Am. J. Hum. Genet. 57:455–464, 1995

The Transmission/Disequilibrium Test: History, Subdivision, and Admixture

Warren J. Ewens[1] and Richard S. Spielman[2]

[1]Department of Biology, University of Pennsylvania; and [2]Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia

Summary

Disease association with a genetic marker is often taken as a preliminary indication of linkage with disease susceptibility. However, population subdivision and admixture may lead to disease association even in the absence of linkage. In a previous paper, we described a test for linkage (and linkage disequilibrium) between a genetic marker and disease susceptibility; linkage is detected by this test only if association is also present. This transmission/disequilibrium test (TDT) is carried out with data on transmission of marker alleles from parents heterozygous for the marker to affected offspring. The TDT is a valid test for linkage and association, even when the association is caused by population subdivision and admixture. In the previous paper, we did not explicitly consider the effect of recent history on population structure. Here we extend the previous results by examining in detail the effects of subdivision and admixture, viewed as processes in population history. We describe two models for these processes. For both models, we analyze the properties of (*a*) the TDT as a test for linkage (and association) between marker and disease and (*b*) the conventional contingency statistic used with family data to test for population association. We show that the contingency test statistic does not have a χ^2 distribution if subdivision or admixture is present. In contrast, the TDT remains a valid χ^2 statistic for the linkage hypothesis, regardless of population history.

Introduction

The availability of microsatellite DNA polymorphisms has made it possible to identify markers in almost any region of the genome. This development has led to numerous studies that test for association between a disease phenotype and a DNA marker at or near a gene of interest (a "candidate" gene). The finding of an association is taken as tentative evidence that the marker is *linked* to a disease gene, perhaps implying this role for the candidate gene itself. Often the next step is to try to confirm this inference by a standard test for linkage.

It is now recognized, however, that this direct test may often fail; even when marker and disease locus are closely linked, linkage may be undetectable in conventional linkage studies (e.g., with lods or affected sib pairs [Cox et al. 1988]). This situation is illustrated by findings from the insulin gene region in insulin-dependent diabetes mellitus (IDDM) (Bell et al. 1984), where evidence from conventional linkage studies (Julier et al. 1991; Bain et al. 1992) lagged far behind evidence from association studies (Cox and Spielman 1989; Spielman et al. 1989).

Since population associations can occur even for unlinked loci, a demonstration of linkage must always supplement evidence from population association. But the example of the insulin gene region and IDDM shows that a population association may provide a more sensitive test, especially in cases where linkage is difficult to detect because the disease alleles are common and have modest effects. Thus it would be desirable to identify disease genes by a method that combines the advantages of the linkage and population-association approaches.

Several such approaches have been described (Rubinstein et al. 1981; Field et al. 1986; Falk and Rubinstein 1987; Thomson 1988; Thomson et al. 1989; Field 1991; Terwilliger and Ott 1992; Spielman et al. 1993; Schaid and Sommer 1994). The methods differ in detail, but all have the following procedure in common: they consider alleles found in the parents of an affected offspring and, in various ways, compare alleles transmitted—versus alleles not transmitted—to the affected offspring.

We show here that the way in which this comparison is carried out determines the kind of inferences that can be drawn about the presence of linkage or of association. Furthermore, the validity of these inferences depends on the underlying population structure and history, in ways that may not be obvious; this dependence is the main subject of the present paper.

Historical Models

It is well known that population admixture and other processes reflecting population history may give rise to disease association even for *unlinked* loci ("spurious" association). Similarly, aspects of population history

Received March 8, 1995; accepted for publication May 18, 1995.

Address for correspondence and reprints: Dr. W. J. Ewens, Department of Biology, University of Pennsylvania, Philadelphia, PA 19104-6018. E-mail: Wewens@mail.sas.upenn.edu

0002-9297/95/5702-0029$02.00

may contribute to association between *linked* loci. In order to establish that loci that exhibit association are in fact linked, it is essential to take into account the effect of population history and structure. In this section we set up and analyze two multigeneration models of population history and subdivision; from these we will deduce properties of several statistics designed to test for linkage and for association between disease and marker loci when population subdivision is present. (We use "subdivision" broadly for the effects of stratification and population heterogeneity, as well as for actual separation into distinct subpopulations.) There are of course infinitely many possible models of population subdivision. We chose simplified models whose properties are clear-cut, which will allow us to make comments applying to more realistic cases.

We assume, for simplicity, a recessive disease whose genetic basis is a single susceptibility allele at a single locus. (The analysis for more general modes of inheritance—and more than one disease allele at the locus—is straightforward and will be indicated later. Models in which multiple loci contribute simultaneously to disease susceptibility are much more complex and are not considered here.) Specifically, we assume a disease locus with alleles D and d, such that only DD individuals can be affected by the disease in question. No generality is lost in the mathematical analysis by assuming that every DD individual must be affected, implying that the penetrance for this genotype is unity, so we make this assumption in the analysis. We assume also a marker locus M, with alleles M and m: the extension to more than two alleles is discussed later also. The recombination fraction between M and D loci is denoted θ.

Model 1, a model of "immediate admixture," is as follows (fig. 1). In generation 0, individuals are assumed to live in a collection of k subpopulations, with random mating within subpopulations but no mating between subpopulations. The relative sizes of the subpopulations are denoted $\alpha_1, \alpha_2, \ldots, \alpha_k$. It is necessary to specify the frequencies, in generation 0, of the four gametes MD, Md, mD, and md in each of these subpopulations, and these are denoted, in subpopulation i,

$$x_{i1}, x_{i2}, x_{i3}, x_{i4} . \quad (1)$$

The frequencies p_i of D and q_i of M in subpopulation i are, respectively, $p_i = x_{i1} + x_{i3}$ and $q_i = x_{i1} + x_{i2}$. The coefficient of gametic disequilibrium δ_i within this subpopulation is defined by

$$\delta_i = x_{i1}x_{i4} - x_{i2}x_{i3} . \quad (2)$$

(Although the following usage is not universal, in this paper we use the term "linkage disequilibrium" for association between loci *only* when the loci are in fact linked. Association between loci not known to be linked is designated "gametic association" (Lewontin 1988) or "population association.")

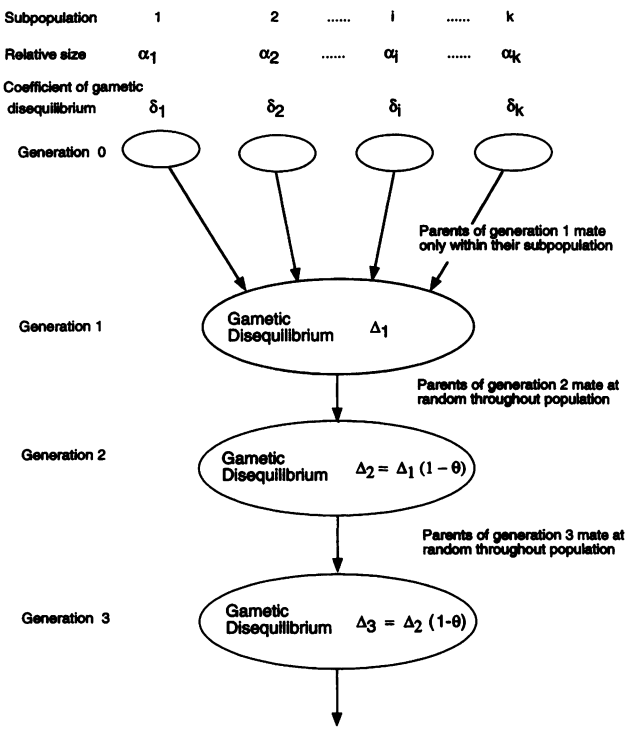

Figure 1 Model 1: population structure/history—the "immediate" admixture model. Initially, mating is only within subpopulations. Each subpopulation contributes a proportion α_i to generation 1. Parents of generation 2 mate at random, without regard to subpopulation of origin, and random mating continues in succeeding generations. For details, see text.

The children of individuals in generation 0, namely generation 1, are assumed to migrate to a common area. The frequencies of D and M in this generation are denoted p and q, respectively, where

$$p = \sum \alpha_i p_i, \quad q = \sum \alpha_i q_i . \quad (3)$$

(Here and throughout, all summations are over subpopulations.) There will also exist in generation 1 a coefficient of gametic disequilibrium Δ_1, defined in a manner analogous to equation (2) and determined in part by the various δ_i values and in part by the admixture process (so that, even if all the δ_i are zero, this coefficient will usually not be zero).

The members of generation 1 mate at random, without regard to the subpopulation origin of their mates. The population consisting of their children, namely generation 2, will have the same gene frequencies (3) as their parents. However the coefficient of gametic disequilibrium in generation 2, denoted Δ_2, is (from standard population genetics theory) $\Delta_2 = \Delta_1(1 - \theta)$ and, as with Δ_1, will contain a component due to the admixture

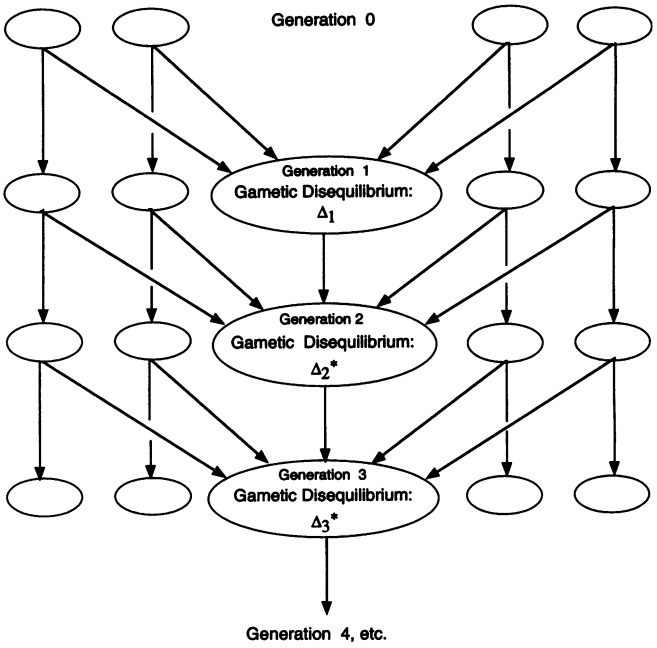

Figure 2 Model 2: population structure/history—the "gradual" admixture model. Initially (generation 0), mating is only within subpopulations. The proportion of generation 1 contributed by each subpopulation is α_i. The members of generation 1 mate randomly and contribute to generation 2, but unmixed descendants of the founder populations also contribute to generation 2. This pattern continues through succeeding generations, so that each generation has some parents who are immigrants from the current generation of the separate subpopulations.

process. We assume that individuals in generation 2 also mate at random in the common area, giving rise to generation 3. In our model, generation 2 is the first generation of parents whose genotype frequencies are in Hardy-Weinberg proportions, since it is the first generation produced by random mating throughout the population sampled. Individuals in generation 3 mate at random to produce generation 4, and so on.

Model 2 is one of "gradual admixture" (fig. 2). As in model 1, individuals in generation 1 move to a common mating area and mate at random there. The coefficient of gametic disequilibrium Δ_1 among these individuals is the same as that for model 1. In contrast to model 1, however, their offspring are joined by offspring of further immigrants who have mated previously within their own subpopulations, so that the individuals in generation 2 consist of offspring from both groups. The coefficient of gametic disequilibrium Δ_2^* is now a function of Δ_1 and the parameters describing the genetic composition and contributions from the initial subpopulations. In the next generation, individuals in generation 2 in the common mating area mate at random, and (as in the previous generation) their offspring are joined by those of further immigrants from the separate subpopulations who have mated previously within those subpopulations. Generation 3 thus consists of the offspring of both groups. This consistent migration is maintained in all future generations, so that a fixed fraction of the children in any generation have as their parents immediate immigrants from the separate subpopulations.

We regard model 2 as being more relevant from a real-world point of view. However, examination of the properties of various statistics discussed below is more straightforward in model 1, so we give full details of the analysis for that case and only quote corresponding results for model 2.

Statistical Notes

It is assumed that the sample of affected children is taken from the common mating area. We will examine later properties of various statistics proposed for testing specified genetic hypotheses. To do this, we note several points of statistical theory. First, to be a valid χ^2 (with 1 df), a statistic must be of the form

$$(X - Y)^2/\mathrm{Var}(X - Y)\,, \qquad (4)$$

where X and Y are random variables having the same mean under the null hypothesis being tested, and $\mathrm{Var}(X - Y)$ is the variance (or, in practice, an unbiased estimate of the variance) of $X - Y$. The validity of any statistic that is claimed to be a χ^2 under a certain hypothesis is checked by assessing whether these requirements are satisfied. (Strictly speaking, we also require X and Y to have normal distributions, but, in view of the close approximation of the normal to the binomial, we do not insist on this requirement. An exact binomial procedure may also be used, as indicated by Spielman et al. [1993].)

Second, consider a set of $2n$ (multinomial) trials, with four possible outcomes, having respective probabilities x, y, y, and z ($x + 2y + z = 1$), on each trial. If a, b, c, and d are the numbers of outcomes in the four categories, then $\mathrm{Var}(b) = 2ny(1 - y)$, $\mathrm{Var}(c) = 2ny(1 - y)$, and $\mathrm{Cov}(b,c) = -2ny^2$. Thus the variance of $b - c$, found from the standard formula $\mathrm{Var}(b - c) = \mathrm{Var}(b) + \mathrm{Var}(c) - 2\mathrm{Covar}(b,c)$, is $4ny$. Since b and c both have mean values $2ny$, an unbiased estimator of this variance is $b + c$. Except for one special case, this is the "best" (minimum variance unbiased) estimator of the variance of $b - c$ and is thus, except for this special case, the estimator that we use.

The one case in which $b + c$ is not the "best" estimate of the variance of $b - c$ occurs if the probabilities x, y, and z are all functions of some single parameter q. In this case a, b, c, and d can be used to find the maximum-likelihood estimate of q, and from this we can find the maximum-likelihood estimate of the variance of $b - c$ as a function of q. Under the optimality theory of maximum-likelihood estimation, this is the "best" estimator

Table 1

Numbers of Marker Alleles M and m among n_1 Affected and n_2 Unaffected Controls in Random Samples of Unrelated Individuals

	M	m	Total
Affected	x_1	$2n_1 - x_1$	$2n_1$
Control	x_2	$2n_2 - x_2$	$2n_2$

of the variance of $b - c$. If x, y, and z are each functions of two parameters q and r, finding the maximum-likelihood estimators of q and r and from this estimating the variance of $b - c$ will result in the same estimate $(b + c)$ as that found above. If x, y, and z are functions of three or more parameters, individual maximum-likelihood estimation of these parameters is in general impossible, and the only way to estimate the variance of $b - c$ is by using $b + c$. The first and third of these cases will occur in the discussion below.

Tests for Association and Linkage

The Population-Association Statistic

Various statistics have been proposed to test for disease association in populations. The prototype for this procedure is one that we call the "relative risk" (RR) statistic, calculated as follows. A sample of n_1 affected and n_2 control individuals is taken, and the total number of M genes in both groups is counted, leading to the data in table 1. The RR (or contingency) statistic for association between disease and marker allele status, when these data are used, can be written in the form $(x_1/2n_1 - x_2/2n_2)^2/\text{Den}$, where Den is a denominator term that is not important to our discussion. Suppose first that the affected individuals are taken from generation 1. Then the mean value of $x_2/2n_2$, (the sample frequency of M among controls), is, for all practical purposes, the population frequency q, given by equations (3), while the mean value of $x_1/2n_1$ is [freq(MMDD) + freq(MmDD)]/[freq(DD)], the frequencies being taken from the individuals in generation 0. Since any individual in this generation has received his genetic makeup from two parents from the same subpopulation, and since the frequencies of the gametes MD and mD among the offspring of individuals in subpopulation i are, respectively, $x_{i1} - \theta\delta_i$ and $x_{i3} + \theta\delta_i$, the mean value of $x_1/2n_1$ is

$$\{\sum \alpha_i[(x_{i1} - \theta\delta_i)^2 + (x_{i1} - \theta\delta_i)(x_{i3} - \theta\delta_i)]\}/\sum \alpha_i p_i^2$$

$$[\sum \alpha_i(x_{i1} - \theta\delta_i)p_i]/\sum \alpha_i p_i^2$$

$$\{\sum \alpha_i[p_i q_i + (1 - \theta)\delta_i]p_i\}/\sum \alpha_i p_i^2 .$$

The mean of $x_1/2n_1 - x_2/2n_2$, the function that is squared in the numerator in the RR statistic, is thus

$$[(\sum \alpha_i p_i^2 q_i) - q(\sum \alpha_i p_i^2)]/(\sum \alpha_i p_i^2) + (1 - \theta)(\sum \alpha_i p_i \delta_i)/(\sum \alpha_i p_i^2) \quad (5)$$

The first term in formula (5), which does not contain δ_i, represents a "spurious" association between marker and disease gene frequencies that is due solely to the admixture process; the second term derives from associations in the original subpopulations, as measured by the various δ_i values. It is because of the inclusion of the spurious association that the population-association statistic is not an appropriate test of the hypothesis of no association in subdivided populations. (This comment also applies when the affected children are taken from generation 2 or from generation 3.) Note also that, even if no such spurious association exists, the second term is not necessarily zero when $\theta = \frac{1}{2}$, so that the statistic does not test directly the "no linkage" hypothesis $\theta = \frac{1}{2}$. We do not, of course, expect the test to do this, since it is purely a test of association.

The Within-Family Contingency Statistic and the TDT

There is a direct analogue, for within-family data, of the population RR contingency statistic discussed above. This may be called the "HRR" (haplotype relative risk) or "contingency" statistic, since it uses the same within-family data as does the "haplotype relative risk" measure proposed by Falk and Rubinstein (1987). The contingency statistic compares the frequency of the allele M among parental alleles transmitted to affected children versus the frequency of the allele M among alleles not transmitted to affected children. (The same statistic has been called the "AFBAC," by Thomson [1988] and is discussed in detail by Thomson [1995].) It is calculated by using the data in table 2, deriving from the $2n$ genes transmitted to n affected children by their parents and from the $2n$ genes not transmitted by these parents.

The standard contingency statistic calculated from the values in this table is

$$4n(w - y)^2/[(w + y)(4n - w - y)] . \quad (6)$$

Table 2

Marker Alleles M and m among the $2n$ Transmitted and $2n$ Nontransmitted Alleles in Parents of n Affected Children

	M	m	Total
Transmitted	w	$2n - w$	$2n$
Nontransmitted	y	$2n - y$	$2n$
Total	$w + y$	$4n - w - y$	$4n$

Table 3

Combinations of Transmitted and Nontransmitted Marker Alleles M and m among $2n$ Parents of n Affected Children

Transmitted Allele	Nontransmitted Allele: M	Nontransmitted Allele: m	Total
M	a	b	$a + b$
m	c	d	$c + d$
Total	$a + c$	$b + d$	$2n$

As in the work of Spielman et al. (1993), it is convenient to rewrite the data in table 2 in the form given in table II of Ott (1989). This is done in our table 3. Here we focus on the $2n$ parents (rather than on the $4n$ parental genes) and describe each parent in terms of both the marker allele transmitted to the affected child and the allele not transmitted. The key relation between the values in table 3 and those in table 2 is that $w = a + b$ and $y = a + c$. This implies that, in terms of the quantities in table 3, the contingency statistic (6) is

$$4n(b - c)^2/[(2a + b + c)(b + c + 2d)] . \quad (7)$$

We discuss in some detail below the circumstances under which this statistic may be used as a valid χ^2 test statistic for association between disease and marker loci.

On the other hand, Spielman et al. (1993), focusing on linkage rather than on association, proposed a statistic to test formally the (null) hypothesis $\theta = \frac{1}{2}$ for marker and disease loci. This is the TDT (transmission/disequilibrium test) statistic

$$(b - c)^2/(b + c) , \quad (8)$$

and, as might be expected of a test statistic for linkage, uses data only from heterozygous (Mm) parents.

The Effects of History and Admixture

We now discuss which hypotheses the contingency statistic (7) and the TDT statistic (8) could test in the context of population subdivision and admixture. We show below that the properties of these two statistics depend on which of our two models is appropriate and, in the case of expression (7), on how many generations of random mating have occurred before the sample is taken. To demonstrate these properties, we consider, in turn, the cases where the affected individuals sampled come from generations 1, 2, 3, and 4, both in model 1 and in model 2.

Scenario 1

In this scenario the affected offspring come from generation 1 (fig. 1). Consider the event that a parent transmits an M allele to an affected child and does not transmit an m allele. The probability of this event is as follows: P(M transmitted, m not transmitted, child affected)/P(child affected). In the numerator we have the event that the parent is Mm at the marker locus, passes on an MD gamete to the child, and does not pass on an mX gamete (where X can be either D or d) and that the other parent transmits a D allele. The desired probability, $P(b)$, is given by

$$\begin{aligned} P(b) &= \{\sum \alpha_i[x_{i1}x_{i3} + x_{i1}x_{i4}(1 - \theta) \\ &\quad + x_{i2}x_{i3}\theta]p_i\}/(\sum \alpha_i p_i^2) \\ &= \{\sum \alpha_i[p_iq_i(1 - q_i) \\ &\quad + \delta_i(1 - \theta - q_i)]p_i\}/(\sum \alpha_i p_i^2) . \end{aligned} \quad (9)$$

The mean value of the quantity b in table 3 is therefore $2nP(b)$. Similarly the mean value of c is $2nP(c)$, where

$$\begin{aligned} P(c) &= \{\sum \alpha_i[p_iq_i(1 - q_i) \\ &\quad + \delta_i(\theta - q_i)]p_i\}/(\sum \alpha_i p_i^2) . \end{aligned} \quad (10)$$

These values allow us to calculate the mean value of $b - c$, the term whose square appears in the numerator of both expression (7) and expression (8). From equations (9) and (10), the mean of $b - c$ is inferred to be

$$2n(1 - 2\theta)(\sum \alpha_i p_i \delta_i)/(\sum \alpha_i p_i^2) . \quad (11)$$

The expression (11) is zero only when $\theta = \frac{1}{2}$ or $\Sigma\alpha_i p_i \delta_i = 0$ (or both). Thus the only hypotheses that are candidates for testing by expression (7) or expression (8) are the *linkage hypothesis* $\theta = \frac{1}{2}$ (which we denote $H(\theta)$) and the *association hypothesis* $\Sigma\alpha_i p_i \delta_i = 0$. The latter hypothesis is of no direct interest and, in practice, would normally be replaced by the more interesting hypothesis $\delta_i = 0$, $(i = 1,2, \ldots, k)$, which we denote $H(\delta)$. Note that, in contrast to the corresponding population expression (5), the expression (11) contains, in this scenario, no spurious association term and, further, is zero when disease and marker loci are unlinked. Thus the "within-family" data in table 3 potentially enable us to test, in this scenario, simultaneously for association arising in the original subpopulations (uncontaminated by any association due to admixture) and for linkage between disease and marker loci.

To arrive at a test of the linkage hypothesis $H(\theta)$ in the form of equation (4), we must calculate the variance of $b - c$ when $\theta = \frac{1}{2}$. To do this, we note that, when $\theta = \frac{1}{2}$, the marker alleles transmitted from two heterozygous parents to the same child are independent (Spielman et al. 1993), as are the marker alleles transmitted by a heterozygous parent to two affected offspring.

Thus, when $\theta = \frac{1}{2}$, the data in table 3 come from a multinomial distribution with four cells, in which the probabilities of the cells corresponding to b and c are equal when H(θ) is true. The variance of $b - c$ is thus found from the statistical notes above, together with the expressions (9) and (10), for P(b) and P(c), respectively, to be

$$\mathrm{Var}(b - c) = 4n \sum \alpha_i[p_i q_i(1 - q_i) + \delta_i(\tfrac{1}{2} - q_i)]p_i / \sum \alpha_i p_i^2$$

This is a function of many unknown parameters and thus, from the statistical notes, can only be estimated by $b + c$. This leads to the TDT statistic (8) as a valid test statistic for the hypothesis H(θ) of no linkage between disease and marker loci.

We now consider the hypothesis H(δ). Under this hypothesis, the probabilities of the cells corresponding to b and c are again equal, so that the appropriate numerator in any proposed test statistic for this hypothesis is $(b - c)^2$. However, two distinct problems arise with finding an unbiased estimator of the denominator of the statistic, which we now discuss in detail.

First, suppose, for simplicity, that all families in the sample are "simplex" (i.e., there is only one affected offspring in each family). Then the multinomial distribution discussed above is appropriate, and expressions (9) and (10) show that, when H(δ) is true, the variance of $b - c$ is

$$\mathrm{Var}(b - c) = 4n \sum \alpha_i p_i q_i (1 - q_i) / \sum \alpha_i p_i^2 . \quad (12)$$

As with the test of H(θ), this is again a function of many parameters and again can be estimated only by $b + c$. Thus, provided that the sample contains only simplex families, the TDT is the valid test statistic for H(δ) also.

Note that the denominator in the contingency statistic (7) is not appropriate, since (as we note later) this denominator is calculated under the assumption of Hardy-Weinberg frequencies, which, as shown by the Wahlund principle, do not hold in our subdivided population. This principle shows that the Hardy-Weinberg variance $4nq(1 - q)$ exceeds the true value (12), so that the denominator in the contingency statistic (7) overestimates the true variance. As a result, use of this statistic underestimates the correct χ^2 for testing H(δ); that is, it leads in scenario 1 to an inexact conservative test (as we confirm later in the Numerical Examples section).

Second, apart from this problem, when H(δ) is true the marker alleles transmitted by a parent to two affected children are not independent, thus invalidating the multinomial assumption implicit in the denominator of the contingency statistic (7). In this case, estimation of the variance term appears to be very difficult, and there is possibly no simple test of the hypothesis H(δ). Note that this problem does not arise in the testing of H(θ), since the transmitted marker alleles are independent under this hypothesis. In scenario 1, there is no difference between models 1 and 2, so the remarks above apply also for model 2.

Scenario 2

Here the affected individuals come from generation 2. The properties of the sampling process now depend on the model analyzed, and, for simplicity, we carry out a detailed analysis only for model 1. In this model, each parent (in generation 1) of an affected child derived all his or her genetic material from one or other of the original subpopulations. This implies that, in computing the scenario 2 analogues of expressions (9) and (10), we should replace, for those (generation 1) parents deriving their genetic material from subpopulation i, the four gametic frequencies $x_{i1}, \ldots, x_{i4}$ by values updated by one generation, namely $x_{ij} - \theta\delta_i (j = 1,4)$ and $x_{ij} + \theta\delta_i (j = 2,3)$. Calculations similar to those leading to expressions (9) and (10) then give, for this scenario,

$$\mathrm{P}(b) = \{\sum \alpha_i[p_i q_i(1 - q_i) + (1 - \theta)\delta_i(1 - \theta - q_i)]\}/p \quad (13)$$

$$\mathrm{P}(c) = \{\sum \alpha_i[p_i q_i(1 - q_i) + (1 - \theta)\delta_i(\theta - q_i)]\}/p \quad (14)$$

so that the mean value of $b - c$ is now

$$2n(1 - 2\theta)(1 - \theta) \sum \alpha_i \delta_i / p . \quad (15)$$

Clearly, as in scenario 1, this is zero if and only if H(θ) or H(δ) is true (or both). (As in scenario 1, we use H(δ) for the interesting case $\delta_i = 0$, $(i = 1, \ldots, k)$.)

We now discuss how these hypotheses may be tested in this scenario. Consider first the hypothesis H(θ). As in scenario 1, the variance of $b - c$ is a complicated function of many parameters, and the only unbiased estimate of the variance of $b - c$ is $b + c$. This leads once more to the TDT statistic (8) as the valid χ^2 test of the hypothesis H(θ). If there is one affected child per family, then, under the hypothesis H(δ), the variance of $b - c$ is also a complicated function of many parameters. By the same argument as that used in scenario 1, the correct test statistic for H(δ) is, once more, the TDT; the contingency statistic (7) is again not a valid χ^2. If the data contain multiplex families, then, as in scenario 1, it appears very difficult to find any valid test of H(δ).

The details of the analysis are slightly different in model 2 (gradual admixture), with slight changes to equations (13)–(15). However, the broad conclusions reached for model 1—in particular, that the TDT statis-

tic is a valid test of H(θ) and that the contingency statistic (7) is not a valid χ^2—continue to hold for model 2.

Scenario 3

Here the affected individuals are from generation 3. In model 1 the individuals in generation 2 were produced by random mating, so that the population as a whole exhibits Hardy-Weinberg genotype frequencies q^2, $2q(1-q)$, and $(1-q)^2$ at the marker locus, where q is defined in equations (3). However, the genotype frequencies of parents of affected children are not generally in Hardy-Weinberg form. The frequencies P(a), P(b), P(c), and P(d) are

$$P(a) = q^2 + q\Delta_2/p\,, \tag{16}$$

$$P(b) = q(1-q) + (1-\theta-q)\Delta_2/p\,, \tag{17}$$

$$P(c) = q(1-q) + (\theta-q)\Delta_2/p\,, \tag{18}$$

$$P(d) = (1-q)^2 - (1-q)\Delta_2/p\,, \tag{19}$$

as given by Ott (1989). Here Δ_2 is the coefficient of gametic disequilibrium in the parental generation and contains a component from the admixture process. Equations (16)–(19) show that genotype frequencies of parents of affected children are in Hardy-Weinberg form if $\Delta_2 = 0$ but not if $\theta = \frac{1}{2}$. From probabilities (17) and (18), the mean value of $b - c$ is inferred to be

$$2n(1-2\theta)\Delta_2/p\,, \tag{20}$$

and this is zero when $\theta = \frac{1}{2}$ (i.e., H(θ) is true) or when $\Delta_2 = 0$. Thus the only two hypotheses that can potentially be tested by expression (7) or expression (8) are $\theta = \frac{1}{2}$ and $\Delta_2 = 0$. The hypothesis H(θ) is, as in previous scenarios, still of major interest. In contrast, the hypothesis $\Delta_2 = 0$ might not be of much interest, since Δ_2 contains (perhaps substantial) spurious components from the admixture process. We consider the test of this hypothesis below.

Under H(θ), P(b) and P(c) are functions of three unknown parameters (Δ_2, p, and q), and the statistical notes above show again that the only unbiased estimator of the variance of $b - c$ is $b + c$. Thus, as in scenarios 1 and 2, the TDT statistic (8) is the valid χ^2 test of the hypothesis H(θ).

Clearly, the power of the TDT test depends on Δ_2, one component of which derives from admixture. Thus, not only is the TDT statistic valid under population subdivision and admixture, it is actually made more powerful by the admixture process. We illustrate this below, with a numerical example.

We next consider the test of the hypothesis $\Delta_2 = 0$. Under this hypothesis, the probabilities (16)–(19) depend only on q; as a result, the estimator of the variance of $b - c$ will not be the same as that used in the testing of H(θ). When there is only one affected child in each family in the sample, the various observations are independent and the multinomial distribution applies. The variance of $b - c$ is therefore $4nq(1-q)$, the maximum-likelihood estimator of q is $(2a + b + c)/4n$, and thus the estimate of the variance of $b - c$ is $(2a + b + c)(b + c + 2d)/4n$, as in expression (7). It follows that, in scenario 3 with simplex families, the contingency statistic (7) *does* provide a valid χ^2 test for the hypothesis $\Delta_2 = 0$—and that it actually has slightly greater power than the TDT, since the former makes use of all the data.

When some families in the sample have more than one affected child, the multinomial assumption is no longer valid and there does not appear to be any obvious test, which uses all the data, of the hypothesis $\Delta_2 = 0$. It is therefore possible, in this scenario, to form valid tests for both linkage ($\theta = \frac{1}{2}$) and, when there is only one affected child in each family, association ($\Delta_2 = 0$). As we remarked above, however, the hypothesis $\Delta_2 = 0$ might not be of much interest. Accordingly, one might be tempted in this case to use the contingency statistic (7), not as a test of association but as a test of linkage, treating this statistic as though it were a valid χ^2 under H(θ). The rationale would be that, apart from statistical fluctuations, the contingency statistic (7) can attain significance only if marker and disease loci are linked. However, this procedure is not valid, since this statistic does not have a χ^2 distribution when H(θ) is true; when H(θ) is true, the parental probabilities (16)–(19) differ from those when $\Delta_2 = 0$. Consequently, sampling is from two different background distributions in the two cases, implying that the contingency statistic (7) is not valid as a test for linkage. We confirm this conclusion in example 2 described in the Numerical Examples section below.

The corresponding conclusions in model 2 are more straightforward. Hardy-Weinberg frequencies have not been attained among the members of generation 2—and, in particular, among parents of affected children. Transmission frequencies more complex than those in probabilities (16)–(19) apply, and the most important implication of this is that the contingency statistic (7) is not a valid χ^2 as a test of association. On the other hand, TDT statistic (8) is a valid χ^2 as a test for linkage.

Scenario 4

Here the affected children are taken from generation 4. The admixture process, in both models, reached a "steady state" in scenario 3, so that properties of scenario 4 are very similar to those of scenario 3. All the conclusions reached above for scenario 3 continue to hold, for both models, in scenario 4. The main quantitative difference will arise from a decreased coefficient of association due to recombination, and the effect of this will be to decrease both statistics (7) and (8) to approximately $(1-\theta)^2$ of their scenario 3 values.

Table 4

Values of Contingency Statistic (7) and TDT χ^2 (8), Corresponding to Expected Values of the Observations (see Text)

	Model 1		Model 2	
Generation	Contingency Statistic (7)	TDT χ^2 (8)	Contingency Statistic (7)	TDT χ^2 (8)
1	1.30	1.48	1.30	1.48
2	1.63	2.07	1.48	1.83
3	16.30	15.34	8.26	8.53
4	13.06	12.43	6.69	6.99

Numerical Examples

We confirm the main points made in the above discussion, by two numerical examples.

Example 1: Properties of the Contingency Statistic and the TDT When Both Association and Linkage Are Present: Power and Validity

Suppose that in generation 0 there are two subpopulations, of equal sizes. In population 1 the gametic frequencies (1) are .68, .12, .12, and .08, while in population 2 they are .08, .12, .12, and .68. (The values chosen are not meant to be realistic but simply to illustrate the model.) Thus $p_1 = q_1 = .8$, $\delta_1 = .04$, $p_2 = q_2 = .2$, and $\delta_2 = .04$. The recombination fraction θ is assumed to be .1. The admixture processes for models 1 and 2 are as described in the above sections; in model 1, admixture is immediate, whereas, in model 2, we assume that 20% of the individuals in any generation are new immigrants from the original subpopulations.

Suppose first that 100 affected individuals, all from different families, are taken from generation 1 (i.e., scenario 1). When expressions (9) and (10) and similar expressions for P(a) and P(d) are used, the mean values of the quantities a, b, c, and d in table 3 are found to be 128.94, 34.59, 25.18, and 11.30. If the observed data take these mean values, the contingency statistic (7) and the TDT statistic (8) take the values shown in line 1 of table 4. By conventional criteria for significance ($\chi^2 = 3.84$, df = 1) the TDT does not detect the linkage (or the association) between the two loci, essentially because the genetic material in each affected individual comes from within one or the other subpopulation, where the coefficient of association is small. The relative values of the statistics (7) and (8) also illustrate the conservative nature of the (incorrect) statistic (7) as a test for H(δ), as predicted above for scenario 1.

Suppose next that the 100 affected individuals are taken from generation 2 (scenario 2). For model 1, the mean values of a, b, c, and d are found from equations (13) and (14) and similar expressions for P(a) and P(d). If the observed data take these mean values, the association statistic (7) and the TDT statistic (8) are as shown in table 4 (line 2). Similar values are found for model 2. The TDT does not detect the linkage (or the association) between the two loci under either model, again because the associations that it uses are still the small values in the original subpopulations. As in scenario 1, the contingency statistic (7) is too conservative as a test of H(δ).

The coefficient of gametic association Δ_2 within generation 2 is .1134, far larger than the value $\delta = .04$ in the original subpopulations. This increase is derived largely from the admixture process, and the large value will become relevant in the testing procedure when affected individuals come from generation 3.

Suppose next that the 100 affected individuals come from generation 3 (scenario 3). If, once again, the observed data take the corresponding mean values, the contingency statistic (7) and the TDT statistic (8) are as shown, for both models, in table 4 (line 3). The TDT statistic is now significant, in both models, detecting the linkage (and the association) between the two loci. Its power to do this derives largely *from the admixture process* and the gametic association that this process generates. The value in model 2 is rather less than that in model 1, since some individuals in this model are recent immigrants and the increased association due to admixture is not relevant to them.

In generation 3, in contrast to the preceding cases, the contingency statistic (7) is a valid χ^2 for testing the hypothesis $\Delta_2 = 0$ in model 1. It is significant and thus has detected the association between the loci. The value of the contingency statistic slightly exceeds that of the TDT statistic, illustrating its slightly greater power as a test of the hypothesis $\Delta_2 = 0$. As discussed above, however, the contingency statistic should not be used as an indirect test statistic for linkage, since contingency statistic (7) does not have the nominated type I error as a test of the hypothesis $\theta = \frac{1}{2}$. (See example 2 and comments after eq. [20] above.) In the more realistic model 2, the contingency statistic is not a valid χ^2 test for any hypothesis, and, as in scenario 1 and 2, its value is *less than* that of the valid χ^2 statistic (8).

The properties of scenario 4 (table 4, line 4) are similar to those of scenario 3, as predicted above. The main numerical effect is that the coefficient of gametic association has decreased by 10%, from .1134 to .1021, and, as a result, the values of both contingency statistic (7) and χ^2 statistic (8), in both models, have decreased to about $(1 - 0.1)^2$, or 81% of their values in scenario 3.

Example 2: Properties of the Contingency Statistic and TDT When Association Is Present but Linkage Is Not: Type I Error

In the second numerical example, we examined the properties of the contingency and TDT statistics when

Table 5

Frequency (%) with Which Contingency (7) and TDT (8) Test Statistics Exceeded 5% Value (3.84) of χ^2 with df = 1, in Simulations with $\theta = \frac{1}{2}$ (Scenario 3, model 1)

Gametic Disequilibrium Δ_2	Contingency Statistic (7) ± Standard Error	TDT χ^2 (8) ± Standard Error
.010	5.66 ± .06	5.01 ± .07
.015	6.17 ± .06	5.08 ± .07
.020	7.09 ± .07	5.04 ± .07
.025	8.34 ± .11	5.02 ± .07

the null hypothesis ($\theta = \frac{1}{2}$) is in fact true. Here we used simulation to check the claim, made above, that in scenario 3, model 1 (i.e., immediate admixture), the contingency statistic (7) does not have a χ^2 distribution under the no-linkage hypothesis $H(\theta)$. (At the same time we checked the validity of the TDT statistic [8] for this hypothesis.) We used a disease allele frequency (p) of .05, a marker allele frequency (q) of .4, and the four alternative values Δ_2 = .010, .015, .020, and .025.

With each value of Δ_2, and using equations (16)–(19) with $\theta = \frac{1}{2}$, we constructed by Monte Carlo simulation a data set (a,b,c,d) as in table 3, with $a + b + c + d = 1{,}000$. We then calculated the values of the statistics (7) and (8) for these data and noted whether the calculated value exceeded the significance value of χ^2, namely 3.84. This procedure was repeated 100,000 times, giving a very accurate estimate of the true type I error of each test statistic. The results are given in table 5.

We note that, whereas the type I error of the TDT statistic never differs appreciably from the value 5%, the type I error for the contingency statistic (7) increases steadily from 5.66% (when $\Delta_2 = 0.010$) to 8.34% (when $\Delta_2 = 0.025$), confirming our claim that the contingency statistic should not be used as a test for linkage. Specifically, use of contingency statistic (7) will lead to more than the nominal frequency (e.g., 5%) of false-positive findings. In practice, this means that use of contingency statistic (7) will exaggerate the apparent significance of results.

Discussion

We have analyzed two multigeneration models of population subdivision and admixture, to show the consequences for within-family tests of association and linkage disequilibrium between genetic disease and marker locus. Our analysis makes some simplifying assumptions, and we discuss several of those here.

Mode of Inheritance and Multiple Marker Alleles

It is straightforward to generalize the above analysis to the case of an arbitrary mode of inheritance, and it is found that the TDT is a valid test statistic whatever the mode of inheritance. (Details are available from W.J.E.) This does not imply that a more powerful testing procedure cannot be found when the mode of inheritance is known. Indeed, Schaid and Sommer (1994) provide statistics analogous to the TDT that provide more powerful tests than the TDT when the mode of inheritance is known. However, even in these cases, the decrease in power is only modest if the TDT is used. Furthermore, if the true mode of inheritance is additive, the TDT is the most powerful test.

Several authors have discussed generalizations of the TDT to the case of multiple marker alleles (see, in particular, the work of Bickeböller and Clerget-Darpoux [in press] and Rice et al. [in press]). The conclusions reached above on the effects of admixture on the TDT will continue to apply for any such generalization.

Population Structure

The model of "immediate" admixture is obviously unrealistic for human and most other populations, but it highlights the aspects of admixture that are important for understanding gametic association. Even if complete panmixia with no further admixture (i.e., model 1) begins in generation 1, the effect of admixture on marker genotype frequencies persists in generation 2, among the parents of affected individuals, and does not disappear until generation 3; as a result, the contingency test of association is not a valid χ^2 until generation 3. As long as further admixture continues (model 2), the contingency test is not valid.

In practice, the population structure and the migration processes affecting admixed populations will, in humans, be far more complex than those described by the model above. It is difficult to predict the magnitude of the error caused by using a test that, as a result of admixture, is not strictly valid. In some cases where admixture is (or was) present, the error resulting from use of the contingency statistic (7), when it is not a valid χ^2, might be small. In this regard, our tables 4 and 5 cover only a small number of possibilities.

Independence

Even if some families in the sample have more than one affected child, the TDT remains a valid test of the linkage hypothesis. However, the presence of multiplex sibships makes the contingency statistic invalid as a test of association, even if there are no problems resulting from recent admixture.

Conclusions

We summarize our conclusions and recommendations regarding the tests discussed above, as follows:

1. Even when the contingency statistic is valid as a test of association, it is not valid as a test of linkage. Thus, in model 1, scenario 3, table 5 shows that, when this statistic is used as a test of linkage, the actual type I error will

exceed the nominal value, leading to an excess of false-positive results. This excess increases appreciably as the coefficient of gametic disequilibrium increases.

2. The contingency statistic is also not valid, in general, as a test for association, since it requires random mating in the population and no admixture for at least two generations before the sample of affected offspring is taken. In practice, of course, we usually do not know how large or how recent the admixture effects may be. In view of the substantial admixture occurring in modern populations, among previously separated ethnic groups, the effects may be large.

3. There is a further difficulty with the contingency statistic; it does not provide a valid test for association when the sample contains multiplex families, unless only one affected child from each family is used in the analysis.

4. For the TDT, in contrast, the preceding considerations do not apply. Whether association is spurious or due to linkage (with disequilibrium), the TDT is a valid test for the null hypothesis $\theta = \frac{1}{2}$ (no linkage). Indeed the power of the TDT requires—and is enhanced by—the presence of association, whatever its cause. Furthermore, the TDT remains valid as a test for linkage even when some families in the sample contain two or more affected children.

5. Accordingly, when a test for linkage in the presence of association is desired, we recommend the TDT as the test of choice.

Acknowledgments

We thank Ralph McGinnis for valuable comments. This research was supported by NIH grants GM21135 (to W.J.E.) and DK46618 and DK47481 (to R.S.) and by Australian Research Council grant 20.164.075 (to W.J.E.).

References

Bain SC, Prins JB, Hearne CM, Rodrigue NR, Rowe BR Pritchard LE, Ritchie RJ, et al (1992) Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to HLA-DR4-positive individuals. Nat Genet 2:212–215

Bell GI, Horita S, Karam JH (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. Diabetes 33:176–183

Bickeböller H, Clerget-Darpoux F. Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Proceedings of Genetic Analysis Workshop 9, Montreal, 1994. Genet Epidemiol (in press)

Cox NJ, Baker L, Spielman RS (1988) Insulin-gene sharing in sib pairs with insulin-dependent diabetes mellitus: no evidence for linkage. Am J Hum Genet 42:167–172

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

Falk CT, 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

——— (1991) Non-HLA region genes in insulin dependent diabetes mellitus. Baillieres Clin Endocrinol Metab 5:413–438

Field LL, Fothergill-Payne C, Bertrams J, Baur MP (1986) HLA-DR effects in a large German IDDM dataset. Genet Epidemiol Suppl 1:323–328

Julier C, Hyer RN, Davies J, Merlin F, Soularue P, Briant L, Cathelineau G, et al (1991) Insulin-IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4-dependent diabetes susceptibility. Nature 354:155–159

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6:127–130

Rice JP, Neuman RJ, Hoshaw SL, Daw EW, Gu G. TDT tests with covariates and genome screens with MOD scores: their behavior on simulated data. Proceedings of Genetic Analysis Workshop 9, Montreal, 1994. Genet Epidemiol (in press)

Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk C, Ginsberg F (1981) Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups. Hum Immunol 3:384

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

Spielman RS, Baur, MP, Clerget-Darpoux F (1989) Genetic analysis of IDDM: summary of GAW5-IDDM results. Genet Epidemiol 6:43–58

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

Terwilliger JD, Ott J (1992) A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337–346

Thomson G (1988) HLA disease associations: models for insulin dependent diabetes mellitus and the study of complex human genetic disorders. Annu Rev Genet 22:31–50

——— (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498 (in this issue)

Thomson G, Robinson WP, Kuhner MK, Joe S (1989) HLA, insulin gene, and Gm associations with IDDM. Genet Epidemiol 6:155–160