European Journal of Human Genetics (2004), 12, 105–114

www.nature.com/ejhg

ARTICLE

Transmission disequilibrium test (TDT) for case–control studies

Edward A Ruiz-Narváez[1] and Hannia Campos*[,1,2]

[1]*Department of Nutrition, Harvard School of Public Health, Boston, MA, USA;* [2]*Centro Centroamericano de Población, Universidad de Costa Rica, San Pedro, Costa Rica*

Genetic association, case–control studies are becoming a major instrument in the attempt to identify disease susceptibility markers of complex diseases. However, a major drawback of population-based studies of genetic association is the confounding effect of the population subdivision. We developed a statistic named *T*-value that estimates the differential transmission of marker alleles from heterozygous parents to the affected offspring, based on population data. Our method does not assume Hardy–Weinberg equilibrium and it can be used in very different population structures. A great advantage of this approach is that the genetic structure of the population can be assessed with a few unlinked loci and using classical population genetics theory (ie Wright's F-statistics). Four general models, assuming either one population with random mating, or one population without random mating, or several populations with random mating within them, or several populations without random mating within them, were developed to determine the behavior of the *T*-value under different mating conditions. Although a complete knowledge of the population structure is ideal to choose the best model, the simulations show that for a total inbreeding of 0.30 or less the last three models gave very similar estimates of the *T*-value. The model that assumed that total departure of Hardy–Weinberg proportions is due to population subdivision was the most robust under different scenarios of population structure. In sum, this study describes a novel procedure that can be used to identify the transmission of disease susceptibility markers in population-based studies.
European Journal of Human Genetics (2004) **12,** 105–114. doi:10.1038/sj.ejhg.5201099

Keywords: case–control studies; transmission disequilibrium test; genetic epidemiology

Introduction

Understanding the genetic basis of human diseases is a major goal of the modern genetic research. For genetically simple diseases like those with a pattern of Mendelian inheritance, that is, high penetrance and early onset, linkage analysis is a simple approach to detect the cosegregation of a marker locus and the disease through a pedigree. Unfortunately, the monogenic diseases are only a small fraction of all the human diseases in the world today, where the most common diseases, that is, cardiovascular disease, cancer, and neuropsychiatric disorders, have a polygenic basis and show complex interactions with environmental factors.[1,2] Other complications, such as unclear Mendelian inheritance, low penetrance, and late onset, restrict the capacity of the traditional linkage analysis to uncover the genetic basis of many human diseases. Thus, there is a need to develop novel analytic tools that can immerse traditional family-based genetic analysis of human disease into population-based human studies of complex disease.[3,4]

*Correspondence: Dr H Campos, Department of Nutrition, Room 353A, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA. Tel: +1 617 432 0100; Fax: +1 617 432 2435;
E-mail: hcampos@hsph.harvard.edu
Received 24 March 2003; revised 28 August 2003; accepted 11 September 2003

Owing to interindividual genetic variation, population-based studies are a promising approach to increase our understanding of complex genetic diseases. For example, genome-wide association studies using population samples, may be used to map loci affecting complex traits.[5] However, false-positive associations may be obtained if the population under study is stratified. The transmission disequilibrium test (TDT) avoids the problem of ethnic from association by testing the difference between the frequency of marker alleles transmitted from heterozygous parents to the affected offspring and the frequency of marker alleles not transmitted.[6] Although the original use of the TDT was to test for linkage in the presence of population association, it can be used to test any marker even if there is no prior evidence for association.[7]

However, the TDT is limited by the availability of DNA samples from parents of the affected individuals. This becomes a serious drawback particularly in epidemiological studies of late-onset diseases. A population-based approach does not need case-relative pairs, and careful matching for ethnic background may circumvent the confounding by ethnicity. But, because the description of genetic variation in human populations is an important prerequisite for the development of mapping strategies, an obvious question is whether an analogous test of the TDT can be carried out on population-based studies.

Mitchell[8] developed a statistic (T), which measures disease–marker associations, and it can be estimated from case–control data. However, a test of significance for the T-statistic was not developed. Furthermore, this statistic can be estimated only in the unlikely situation of a population that has no serious deviations from Hardy–Weinberg proportions, a premise that cannot be achieved if the population is stratified. Thus, a more general method is needed to estimate the differential transmission of marker alleles to the affected offspring, based on population data. We developed a statistic (T-value) to estimate the proportion of transmission of a potential high-risk marker allele from heterozygous parents to the affected offspring. To estimate the T-value, it is not necessary to assume that Hardy–Weinberg equilibrium holds; in fact, it can be estimated in very diverse scenarios of population structure.

Methods

Let us suppose a genetic marker with two codominant alleles, A_1 and A_2. The TDT uses heterozygous parents for the genetic marker and it compares the frequency by which a 'high-risk' marker allele (A_1) is transmitted to the affected offspring, to the frequency by which the alternate marker allele (A_2) is transmitted to the same offspring.[6] To carry out the TDT in a case–control study, we have to estimate the number of alleles, both A_1 and A_2, transmitted from heterozygous parents to their affected offspring.

The T-value can be written as:

$$T = \frac{x_1}{x_1 + x_2} \tag{1}$$

where x_1 and x_2 are, respectively, the number of alleles A_1 and A_2 transmitted from heterozygous parents to the affected offspring. By definition, x_1 and x_2 can be calculated as

$$x_1 = x_{1,11} + x_{1,12}, \tag{2}$$

$$x_2 = x_{2,22} + x_{2,12} \tag{3}$$

where $x_{i,ij}$ $(i, j = 1, 2)$ is the number of alleles A_i inherited from heterozygous parents to the affected offspring with genotype A_iA_j.

In population-based studies, we do not have information about the genotypes of the parents of the cases; therefore, we can assume that the different $x_{i,ij}$ values, as well as the T-value, are random variables. If we define $\varphi_{i,ij}$ as the probability that one allele A_i taken from a patient with genotype A_iA_j had been inherited from a heterozygous parent, the probability distributions of the different $x_{i,ij}$ are

$$P(x_{1,11} = r) = \binom{2P}{r} \varphi_{1,11}^{r} (1 - \varphi_{1,11})^{2P-r}, \quad r \in [0, 1, \ldots, 2P] \tag{4}$$

$$P(x_{1,12}{=}s) = \binom{H}{s} \varphi_{1,12}^{s} (1 - \varphi_{1,12})^{H-s}, \quad s \in [0, 1, \ldots, H] \tag{5}$$

$$P(x_{2,12} = t) = \binom{H}{t} \varphi_{2,12}^{t} (1 - \varphi_{2,12})^{H-t}, \quad t \in [0, 1, \ldots, H] \tag{6}$$

$$P(x_{2,22} = u) = \binom{2Q}{u} \varphi_{2,22}^{u} (1 - \varphi_{2,22})^{2Q-u}, \quad u \in [0, 1, \ldots, 2Q] \tag{7}$$

where P, H, and Q are the number of sampled patients with genotypes A_1A_1, A_1A_2, and A_2A_2, respectively.

The above probability distributions depend on the system of matings and the genetic structure of the sampled population. Although more complex situations can be envisioned, we will focus our analysis on four general models: (1) one population with random mating, (2) one population without random mating, (3) several populations with random mating within them, and (4) several populations without random mating within them.

One population with random mating (model T1)

The simplest situation occurs when the sampling is made on a single panmictic population; therefore, the $\varphi_{i,ij}$ probabilities can be calculated using the Hardy–Weinberg genotype frequencies. Table 1 shows the frequencies of the

Table 1 Frequency of the different matings occurring within a panmictic population (model T1) and within an inbred population (model T2), as well as the number of alleles A_1 and A_2 transmitted from heterozygous parents to the different kinds of offspring[a]

	Frequency		*Offspring proportion*			*Number of alleles A_1 from heterozygous parents*		*Number of alleles A_2 from heterozygous parents*	
Mating (1)	*Model T1 (2a)*	*Model T2 (2b)*	*A_1A_1 (3)*	*A_1A_2 (4)*	*A_2A_2 (5)*	*A_1A_1 (6)*	*A_1A_2 (7)*	*A_2A_2 (8)*	*A_1A_2 (9)*
$A_1A_1 \times A_1A_1$	p_1^4	$p_1^4+p_1p_2F$	1	0	0	0	0	0	0
$A_1A_1 \times A_1A_2$	$4p_1^3p_2$	$4p_1^3p_2(1-F)$	$\frac{1}{2}$	$\frac{1}{2}$	0	1	0	0	1
$A_1A_1 \times A_2A_2$	$2p_1^2p_2^2$	$2p_1^2p_2^2(1-F)$	0	1	0	0	0	0	0
$A_1A_2 \times A_1A_2$	$4p_1^2p_2^2$	$4p_1^2p_2^2(1-F)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	2	1	2	1
$A_1A_2 \times A_2A_2$	$4p_1p_2^3$	$4p_1p_2^3(1-F)$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	1	1	0
$A_2A_2 \times A_2A_2$	p_2^4	$p_2^4+p_1p_2F$	0	0	1	0	0	0	0

[a]The different probabilities $\varphi_{i,ij}$ (see text) can be calculated according to the equations: $\varphi_{1,11} = \frac{1}{2}\sum(\text{column}(2)\times \text{column}(3)\times\text{column}(6))/\sum(\text{column}(2)\times\text{column}(3))$, $\varphi_{1,12} = \sum(\text{column (2)} \times \text{column (4)} \times \text{column (7)})/\sum(\text{column (2)} \times \text{column (4)})$, $\varphi_{2,12} = \sum(\text{column (2)} \times \text{column (4)} \times \text{column (9)})/\sum(\text{column (2)} \times \text{column (4)})$, $\varphi_{1,11} = \frac{1}{2}\sum(\text{column}(2)\times\text{column}(5) \times\text{column}(8))/\sum(\text{column}(2)\times \text{column}(5))$.

different matings occurring in the population, the offspring proportions of each mating, as well as the number of A_1 or A_2 alleles inherited by each offspring from A_1A_2 heterozygous parents. According to Table 1, the different $\varphi_{i,ij}$ probabilities are $\varphi_{1,11(1)} = \varphi_{1,12(1)} = p_2$ and $\varphi_{2,12(1)} = \varphi_{2,22(1)} = p_1$, where p_1 and p_2 are the frequencies of the A_1 and A_2 alleles in the sampled population, and the number 1 in parentheses refers to the first considered model. These results show that, within a random mating population, the probability of an A_1 allele from a patient with genotype A_1A_1 or A_1A_2 being inherited from a heterozygous parent equals the frequency of the allele A_2 in the general population. Likewise, the probability of an A_2 allele from a patient with either genotype A_2A_2 or A_1A_2 being inherited from a heterozygous parent is equal to the frequency of the allele A_1 in the general population.

The expected value (μ) and variance (σ^2) of the random variables x_1 and x_2 under this model can be calculated as

$$\mu_{x1(1)} = 2P\varphi_{1,11(1)} + H\varphi_{1,12(1)}, \tag{8}$$

$$\sigma^2_{x1(1)} = 2P\varphi_{1,11(1)}(1-\varphi_{1,11(1)}) + H\varphi_{1,12(1)}(1-\varphi_{1,12(1)}) \tag{9}$$

$$\mu_{x2(1)} = 2Q\varphi_{2,22(1)} + H\varphi_{2,12(1)}, \tag{10}$$

$$\sigma^2_{x2(1)} = 2Q\varphi_{2,22(1)}(1-\varphi_{2,22(1)}) + H\varphi_{2,12(1)}(1-\varphi_{2,12(1)}) \tag{11}$$

The probability distribution of T can be determined by the Monte–Carlo method, according to equation (1) and the different probability distributions of the $x_{i,ij}$. An estimator of the proportion of A_1 alleles transmitted from heterozygous parents to the affected offspring can be calculated by taking the expectation $\mu_{T(1)}$ of the distribution of T; therefore

$$\mu_{T(1)} = \frac{\mu_{x1(1)}}{\mu_{x1(1)} + \mu_{x2(1)}} \tag{12}$$

The null hypothesis ($\mu_{x1(1)} - \mu_{x2(1)} = 0$ or $\mu_{T(1)} = \frac{1}{2}$) of nondifferential transmission of the allele A_1 from heterozygous parents to the affected offspring can be tested by the statistic

$$\Delta_{(1)} = \frac{\mu_{x1(1)} - \mu_{x2(2)}}{\sqrt{\sigma^2_{x1(1)} + \sigma^2_{x2(1)}}}, \tag{13}$$

that follows, approximately, a standard normal distribution if the number of sampled gene copies is large enough. A permutation test can be performed by simulations of the probability distribution of T according to equation (1).

One population without random mating (model T2)

The departure from the Hardy–Weinberg equilibrium causes a correlation between uniting gametes within the population, which can be measured by the inbreeding index F.[9] Table 1 shows the mating frequencies with an inbreeding index F. According to the values in Table 1, the different $\varphi_{i,ij}$ probabilities are

$$\varphi_{1,11(2)} = p_2\left(\frac{p_1(1-F)}{p_1(1-F)+F}\right) \tag{14}$$

$$\varphi_{1,12(2)} = p_2, \tag{15}$$

$$\varphi_{2,12(2)} = p_1, \tag{16}$$

$$\varphi_{2,22(2)} = p_1\left(\frac{p_2(1-F)}{p_2(1-F)+F}\right) \tag{17}$$

where the number 2 in parentheses refers to the second used model. It is noteworthy that the probabilities $\varphi_{1,12(2)}$ and $\varphi_{2,12(2)}$ are the same as those calculated under the one-population-random-mating model (MODEL T1).

Compared to model T1, the above equations show that a general effect of positive inbreeding is to decrease the

probabilities of transmission from heterozygous parents to the homozygous offspring. The probabilities of transmission to the heterozygous offspring are not affected by inbreeding. In other words, because of the reduction in heterozygosity under positive inbreeding, the number of alleles transmitted from heterozygous parents will be reduced under model T2 compared to model T1.

The expectations and variances of the number of alleles A_1 and A_2 transmitted from heterozygous parents to the affected offspring can be estimated by equations (8)–(11). The testing of the null hypothesis of nondifferential transmission can be performed using the method in the one-population-random-mating model.

Several populations with random mating within them (model T3)

Let us suppose that the population under study is divided into k subpopulations of equal size and are in Hardy–Weinberg equilibrium. If the subpopulation source of each sampled individual is unknown, a correction that takes into account the allele frequency differences between the subpopulations must be performed and the $\varphi_{i,ij}$ probabilities must be averaged across all the subpopulations. After averaging the $\varphi_{i,ij}$ probabilities, we obtain

$$\varphi_{1,11(3)} = p_2\left(\frac{p_1(1-3F_{ST}) + F_{ST}(1-\gamma_1\sqrt{F_{ST}p_1p_2})}{p_1(1-F_{ST}) + F_{ST}}\right) \quad (18)$$

$$\varphi_{1,12(3)} = \frac{p_2(1-3F_{ST}) + F_{ST}(1+\gamma_1\sqrt{F_{ST}p_1p_2})}{1-F_{ST}}, \quad (19)$$

$$\varphi_{2,12(3)} = \frac{p_1(1-3F_{ST}) + F_{ST}(1+\gamma_2\sqrt{F_{ST}p_1p_2})}{1-F_{ST}} \quad (20)$$

$$\varphi_{2,22(3)} = p_1\left(\frac{p_2(1-3F_{ST}) + F_{ST}(1-\gamma_2\sqrt{F_{ST}p_1p_2})}{p_2(1-F_{ST}) + F_{ST}}\right) \quad (21)$$

where p_1 and p_2 refer to the average gene frequencies of the alleles A_1 and A_2 in the total population, and the number 3 in parentheses refers to the third studied model. F_{ST} is the correlation between two gametes drawn at random from each subpopulation, and it measures the degree of genetic differentiation between subpopulations.[9,10] F_{ST} is equal to σ_p^2/p_1p_2, where σ_p^2 is the variance of the allele frequencies across the subpopulations. γ_1 and γ_2 are values of skewness of the frequency distributions of the alleles A_1 and A_2 over all the subpopulations; for a diallelic locus $\gamma_1 = -\gamma_2$.

The expectations and variances of the number of alleles A_1 and A_2 transmitted from heterozygous parents to the affected offspring can be estimated using the corresponding $\varphi_{i,ij(3)}$ probabilities according to equations (8)–(11). The hypothesis testing of the nondifferential transmission can be done as explained in the last two models.

Several populations without random mating within them (model T4)

If the population under study is divided into k subpopulations of equal size and within them the Hardy–Weinberg equilibrium does not hold, the effects of population structure and inbreeding must be taken into account. The $\varphi_{i,ij}$ probabilities within each subpopulation are given according to the one-population-nonrandom-mating model (T2); therefore, the corresponding values in the general population are the weighted averages of the $\varphi_{i,ij}$'s across the subpopulations. Namely,

$$\varphi_{1,11(4)} = \varphi_{1,11(3)}(1-F_{IS})\left(\frac{p_1(1-F_{ST}) + F_{ST}}{p_1(1-F_{IT}) + F_{IT}}\right) \quad (22)$$

$$\varphi_{1,12(4)} = \varphi_{1,12(3)} \quad (23)$$

$$\varphi_{2,12(4)} = \varphi_{2,12(3)} \quad (24)$$

$$\varphi_{2,22(4)} = \varphi_{2,22(3)}(1-F_{IS})\left(\frac{p_2(1-F_{ST}) + F_{ST}}{p_2(1-F_{IT}) + F_{IT}}\right) \quad (25)$$

where F_{IT} is the correlation between two uniting gametes relative to the total population and measures the reduction in heterozygosity of an individual relative to the total population. F_{IT} takes into account both the effects of nonrandom mating within subpopulations (F_{IS}) and the effects of population subdivision (F_{ST}), and they are related by the equation $(1-F_{IT}) = (1-F_{IS})(1-F_{ST})$.[10] It is noteworthy that the probabilities $\varphi_{1,12(4)}$ and $\varphi_{2,12(4)}$ are not affected by nonrandom mating within subpopulations, and they are the same as the corresponding probabilities on the several populations-random-mating model.

The expectations and variances of the number of alleles A_1 and A_2 transmitted from heterozygous parents to the affected offspring can be estimated using the corresponding $\varphi_{i,ij(4)}$ probabilities according to equations (8)–(11). The hypothesis testing of the nondifferential transmission can be done as explained in the last three models.

Results

As previously described, there are several parameters that determine the value of the probabilities $\varphi_{i,ij}$. These include the frequency of the suspected allele in the general population (p_1), the frequency of the same allele in the case sample (q_1), and the different F values. To examine the behavior of the T-value, we calculated the probabilities $\varphi_{i,ij}$ under three general conditions:

1. The p_1 and q_1 frequencies were allowed to change, but the increased frequency of the allele A_1 in cases relative to general population was fixed at 20%.

2. The frequency of A_1 in the general populations was fixed at 0.40, and the frequency of A_1 in the cases was allowed to take various values.
3. The p_1 and q_1 were, respectively, fixed to 0.40 and 0.50, but the vector (P, H, Q) was variable.

A drawback in using the different models in a case–control study is that the total departure from Hardy–Weinberg proportions in the general population can only be estimated using the control genotypes. In other words, only one estimate of the F_{IT} value can be obtained, yet different F_{IS}, F_{ST} pairs can result in the same F_{IT}. Therefore, if we only know the total reduction of heterozygosity in the general population (F_{IT}), we are not aware of the differential contributions of both the effects of nonrandom mating within subpopulations (F_{IS}) and the effects of population subdivision (F_{ST}). A conceivable solution to this problem is to estimate the T-value under two different assumptions: one, that there is no population subdivision ($F_{ST}=0$) and that all the departure from Hardy–Weinberg proportions is due to the effects of nonrandom mating within the population ($F_{IS}=F_{IT}>0$); two, the general population is subdivided ($F_{ST}=F_{IT}>0$), but there is random mating within each subpopulation ($F_{IS}=0$).

Figure 1 shows the calculated T values by fixing the relative increase in the frequency of the allele A_1 in cases relative to the general population at 20%. Since the models T2 and T3 are particular cases of the model T4, they are the upper and lower bounds of the estimated T-values according to the model T4. It is noteworthy that the effects of changes in the F_{IT} are greater in model T3 than in model T2. In other words, model T2 is less affected by changes in the structure of the population. One reason for this result is that nonrandom mating within subpopulations does not affect the probabilities that one heterozygous case has received alleles from heterozygous parents (see equations (15) and (16)).

If we know the F_{IT} value, but do not know the F_{IS} and F_{ST}, the difference between the estimates of T according to models T2 and T3 can be understood as the degree of accuracy in our estimates. As the effect of nonrandom mating on the T-value is lesser than the effect of the subpopulation differentiation (Figure 1), our main concern is the latter factor. According to Figure 1, model T2 and model T3 produce similar estimates of the T-value when F_{IT} values are less or equal to 0.25 and allele frequencies are at least 0.10. Slatkin[11,12] has shown that the ratio $F_{ST}/(1-F_{ST})$ is roughly proportional to the divergence time between two populations, and it is equal to $\tau/2Ne$, where τ is the divergence time in generations and Ne is the effective population size of each subpopulation, assuming equal sizes for both. For example, for an F_{ST} equal to 0.25 and an effective size equal to 500, the expected divergence time is $\sim$300 generations or $\sim$7500 years, assuming 25 years per generation. Although these estimates of divergence times

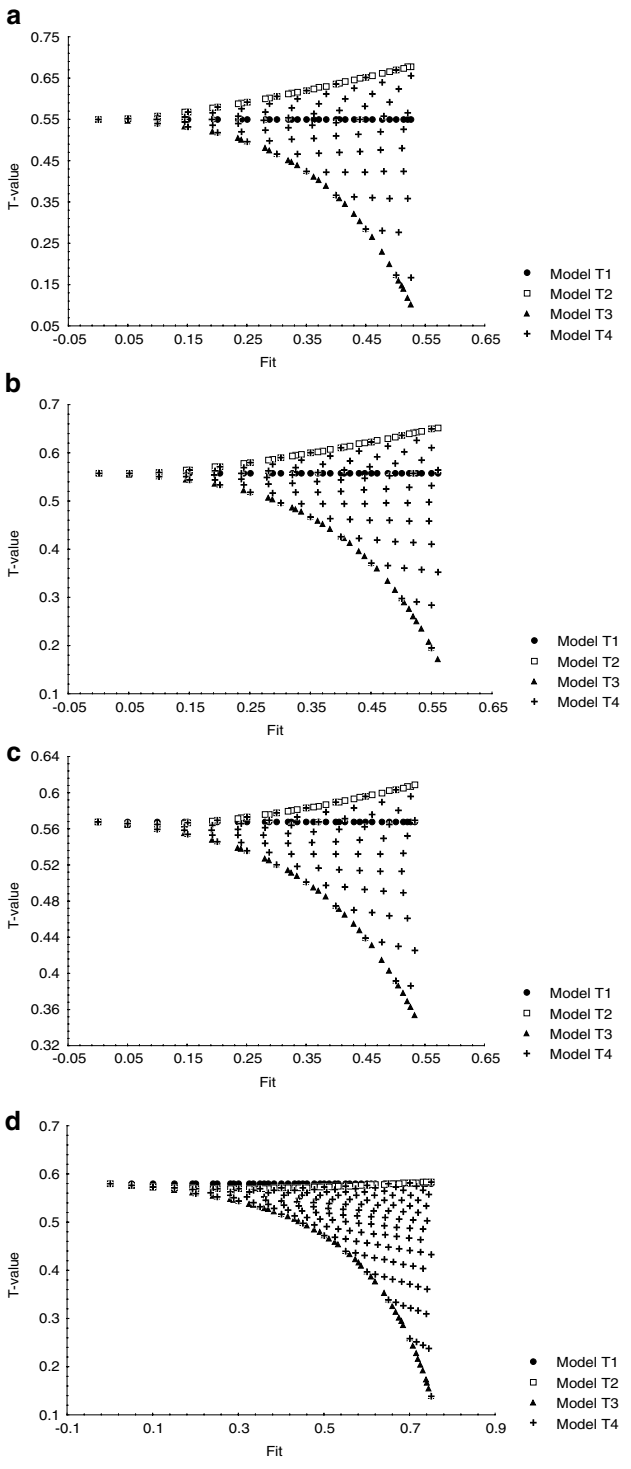

Figure 1 Graphs show the effect of the total population inbreeding (F_{IT}) on the T-value when the increase in the frequency of the high-risk allele in cases is fixed at 20% relative to controls, and the allele frequencies in cases and controls are allowed to change: (**a**) $p_1=0.10$ and $q_1=0.12$; (**b**) $p_1=0.20$ and $q_1=0.24$; (**c**) $p_1=0.30$ and $q_1=0.36$; (**d**) $p_1=0.40$ and $q_1=0.48$.

must be taken with caution because of their unknown error, they do indicate that for small divergence times, the models T2, T3, and T4 produce very similar estimates of the *T*-value.

Another factor affecting the *T*-value is the degree of increase in the frequency of the high-risk allele in cases relative to the general population. To explore this effect, we fixed p_1 at 0.40, while q_1 was allowed to change from 0.50 to 0.80. According to Figure 2, a $F_{IT} \leqslant 0.30$ combined with any allele frequency in cases produce very similar estimates of the *T*-value in the three models T2, T3, and T4. Although we chose a frequency of 0.40 for the high-risk allele in the general population, less frequent alleles gave the same results (data not shown).

If the allele frequencies are fixed both in the general population and in the cases, there is still another factor that can influence the estimates of the *T*-value; that is, the vector (*P*, *H*, *Q*) in cases, or the number of people with genotypes A_1A_1, A_1A_2, and A_2A_2, respectively. For example, a sample with $P = 250$, $H = 500$, and $Q = 250$ has a different (*P*, *H*, *Q*) vector from a sample with $P = 300$, $H = 400$, and $Q = 300$; yet both samples have the same allele frequencies. The rationale to analyze the (*P*, *H*, *Q*) vector is because if the population from where the cases are sampled is structured, we expect to find a similar degree of structure in the case sample. Figure 3 shows the effect of several combinations of *P*, *H*, and *Q* by fixing, respectively, the frequency of the high-risk allele to 0.40 and 0.50 in the general population and in cases. A striking result is that the three models T2, T3, and T4 converge to the same *T*-value when the degree of departure from Hardy–Weinberg proportions is the same in both cases and the general population. This result has a very important practical implication. For example, a highly recommended first analysis in a case–control study is to test if the Hardy–Weinberg equilibrium holds both in the case and control samples. As it was shown, any deviations from Hardy–Weinberg proportions affect the different probabilities of transmission from heterozygous parents, but, if the factors controlling the genotype frequencies have the same effect both in the general and the affected populations, the differences between the models T2, T3, and T4 disappear.

In most of the situations, the exact structure of the population under study is unknown and the F_{IT} value is commonly the sole available estimate of population stratification. Since the true model may be unknown in the majority of the studies, we assessed type I error and power of the four different models under different population scenarios (Table 2). We assumed the same F_{IT} for both general and affected populations; 100 000 simulations were used and a random sample of 100 cases was taken in each simulation. The frequency of the high-risk allele in the general population was fixed to 0.10. Higher allele frequencies did not affect appreciably type I error, but increased the power of the different models (data not

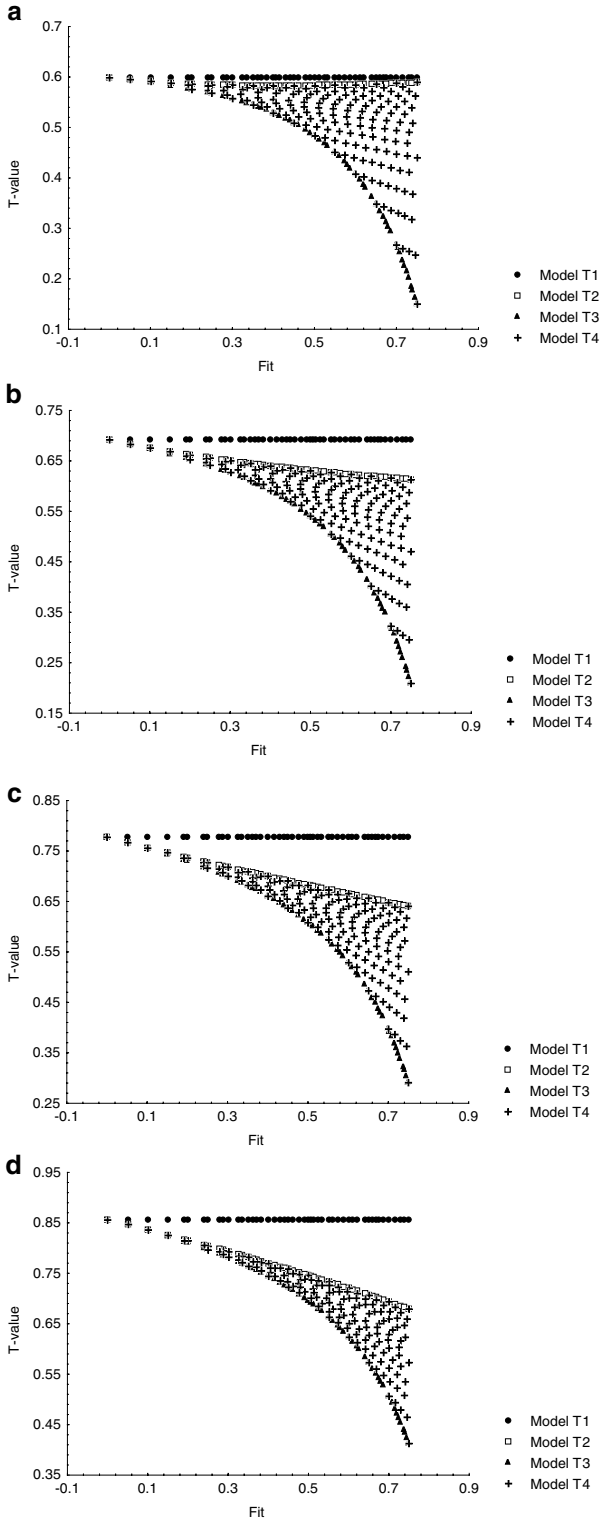

Figure 2 Graphs show the effect of the total population inbreeding (F_{IT}) on the *T*-value when the frequency of the high-risk allele in controls is fixed at 0.40, and the allele frequencies in the cases are allowed to change: (**a**) $q_1 = 0.50$; (**b**) $q_1 = 0.60$; (**c**) $q_1 = 0.70$; (**d**) $q_1 = 0.80$.

shown). The critical point to reject the null hypothesis ($\Delta = 0$, see equation (13)) was chosen to attain a significance level of 0.05 ($\sim \pm 1.96$ standard deviation) for the true model under the different population scenarios. In the simplest situation, there is no population subdivision and there is random mating within a unique population ($F_{IS} = F_{ST} = F_{IT} = 0$). The four models were equivalent in this first population setting. With nonrandom mating within a single population ($F_{IS} = F_{IT} > 0$, $F_{ST} = 0$), the true model T2 was equal to the model T4. In this case, model T1 showed a higher type I error ($\sim 7\%$) and model T3 had a lower significance level (~ 2–4%). The power of the four models was similar under this second population scenario and did not change greatly for F_{IT} values greater than 0.10 (data not shown). When the population is subdivided, but there is random mating within each subpopulation ($F_{ST} = F_{IT} > 0$, $F_{IS} = 0$), the true model T3 was equal to the model T4. T1 and T2 models tended to give a higher false-positive rate. Although the power of the four models was similar, model T1 showed slightly superior power due to its high type I error (~ 7–11%). Large differences were not observed for F_{IT} values greater than 0.10 (data not shown). In the last situation, several subpopulations with nonrandom mating within them ($F_{IS} > 0$, $F_{ST} > 0$, $F_{IT} > 0$); the T1, T2, and T3 models showed a higher false-positive rate compared to the true model T4. However, the difference between the model T3 and the true model T4 was at most 1.5% for $F_{IT} = 0.10$. Similar results were observed for greater values of population stratification (data not shown).

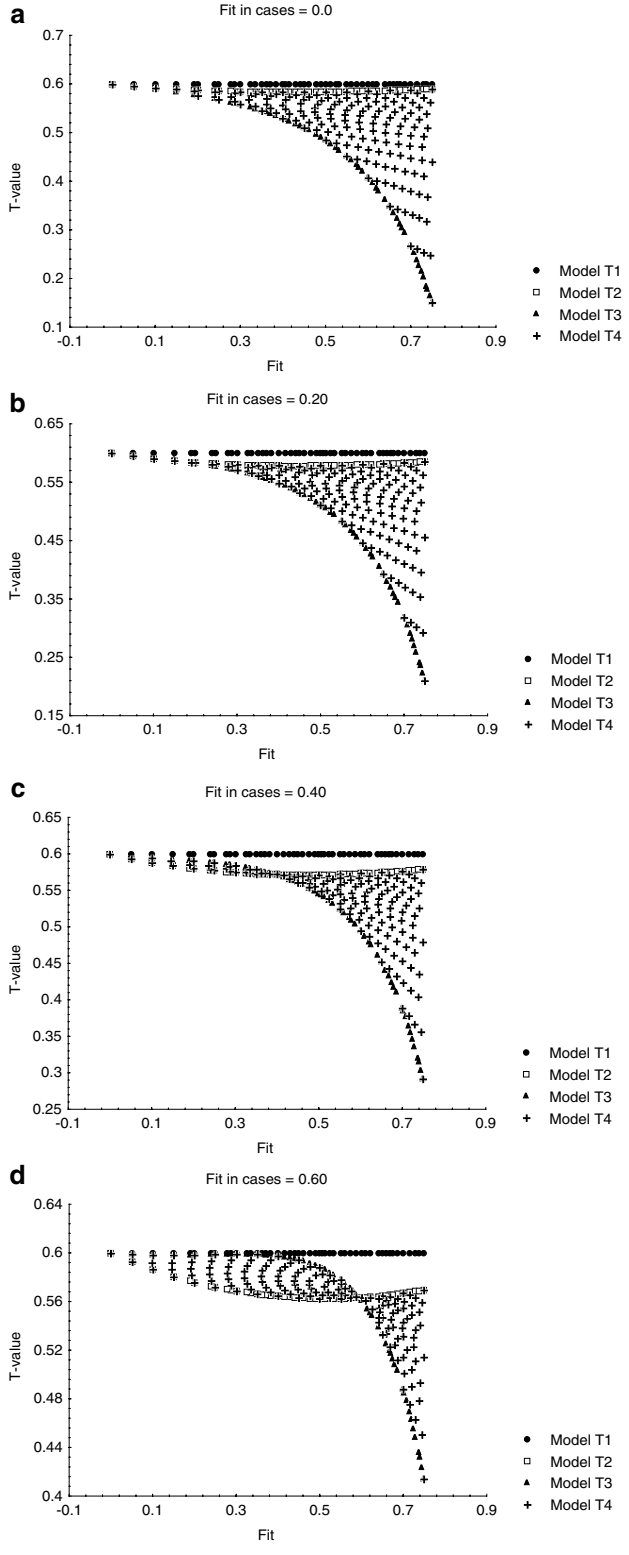

Figure 3 Graphs show the effect of the total population inbreeding (F_{IT}) on the *T*-value when the vector (*P*, *H*, *Q*) is variable in cases. The number of cases was fixed to 1000 is varuable with $p_1 = 0.40$, $q_1 = 0.50$, and different vectors. (**a**) Vector (250, 500, 250); (**b**) vector (300, 400, 300); (**c**) vector (350, 300, 350); (**d**) vector (400, 200, 400).

Conclusions

We presented a novel population-based method to estimate the differential transmission of marker alleles from heterozygous parents to affected offspring. This method expands the previous procedure proposed by Mitchell[8] by providing an estimate of the differential transmission of alleles from heterozygous parent to their affected offspring and a statistical test that can be used under different population structures. Unlike Mitchell's approach, the current test does not assume Hardy–Weinberg proportions and can be adjusted to deal with deeper levels of population stratification. This method attains the same goals of the traditional family-based TDT, but it eliminates the need for recruitment of family members that are particularly difficult to obtain in late-onset diseases or large population-based epidemiological studies. Our analyses showed that model T3 was the most robust approach under several scenarios of population structure. Further-

Table 2 Type I error and power of the Δ-statistic with a random sample of 100 cases under four different population scenarios, and a high-risk allele frequency of 0.10 in the general population

		Power to reject different alternative hypothesis (H_a:$T = T_a$)		
Population scenario	*Type I error*[a]	0.55	0.60	0.65
One population, random mating				
$F_{IS} = F_{ST} = F_{IT} = 0.0$				
Model T1 (T2,T3,T4)	0.050	0.188	0.562	0.906
One population, nonrandom mating				
$F_{IS} = F_{IT} = 0.05$, $F_{ST} = 0.0$				
Model T1	0.066	0.281	0.694	0.957
Model T2 (T4)	0.050	0.246	0.640	0.934
Model T3	0.040	0.235	0.643	0.941
$F_{IS} = F_{IT} = 0.10$, $F_{ST} = 0.0$				
Model T1	0.073	0.232	0.664	0.953
Model T2 (T4)	0.050	0.170	0.546	0.888
Model T3	0.024	0.135	0.533	0.913
Several populations, random mating				
$F_{ST} = F_{IT} = 0.05$, $F_{IS} = 0.0$				
Model T1	0.077	0.309	0.706	0.962
Model T2	0.060	0.278	0.675	0.942
Model T3 (T4)	0.050	0.258	0.677	0.950
$F_{ST} = F_{IT} = 0.10$, $F_{IS} = 0.0$				
Model T1	0.113	0.316	0.770	0.980
Model T2	0.085	0.243	0.663	0.943
Model T3 (T4)	0.050	0.215	0.670	0.960
Several populations, nonrandom matings				
$F_{IT} = 0.05$, $F_{IS} = F_{ST} = 0.025$				
Model T1	0.091	0.355	0.779	0.979
Model T2	0.071	0.325	0.737	0.965
Model T3	0.060	0.301	0.739	0.972
Model T4	0.050	0.278	0.706	0.963
$F_{IT} = 0.10$, $F_{IS} = F_{ST} = 0.051$				
Model T1	0.143	0.406	0.840	0.993
Model T2	0.112	0.332	0.754	0.976
Model T3	0.065	0.294	0.764	0.986
Model T4	0.050	0.232	0.699	0.975

[a]The critical point to reject the null hypothesis corresponds to a significance level of 0.05 for the true model under each population scenario.

more, for F_{IT} values of 0.30 or less, all the models tested gave very similar estimates of the *T*-value. The usefulness of this procedure can be confirmed by comparing the results obtained from traditional family-based TDT with those from population-based TDT.

Population association studies between two loci that are linked are widely used to map loci affecting complex traits.[13,14] One of the major advantages of the population-based studies over the family approaches is that they avoid recruitment of family members, which is particularly difficult when studying diseases with late onset. Case–control studies have often provided the first line of evidence that a putative disease susceptibility locus or a marker in linkage disequilibrium exists; for example, the observed association between the *APOE* genotype and Alzheimer's disease.[15] However, the use of the case–control approach to uncover disease–marker associations has been disappointing. In general, initial reports of strong associations cannot be reproduced or are not supported by larger well-conducted studies.[14,16] The results have been inconsistent perhaps due to modest gene effects *per se*, but more likely to problems with study design such as low statistical power or the lack of a comparable control population to determine the underlying gene frequencies. Another explanation that can limit the validity of the epidemiological case–control design relates to the potential for confounding that can result from population stratification or genetic admixture.[17–19] For example, if the population under study is heterogeneous, if mating does not occur randomly, or if the cases and controls are not ethnically

balanced, a coincidental allele frequency difference can emerge. Such an artifact is most likely to happen when the disease occurs more frequently in an unidentified subpopulation, which also differs, by chance, in the frequency of the tested allele.

To correct for population stratification, it is necessary to detect it and quantify it. Pritchard and Rosenberg[20] have shown that approximately 15–20 unlinked microsatellite loci are needed to test for stratification, and hundreds of markers are required to identify subpopulations of recent divergence time.[21,22] Here we proposed a simpler approach to measure population stratification. Although it is desirable to know the exact genetic structure of the population under study, a simple measure such as F_{IT} of total departure of Hardy–Weinberg proportions is useful as a rough estimate of population stratification. Since our method, specifically model T3, is robust to major departures from Hardy–Weinberg proportions, either by nonrandom mating within subpopulations or population subdivision, the proposed test can be used in multiple epidemiological settings. By using a few unlinked loci, Wright's *F*-statistics[9,10] provide a simple and reliable way to assess the genetic structure of the population under study. A limitation of the use of the *F*-statistics is that without a priori information about the population structure, it would be impossible to discriminate between the effects of the nonrandom mating within subpopulations and those due to the population subdivision. An approximation can be done by assuming that the total departure from Hardy–Weinberg proportions in the general population (F_{IT}) is caused entirely by either nonrandom mating within subpopulations ($F_{ST}=0$, $F_{IS}=F_{IT}>0$) or population subdivision ($F_{IS}=0$, $F_{ST}=F_{IT}>0$). Based on the simulations presented in this study, both approximations can provide essentially the same estimates of the *T*-value when the values for F_{IT} are as large as 0.30, but type I error of the different models may change depending on the exact population structure. Specifically, in the presence of nonrandom mating or population subdivision, model T1 overestimates the number of alleles transmitted from heterozygous parents, leading to a higher type I error. Model T3 showed the lowest false-positive rate ($\sim$2.4–6.5%) with no significant lack of power compared to the true model under the different population scenarios. In other words, when the exact genetic structure of the population is unknown, the most conservative approach is to assume that total departure from Hardy–Weinberg proportions is only due to the effects of population subdivision. Also, due to theoretical reasons, we expect that, within subpopulations, most of the matings occur independently of the marker genotypes; therefore, population stratification will be the main factor affecting the F_{IT} value. Although the degree of genetic differentiation between populations depends, in part, on the marker mutation rate, it has been shown that differences among major human groups constitute only 3–9% of the total genetic variation by using either high-mutation rate loci as microsatellites[23] or low-mutation rate loci as allozyme genes.[24] Therefore, it is very unlikely that F_{IT} exceeds a value of 0.30 in a population-based approach with careful matching for ethnic background. Even though the proposed method can be used when knowledge about population subdivision is limited, this approach is flexible enough to incorporate demographic or historic data about the structure of the population when available. With this approach, it will be also possible to carry out an analysis of molecular variance (AMOVA)[25] to quantify the different components of variance that contribute to the total inbreeding in the general population.

Although the proposed method uses population data, the logistics of this procedure diverge greatly from the traditional association studies. Traditional association studies are designed to test for differences in the gene frequencies between cases and controls. In contrast, the proposed method evaluates the differential transmission of marker alleles from heterozygous parents to the affected offspring. In fact, it is possible, depending on the model used, to have no differences in the allele marker frequencies between cases and controls, and obtain a *T*-value different from $\frac{1}{2}$. It is necessary to emphasize that because the controls are used to estimate population parameters (frequency of the high-risk allele and F_{IT}), a properly conducted study should require that controls represent the population that give rise to the cases. Thus, standard epidemiological methods such as matching by ethnic background must be used in the selection of the controls.

In summary, we presented a population-based TDT that attains the same goals of the traditional family-based TDT but without the recruitment of family members that are particularly difficult to obtain in late-onset diseases. By using simulations, model T3 proved to be the most robust approach under several scenarios of population structure. Although it can be useful to know the exact genetic structure of the sample population, different models gave very similar estimates of the *T*-value for F_{IT} values of 0.30 or less. Further work is needed to compare the results obtained from traditional family-based TDT and population-based TDT.

Acknowledgements

This study was supported by grant HL60692 from the National Institutes of Health.

References

1 Peyser PA: Genetic epidemiology of coronary artery disease. *Epidemiol Rev* 1997; **19**: 80–90.
2 Willett WC: Balancing life-style and genomics research for disease prevention. *Science* 2002; **296**: 695–698.

3 Ellsworth DL, Manolio TA: The emerging importance of genetics in epidemiologic research II. Issues in study design and gene mapping. *Ann Epidemiol* 1999a; **9**: 75–90.
4 Ellsworth DL, Manolio TA: The emerging importance of genetics in epidemiologic research III. Bioinformatics and statistical genetic methods. *Ann Epidemiol* 1999b; **9**: 207–224.
5 Clark AG: Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr Opin Genet Dev* 2003; **13**: 296–302.
6 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–516.
7 Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996; **59**: 983–989.
8 Mitchell LE: Relationship between case–control studies and the transmission/disequilibrium test. *Genet Epidemiol* 2000; **19**: 193–201.
9 Wright S: The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 1965; **19**: 395–420.
10 Wright S: The theory of gene frequencies; in *Evolution and the Genetics of Populations*. Chicago: University of Chicago Press, 1969, Vol 2.
11 Slatkin M: Inbreeding coefficients and coalescence times. *Genet Res* 1991; **58**: 167–175.
12 Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 1995; **139**: 457–462.
13 Botto LD, Yang Q: 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. *Am J Epidemiol* 2000; **151**: 862–877.
14 Eichner JE, Dunn ST, Perveen G, Thompson DM, Stewart KE, Stroehla BC: Apolipoprotein E polymorphism and cardiovascular disease: a HuGE review. *Am J Epidemiol* 2002; **155**: 487–495.
15 Strittmatter WJ, Saunders AM, Schmechel D *et al*: Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci USA* 1993; **90**: 1977–1981.
16 Hassan A, Markus HS: Genetics and ischaemic stroke. *Brain* 2000; **123**: 1784–1812.
17 Cox NJ, Bell GI: Disease association. Chance, artifact, or susceptibility genes? *Diabetes* 1989; **38**: 947–950.
18 Lander ES, Schork NJ: Genetic dissection of complex traits. *Science* 1994; **265**: 2037–2048.
19 Ewens WJ, Spielman RS: The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995; **57**: 455–464.
20 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–228.
21 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am J Hum Genet* 2000; **67**: 170–181.
22 Pritchard JK, Donnelly P: Case–control studies of association in structured or admixed populations. *Theor Popul Biol* 2001; **60**: 227–237.
23 Rosenberg NA, Pritchard JK, Weber JL *et al*: Genetic structure of human populations. *Science* 2002; **298**: 2381–2385.
24 Nei M, Roychoudhury AK: Genetic relationship and evolution of human races. *Evol Biol* 1982; **14**: 1–59.
25 Excoffier L, Smouse PE, Quattro JM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479–491.