals in the admixed population who are new immigrants from founding population i is γ_i, where $i = 1, 2$. We make a distinction between α and γ_i, to allow for the possibility that they can be different, which, for example, is the case for the African American population. Even though in each generation the admixed population may be stratified into groups that do not intermate, we assume that most individuals in each generation are not new immigrants, and, therefore, we ignore the stratification problem and assume that the admixed population is mating randomly. For example, for the African American population, $\gamma_1 + \gamma_2 < .05$ has been estimated (Stephens et al. 1994). By use of equations (4), (5), and (6), I^* for generation 1 of the admixed population can be written as

$$I^* = \frac{X^2(\alpha-\eta)^2}{2} \sum_j \frac{(q_{1j} - q_{2j})^2}{\xi q_{1j} + (1-\xi) q_{2j}} , \quad (7)$$

where the value of parameter $\xi = (1-\nu)\alpha + \eta\nu$ is between 0 and 1 and does not depend on the marker.

Note that assumptions about the disease allele frequencies in the founding populations affect $\alpha - \eta$. In particular, $\alpha - \eta = 0$ if the disease allele frequencies in the founding populations are equal. In the Appendix, the form of I^* in equation (7) is shown to continue to hold for later generations of the admixed population, assuming random mating in the admixed population and recurrent migration. The value of the term outside the sum and that for ξ will change each generation, but the crucial point is that both quantities depend on the marker only through θ.

The sum in equation (7) simplifies if the marker has just two alleles. If $\delta = |q_{21} - q_{11}|$, then

$$\sum_j \frac{(q_{1j} - q_{2j})^2}{\xi q_{1j} + (1-\xi) q_{2j}} = \begin{cases} \dfrac{\delta^2}{[q_{11} + (1-\xi)\delta]\{1 - [q_{11} + (1-\xi)\delta]\}} , & \text{if } q_{21} > q_{11} \\ \dfrac{\delta^2}{(q_{21} + \delta\xi)[1 - (q_{21} + \delta\xi)]} , & \text{if } q_{21} < q_{11} \end{cases} . \quad (8)$$

In order to maximize the right-hand side of equation (8), δ, q_{11}, and q_{21} must be considered. By taking the derivative with respect to δ, we can show that, for fixed q_{11} and q_{21}, the right-hand side of equation (8) is an increasing function of δ. Also, if δ is fixed, then the right-hand side of equation (8) is largest when either $q_{11} + (1-\xi)\delta$ or $q_{21} + \xi\delta$ is minimized. McKeigue's (1997) recommendation is the extreme case of $\delta = 1$ and $q_{11} = 0$ or $q_{21} = 0$.

To get a sense of what values of ξ to expect, we calculated its value in two different migration schemes and many different genetic models (table 1). To find ξ, we first used the model assumptions to calculate the denominator in equation (3), for marker allele 1 (any marker allele works, because ξ does not depend on the marker allele). We then equated the resulting value to $\xi q_{11} + (1-\xi) q_{21}$ and solved for ξ. We found that ξ depends on the admixture assumptions but that it is reasonably insensitive to the genetic assumptions. This result is not unexpected if, for some i, $\Pr(\mathrm{M}_i|\mathrm{D}_1) - q_i$ is small relative to q_i in the admixed population. In this case, the denominator in equation (3), for marker allele i, is approximately equal to q_i. In the Appendix, we show that, in the kth generation ($k \geq 1$) of admixture,

$$q_{(k)i} = \alpha_k q_{1i} + (1-\alpha_k) q_{2i} , \quad (9)$$

where

$$\alpha_k = \alpha(1-\gamma_1-\gamma_2)^{k-1} + \frac{\gamma_1}{\gamma_1+\gamma_2}[1 - (1-\gamma_1-\gamma_2)^{k-1}] . \quad (10)$$

Hence, ξ should be close to α_k. For example, if $\alpha = .5$ and $\gamma_1 = \gamma_2$, then $\alpha_k = .5$ is always true; thus, a value of ξ that is near .5 would be expected. Alternatively, if $\alpha = 1$ and $\gamma_1 = 0$, then $\alpha_k = (1-\gamma_2)^{k-1}$, in the kth generation of admixture. Thus, if $\gamma_2 = .02$ and $k = 10$, which are reasonable values for the African American

population, then $\alpha_{10} = .83$; thus, in this case we would expect the value of ξ to be near .8. In effect, equation (7) shows that, for an admixed population, I^* can be written as a product of two terms—one depending on the marker and the other depending on the disease. This representation of I^* shows that the optimal collapsing is minimally dependent on the disease.

If ξ is close to α_k, then it follows from equation (9) that the denominator of the jth term in the sum in equation (7) can be replaced by $q_{(k)j}$. Furthermore, it follows from equation (9) that $(q_{1j} - q_{2j})^2 = \{[q_{(k)j} - q_{2j}]/\alpha_k\}^2$. Hence, the sum in equation (7) can be written as

$$\sum_j \frac{[q_{(k)j} - q_{2j}]^2}{q_{(k)j}} \quad . \tag{11}$$

Since the constant α_k^2 does not depend on the marker, it is absorbed into the constant in front of the sum in equation (7). In cases for which there is limited information about one of the founding populations, such as the African population for the current African American population, expression of the sum as shown in equation (11) is advantageous, since it can be estimated only with estimates of the allele frequencies in one of the founding populations and in the current admixed population. For a biallelic marker, the sum in equation (11) simplifies to $\delta^2\{q_{(k)1}[1 - q_{(k)1}]\}^{-1}$, where $\delta = q_{(k)1} - q_{(k)2}$. For any microsatellite, the collapsing that is optimal for the TDT maximizes the sum in equation (8) or equation (11), depending on the type of data available. There is no easy way to identify the collapsing appropriate for obtaining the maximum, other than to evaluate equation (8) or equation (11) for all possible collapsings. Since δ is squared in the numerator of equation (8), it seems reasonable to assume that the collapsed two-allele marker with the largest δ might be near optimal. The following simple algorithm leads to the collapsed two-allele marker with the largest δ. If $\hat{q}_{ij}$ is the estimate of the frequency of marker allele j in population i, where $j = 1, ..., m$ and $i = 1, 2$, then all alleles with $\hat{q}_{1j} < \hat{q}_{2j}$ are grouped together. The same algorithm works if the data require that equation (11) be used.

After identification of the optimal way to collapse the marker, one still must decide whether to use the original microsatellite or the collapsed biallelic marker. To do this, we estimate the ratio $I^*(2)/I^*(m)$, where $I^*(m)$ and $I^*(2)$ denote the values of I^* for the m-allele microsatellite and the collapsed two-allele marker, respectively. If the ratio is near 1, then the noncentrality parameter is not reduced by much, and the collapsed two-allele marker would be preferred. To show this, we argue as follows. For a marker with m alleles ($m \geqslant 2$) the power can be calculated from a noncentral χ^2 with $m - 1$ df and noncentrality parameter λ_{m-1}. It follows from equation (1) that

$$\lambda_1 = \frac{I^*(2)}{I^*(m)}\lambda_{m-1} \quad . \tag{12}$$

If $I^*(2)/I^*(m)$ is near 1, then decreasing the df from $m - 1$ to 1, without decreasing λ_{m-1} by much, can increase power.

To calculate the sample size necessary for analysis of a two-allele marker with a recombination fraction from the disease locus of θ, we first find the appropriate χ^2 critical value of the specified significance level. We next find the noncentrality parameter that gives the desired power and equate this parameter to $2N_T(1 - \theta)^2 I^*$. Finally, the calculation of I^* follows from equation (2) and from the recursions in the Appendix.

Results

In table 1, we list the parameters for the genetic and the migration models that we considered in this article. The parameterization of the genetic models is somewhat different from that used by McKeigue (1997), since we want to distinguish between cases in which the disease allele is rare in one of the founding populations and common in the other and in which the disease allele is common in both. On the basis of the discussion by McKeigue (1997), we considered diseases with a relative risk of 2–10. To obtain values in this range, for the relative risk, we needed to adjust the value of the genotype risk ratio f_{11}/f_{22}. Founding population 2 is assumed to be the "high-risk" population, and the prevalence of the disease in the admixed population is assumed to be low. If $\alpha = .5$, then ξ is always near the predicted value of .5, for all generations. For the different combinations of parameter values for generation 10, the range of values of ξ is .52–.58. In contrast, if α is near 1, then the values of ξ are near 1 for the early generations and move away from 1 for the later generations; for example, by generation 10, the range of values of ξ is .83–.87, which agrees quite well with the predicted value of .83, from equation (10).

Equation (12) shows that the collapsing of a microsatellite to a two-allele marker decreases the noncentrality parameter by the factor $I^*(2)/I^*(m)$. In table 2, we examine how this decrease impacts the power of the TDT for the collapsed two-allele marker, assuming that the assigned significance level is .001 and the power of the TDT for the m-allele microsatellite is .5. The results in table 2 show that the power of the TDT for the collapsed two-allele marker increases with an increase in m, and in $I^*(2)/I^*(m)$. Hence, the collapsing strategy increases the power of the TDT if m is large and $I^*(2)/I^*(m)$ is close to 1. In particular, it makes little sense to collapse a three- or four-allele marker.

To examine what values of $I^*(2)/I^*(m)$ might be encountered in practice, we considered published allele-

Table 2

Power of a Collapsed Two-Allele Marker, As a Function of $I^*(2)/I^*(m)$

	Power for $I^*(2)/I^*(m)$ =				
m	.9	.8	.7	.6	.5
3	.54	.46	.39	.30	.22
4	.61	.53	.44	.36	.27
5	.67	.59	.49	.40	.31
6	.71	.64	.54	.44	.34
7	.75	.67	.58	.47	.37
8	.78	.70	.62	.51	.40
9	.81	.73	.64	.54	.42
10	.82	.75	.67	.56	.44
11	.84	.78	.69	.59	.47
12	.87	.81	.73	.63	.51

Note.—The assigned significance level is .001, and the power of the TDT for the m-allele microsatellite is .5.

frequency data for 30 microsatellites from 13 different populations (Jorde et al. 1995). We pooled the data from the five different African populations, and for the Caucasian sample we pooled the Utah and French data. Using the algorithm described in Methods, we identified the collapsing with the largest δ, for each of the 30 microsatellites. We then found the best collapsing for $\xi = .85$, for each microsatellite, assuming 10 generations of admixture. In table 3, estimates of the ratio $I^*(2)/I^*(m)$ are given for all 30 microsatellites, for the best collapsed two-allele marker and for the collapsed two-allele marker with the largest δ. We also calculated the power of the TDT for each of these biallelic markers, assuming that the assigned significance level equals .001 and the power of the TDT for the microsatellite equals .5. We found that for 26 of the microsatellites the optimal collapsed two-allele marker was preferred, whereas the collapsed two-allele marker with the largest δ was preferred for 17 of the microsatellites. The reason for the drop in number is that the value of I^*, as noted earlier, depends on δ and on the allele frequencies. We also performed the same analysis by assuming that we had data from the Caucasian population and from the current admixed population. The data for the admixed population was calculated by use of the data from the two founding populations, $\xi = .83$ and $k = 10$. We obtained essentially the same results, suggesting that use of estimates of allele frequencies in the current admixed population and in the Caucasian population would have worked just as well. For 17 microsatellites, the power of the TDT for the collapsing with the largest δ was essentially the same as that for the collapsing with the largest I^*. However, for nine microsatellites, the power of the TDT for the collapsing with the largest δ was $<.5$, whereas the power for the optimal collapsed two-allele marker was $>.5$. The optimal collapsing performed badly for marker D9S249 only. However, this locus has only three alleles, and the potential improvement, in power, from collapsing is small (see table 2). In contrast, the power of the TDT for the collapsing with the largest δ was $<.3$ for five markers other than D9S248.

Until now, we have discussed only relative power. It is possible that, even after collapsing a microsatellite to a two-allele marker, we still would not have sufficient sample size for a powerful test. To investigate this issue for a population such as the African American population, we calculated the sample size needed to achieve 80% power for a biallelic marker, at an assigned significance level, for each individual test of .001. For this significance level and power, we required a noncentrality parameter of ~17. The value of θ was set at .05, which is appropriate for a 10-cM map. We assumed a migration model appropriate for the African American population—namely, $\alpha = 1$, $\gamma_1 = 0$, $\gamma_2 = .02$, and $k = 10$—and we considered each of the three genetic models in table 1. The disease allele frequencies were assumed to be .001 in the low-risk population and .5 in the high-risk population. Thus, the disease gene is very common in the high-risk population and virtually absent from the low-risk population. We also considered two other scenarios, (.001, .02) and (.2, .7), and our conclusions were essentially the same, so we did not include these calculations. Three hypothetical diseases were considered: one with a high relative risk of 10, one with a medium relative risk of 5, and one with a low relative risk of 2. The genotype risk ratio was determined for each disease, to give the desired relative risk. Sample size was calculated for three types of biallelic markers. The first type was polymorphic in both populations, and the difference in allele frequencies was not very large. The difference in allele frequencies for the second type of marker also was not very large, but the marker was essentially monomorphic in the low-risk population. Finally, the third type of marker was also essentially monomorphic in the low-risk population, but the difference in allele frequencies was large. This last type of marker is the type that McKeigue (1997) considered.

For table 4, the frequencies of M_1, in the low- and high-risk populations, for the examples of the three types of markers are (.1, .4), (.001, .3), and (.001, .9), respectively. The results in table 4 show that, for a relative risk of $\geqslant 10$, 80% power, for the most part, can be achieved with feasible sample sizes, regardless of which type of marker is used. If the relative risk drops to ~5, then markers of the third type are going to be needed to perform the analysis with reasonable sample sizes. With a larger difference in allele frequencies, markers of the first type can also work. For example, if the allele frequencies are (.1, .9) instead of (.1, .4), then the sample size for a recessive model drops from 5,441 to 1,105. Our results suggest that, in the African American population, diseases with a relative risk as small as 2 require

Table 3

Microsatellite Data from the Study by Jorde et al. (1995)

	Optimal Collapsed Two-Allele Marker			Collapsed Two-Allele Marker with Largest δ		
Locus[a]	δ	$I*(2)/I*(m)$	Power	δ	$I*(2)/I*(m)$	Power
D1S407 (9)	.17	.73	.66	.17	.73	.66
D1S399 (10)	.22	.49	.42	.22	.49	.42
D2S273 (8)	.12	.64	.55	.13	.56	.46
D3S1537 (9)	.22	.84	.76	.22	.84	.76
D3S1545 (9)	.28	.86	.77	.28	.85	.77
D4S1525 (11)	.32	.71	.69	.43	.66	.66
D4S1530 (13)	.28	.70	.73	.30	.39	.35
D5S580 (22)	.36	.80	.91	.41	.77	.90
D6S400 (11)	.49	.87	.93	.49	.87	.83
D6S393 (8)	.41	.78	.68	.58	.54	.45
D7S620 (10)	.18	.80	.76	.21	.63	.59
D7S623 (7)	.11	.80	.67	.11	.80	.67
D8S499 (13)	.20	.77	.79	.38	.40	.36
D8S384 (8)	.20	.89	.77	.26	.30	.17
D9S249 (3)	.04	.34	.11	.04	.34	.12
D9S762 (7)	.02	.68	.56	.04	.27	.13
D10S526 (18)	.21	.66	.74	.40	.54	.64
D10S516 (13)	.12	.55	.57	.26	.39	.34
D10S525 (7)	.04	.69	.57	.08	.37	.23
HRAS1 (3)	.13	.99	.59	.13	.99	.60
VWFII (9)	.23	.53	.45	.23	.53	.45
D14S119 (16)	.30	.63	.72	.38	.62	.71
D15S195 (12)	.12	.64	.64	.14	.34	.27
D16S485 (9)	.22	.63	.57	.22	.63	.57
D17S919 (11)	.37	.81	.78	.37	.81	.78
D18S4930 (8)	.07	.57	.46	.16	.35	.23
D19S403 (11)	.45	.80	.78	.45	.78	.76
D19S400 (11)	.35	.55	.53	.35	.55	.53
D20S161 (9)	.38	.85	.77	.38	.85	.77
D20S428 (10)	.17	.90	.83	.17	.90	.83

Note.—For each microsatellite, the values of δ, $I^*(2)/I^*(m)$, and the power of the TDT are given for the optimal collapsed two-allele marker and for the collapsed two-allele marker with the largest δ. The assigned significance level is .001, and the power of the TDT for the m-allele microsatellite is .5. The value of ξ is .85.

[a] The number of alleles is given in parentheses.

too large a sample and therefore cannot be studied by use of this method.

Discussion

Recently admixed populations can be enriched for association, and, therefore, the TDT can be a powerful test for linkage in these populations. The increase in association is a consequence of mixing subpopulations with differences in allele frequencies at the marker and disease loci. Under reasonable assumptions, the larger the relative risk for the disease in the founding populations the more powerful the TDT. When performing genome scans, most investigators probably will use one of the batteries of markers that are currently available (Sheffield et al. 1995; Yuan et al. 1997). Since the markers are microsatellites, it is possible that the TDT may be more powerful for collapsed two-allele markers. For the data from the study by Jorde et al. (1995), when we assumed two major founding populations and random mating in the admixed population, we found that collapsing the microsatellite in the optimal way almost always led to a more powerful test, and, therefore, we would recommend this collapsing strategy. An additional benefit of collapsing markers to two allelic classes is that it is easy to choose, from the collection of available markers, those markers with the best chance of identifying a disease locus. In particular, the best markers will maximize the sum in either equation (8) or equation (11), depending on the available data.

An important feature of the method described in this article is that it does not depend on the disease, and, therefore, it can be used when planning a study. Crucial to the analysis are marker-allele-frequency estimates for

Table 4

Sample-Size (N_T) Calculations for Populations Similar to the African American Population, under the Assumptions of $\theta = .05$ and of the Migration Model and Parameters Given in Results

Relative-Risk Ratio	Marker Type	Sample Size for Each Type of Genetic Model[a]		
		Recessive	Additive	Dominant
10	1	1,496	861	806
	2	744	451	425
	3	204	122	114
		(37)	(19)	(13)
5	1	5,441	1,789	1,528
	2	2,476	877	759
	3	699	242	209
		(17)	(9)	(6)
2	1	69,147	12,554	9,351
	2	28,978	5,501	4,146
	3	8,353	1,570	1,179
		(5)	(3)	(2)

Note.—The assigned significance level is .001, and the desired power is .8. The frequencies of marker allele M_1 in the low- and high-risk populations for three marker types are (.1, .4), (.001, .3), and (.001, .9), respectively. The frequencies of the disease allele in the low- and high-risk populations are .001 and .5, respectively.

[a] The genotype risk ratios are given in parentheses.

the founding populations. Often, one of these populations is Caucasian, and there probably are good estimates available in one of the public databases. This may not be the case for the other founding population, and, therefore, frequency estimates may have to be developed, in order for the method to be implemented. Fortunately, this is not very disease specific, and only one collection of markers is needed. An alternative strategy, which may be easier, is to estimate allele frequencies in the admixed population and to use the representation of the sum in equation (11). Our results for the data from the study by Jorde et al. (1995) suggest that this approximation will give essentially the same results. For the African American population, this approach also would be preferred, because of the multiple African subpopulations from which the founding ancestors could have originated.

Finding the optimal collapsing is not difficult but must be done numerically. An alternative strategy is to use the collapsing that has the largest δ, because there is a simple algorithm for this collapsing. For the data from the study by Jorde et al. (1995), this strategy increased the power of the test for only ~65% of the markers, which was much less than the proportion of markers for which the power was improved by the optimal collapsing (90% of the markers). In addition, the collapsing with the largest δ performed substantially worse than the original microsatellite, for several of the markers, whereas the optimal collapsing performed badly only for a single three-allele marker. The collapsing leading to the largest δ performs badly if the resulting alleles are common in both founding populations. For these cases, the optimal collapsing should be used.

When the noncentrality parameters for two different collapsings are compared, the genetic and migration parameters are all subsumed by a single parameter, ξ. Furthermore, the value of this parameter is very robust in the genetic model and depends primarily on the migration parameters. Hence, for recently admixed populations with a good historical record, reasonably good estimates of ξ, which are not very disease specific, can be obtained . For populations such as the African American population, a value of $\xi = .8$ is reasonable, when admixture is assumed to have been ongoing for ~10–15 generations. It also should be kept in mind that approximate values of ξ are all that are needed to judge whether or not to collapse the microsatellite.

The results in this article were obtained by assumption of random mating in the admixed population. If there is nonrandom mating, then the formula for the noncentrality parameter given by Kaplan et al. (1997) no longer holds. However, small departures from random mating should give similar results, whereas large departures probably would be noticed by the investigator. In the African American population, for example, stratification has not been detected in several studies of markers (Weir 1992; Maiste and Weir 1995).

Even though we made many assumptions during our investigation of sample-size requirements for two-allele markers, several of the conclusions are general and should be kept in mind when a study is being planned. Most importantly, the larger the relative risk for the disease in the founding populations, the more powerful the test. In addition, biallelic markers should have as large a δ as possible, and, for any given δ, the rarer the marker allele in the low-risk population, the better. Our results suggest that, in the African American population, diseases with large relative risk (>10) can be studied by use of biallelic markers with values of δ as small as .3. For diseases with moderate relative risk (~5), markers with a large δ are required. Finally, diseases with small relative risk (<2) appear to require too large a sample size, even for markers with a large δ. This is in contrast with the results from the study by McKeigue (1997), who obtained reasonable sample sizes for diseases with a relative risk as low as 2. One reason for the difference is that McKeigue assumed an equally admixed population—that is, $\alpha = .5$—which is not appropriate for the African American population. Also, McKeigue calculated the number of informative case-parent pairs, which can be much less than the actual sample size.

We endorse the conclusion of Briscoe et al. (1994) that, by choosing marker loci with large allele-frequency differences in the founding populations, it is possible to

exploit association in the resulting admixed population, to enhance the power of association analysis in gene mapping. We, however, have refined their conclusion to mean the choice of allelic classes for the markers at hand. Collapsing alleles to two classes, in an optimal way, can provide a substantial increase in power for the TDT, over that obtained by use of all the original marker alleles. These power studies were specifically for a test of linkage, rather than for a test of linkage disequilibrium (Stephens et al. 1994). Our study has shown that the extra analysis required to consider alternative collapsing schemes, after typing of a panel of markers, may result in increased power for detection of a disease-gene location.

Acknowledgments

This work was supported, in part, by National Institutes of Health grants GM 45344 and NS 23760.

Appendix

The proof of equation (6), for later generations of the admixed population, depends on showing that equations (4) and (5) continue to hold in each generation. The proof is by induction. For each generation, let γ_1 and γ_2 be the portions of the admixed population that are new immigrants from populations 1 and 2, respectively. Suppose that in generation k

$$q_{(k)i} = \alpha_k q_{1i} + (1 - \alpha_k) q_{2i} , \tag{A1}$$

$$p_{(k)i} = \beta_k p_{1i} + (1 - \beta_k) p_{2i} , \tag{A2}$$

and

$$P_{(k)}(\mathrm{M}_i|\mathrm{D}_r) = \eta_k q_{1i} + (1 - \eta_k) q_{2i} , \tag{A3}$$

where α_k, β_k, and η_k are constants that depend on the marker only through θ. When migration is considered, the marker allele frequencies in generation $k + 1$ can be written as

$$\begin{aligned} q_{(k+1)i} &= \gamma_1 q_{1i} + \gamma_2 q_{2i} + (1 - \gamma_1 - \gamma_2) q_{(k)i} \\ &= \alpha_{k+1} q_{1i} + (1 - \alpha_{k+1}) q_{2i} , \end{aligned}$$

where $\alpha_{k+1} = \gamma_1 + (1 - \gamma_1 - \gamma_2)\alpha_k$ and $k \geqslant 1$. This proves equation (A1), and the proof of equation (A2) is analogous. In fact, $\alpha_k = \beta_k$. The proof of equation (A3) is a little more complicated because recombination must be considered. The frequency of any haplotype $\mathrm{M}_i\mathrm{D}_r$ in generation $k + 1$ can be written as

$$\begin{aligned} \mathrm{Pr}_{(k+1)}(\mathrm{M}_i\mathrm{D}_r) = {} & \gamma_1 q_{1i} p_{1r} + \gamma_2 q_{2i} p_{2r} + (1 - \gamma_1 - \gamma_2) \\ & \times [(1 - \theta)\mathrm{Pr}_{(k)}(\mathrm{M}_i|\mathrm{D}_r) p_{(k)r} + \theta q_{(k)i} p_{(k)r}] . \end{aligned}$$

The coefficient of q_{1i} is

$$\eta_{k+1} = \gamma_1 p_{1r} + (1 - \gamma_1 - \gamma_2)[(1 - \theta)\eta_k p_{(k)r} + \theta\alpha_k p_{(k)r}] ,$$

and the coefficient of q_{2i} is

$$\begin{aligned} \eta'_{k+1} = {} & \gamma_2 p_{2r} + (1 - \gamma_1 - \gamma_2) \\ & \times [(1 - \theta)(1 - \eta_k) p_{(k)r} + \theta(1 - \alpha_k) p_{(k)r}] . \end{aligned}$$

The proof of equation (A3) follows from the observation that

$$\begin{aligned} \eta_{k+1} + \eta'_{k+1} &= \gamma_1 p_{1r} + \gamma_2 p_{2r} + (1 - \gamma_1 - \gamma_2) p_{(k)r} \\ &= p_{(k+1)r} . \end{aligned}$$

References

Bodmer WF (1986) Human genetics: the molecular challenge. In: Quantitative biology. Vol 1. Cold Spring Harb Symp Quant Biol 51:1–13

Briscoe D, Stephens JC, O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. J Hered 85:59–63

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85: 3071–3074

Copeman JB, Cucca F, Hearne CM, Cornall RJ, Reed PW, Ronningen KS, Undlien DE, et al (1995) Linkage disequilibrium mapping of type 1 diabetes susceptibility gene (IDDM7) to chromosome 2q31-33. Nat Genet 9:80–85

Cox NJ, Spielman RS (1989) The insulin gene and susceptibility to IDDM. Genet Epidemiol 6:65–69

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, et al (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60:691–702

Maiste PJ, Weir BS (1995) A comparison of tests for independence in the FBI RFLP data bases. Genetica 96:125–138

McKeigue PM (1997) Mapping genes underlying ethnic dif-

ferences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet 60:188–196

Risch N (1992) Mapping genes for complex diseases with recently admixed populations. Am J Hum Genet Suppl 51: A13

Sawcer S, Jones HB, Feakes R, Gray J, Smalden N, Chataway J, Robertson N, et al (1996) A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22. Nat Genet 13:464–468

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multiallele marker loci. Ann Hum Genet 59:323–336

Sheffield VC, Weber JL, Buetow KH, Murray JC, Even DA, Wiles K, Gastier JM, et al (1995) A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genome-wide linkage maps. Hum Mol Genet 4:1837–1844

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

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 52:506–516

Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55:809–824

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56: 777–787

Weir BS (1992) Independence tests for VNTR alleles defined as fixed bins. Genetics 130:873–887

Yuan B, Vaske D, Weber JL, Beck J, Sheffield VC (1997) Improved set of short-tandem-repeat polymorphisms for screening the human genome. Am J Hum Genet 60:459–460