The Relative Power of Family-Based and Case-Control Designs for Linkage Disequilibrium Studies of Complex Human Diseases. II. Individual Genotyping

Jun Teng and Neil Risch

Genome Res. 1999 9: 234-241
Access the most recent version at doi:10.1101/gr.9.3.234

The Relative Power of Family-Based and Case-Control Designs for Linkage Disequilibrium Studies of Complex Human Diseases. II. Individual Genotyping

Jun Teng[1] and Neil Risch[1–4]

[1]Department of Statistics, Stanford University and Departments of [2]Genetics and [3]Health Research and Policy, Stanford University School of Medicine, Stanford, California 94305 USA

In this paper we consider test statistics based on individual genotyping. For sibships without parents, but with unaffected as well as affected sibs, we introduce a new test statistic (referred to as T_{DS}), which contrasts the allele frequency in affected sibs versus that estimated for the parents from the entire sibship. For sibships without parents, this test is analogous to the TDT and is completely robust to nonrandom mating patterns. The efficiency of the T_{DS} test is comparable to that of the T_{HS} test (which compares affected vs. unaffected sibs and was based on DNA pooling), for sibships with one affected child. However, as the number of affected sibs in the sibship grows, the relative efficiency of the T_{DS} test versus the T_{HS} test also increases. For example, for sibships with three affected, one-third fewer families are required; for families with four affected, nearly half as many are required. Thus, when sibships contain multiple affected individuals, the T_{DS} test provides both an increase in power and robustness to nonrandom mating.

In the first paper in this series, Risch and Teng (1998), we considered statistics based on data derived from DNA pooling. Only overall allele frequency estimates for a pool are available from such experiments; hence, only statistics based on pooled allele frequencies are possible, such as the haplotype-based haplotype relative risk (HHRR) (Falk and Rubinstein 1987; Terwilliger and Ott 1992). Such statistics are not automatically robust to nonrandom mating, although they are conservative under population stratification. Furthermore, such statistics may not extract all the available information in some study designs if individual genotyping is performed. Therefore, in this paper we consider analyses of data obtained from individual genotyping of all study subjects. We compare the same family constellations as described in Risch and Teng (1998). As individual genotyping provides more information than DNA pooling, it enables us to improve the statistical treatment in two ways: by increasing robustness and power.

We consider statistics of the form $(\hat{p}_1 - \hat{p}_2)/\hat{\sigma}$, in which the numerator contrasts the estimated allele frequencies in two groups (affected sibs vs. parents) and the denominator is the estimated standard deviation of the numerator. Typically, the variance of $(\hat{p}_1 - \hat{p}_2)$ is a function of genotype frequencies in the parents. When DNA pooling has been performed, this variance has to be estimated based on the assumption of Hardy–Weinberg equilibrium. On the other hand, individual genotyping allows us to get an unbiased estimate of the variance under more general conditions and thus provides further robustness to non-random mating. More importantly, in the case where parents are unavailable, individual genotyping gives us a greater choice of the contrast we can make in the numerator, which potentially can improve the power of the test.

Study designs that include affected offspring with parents lend themselves to the calculation of a TDT statistic, provided individual genotyping is performed. Although the TDT offers additional robustness to nonrandom mating in this case, the power of this test statistic is generally comparable to that of the HHRR statistic, at least when mating is nearly at random. This is because the Hardy–Weinberg estimator of parental heterozygosity, used in the denominator of the HHRR statistic, is close to the directly counted parental heterozygosity estimate used in the TDT (Risch and Teng 1998, formula 4). Thus, sample size requirements using individual genotyping for designs involving affected offspring with parents, based on TDT, are essentially identical to those we have presented previously (Risch and Teng 1998) for the same designs based on DNA pooling and HHRR statistics (calculations performed but not presented). Therefore, we use the sample size requirements for affected sibships with parents derived in Risch and Teng (1998) for comparison with individually genotyped sibships without parents.

Received January 7, 1998; acceped in revised form January 20, 1999.
[4]Corresponding author.
EMAIL risch@lahmed.stanford.edu; FAX (650) 725-1534.

9:234–241 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

In the classic TDT, p_1 is the allele frequency in the affected child (or children) and p_2 the allele frequency in the parents. For sibships without parents, the test described in Risch and Teng (1998) proposes p_1 to be the allele frequency in the affected sibs, and p_2 the allele frequency in the unaffected sibs. When the locus-related penetrance is low, the allele frequency p_2 in unaffected sibs can also be viewed as providing a nearly unbiased estimate of the allele frequency in the parents (in this sense, it is similar to the TDT, in which p_2 is the observed allele frequency in the parents). When more than one child has been individually genotyped, however, it is possible to obtain a more efficient estimate of the parent allele frequency p_2, as well as an estimate of the variance of $\hat{p}_1 - \hat{p}_2$ that is robust to nonrandom mating. We derive such a statistic below and describe its properties.

We use the same notation as given in Risch and Teng (1998); namely, m_{ij} denotes the conditional probability of mating type (i,j) given an affected child (and similarly $m^{(r)}_{ij}$ for r affected children), in which i and j are the number of A alleles in the two parents (we use parentheses in subscripts to denote unordered genotypes); f_k is the ratio of penetrance in individuals with k D alleles compared with dd individuals; hats over letters (circumflexes) denote sample estimates. To simplify some formulas, we also introduce the following notation:

$$c_{21} = \frac{f_2}{f_2 + f_1},\ c_{20} = 1 - c_{21} = \frac{f_1}{f_2 + f_1},$$

$$c_{01} = \frac{f_1}{f_1 + 1},\ c_{00} = 1 - c_{01} = \frac{1}{f_1 + 1},$$

$$c_{12} = \frac{f_2}{f_2 + 2f_1 + 1},\ c_{11} = \frac{2f_1}{f_2 + 2f_1 + 1},\ c_{10} = \frac{1}{f_2 + 2f_1 + 1}$$

We assume, as in Risch and Teng (1998), that unaffected sibs have a random genotype distribution (low penetrance) given the parental mating type.

Affected–Unaffected Sib Pairs

We first examine the case of one affected and one unaffected sib, without parents. For this case, there are nine possible marker genotype outcomes for the sib pair, as listed in Table 1, along with their probabilities of occurrence. To estimate the frequency of allele A in the parents (p_2), we notice that under the null hypothesis, $f_2 = f_1 = 1$ and the affected and unaffected sibs become symmetric; so Table 1 can be simplified to six possible outcomes: (1) Both sibs are *AA;* (2) both sibs are *aa;* (3) both sibs are *Aa;* (4) one is *AA,* the other is *Aa;* (5) one is *Aa,* the other is *aa;* and (6) one is *AA,* the other is *aa*. There are also the same six possible genotype combinations (mating types) for the parents with respective probability $m_{(ij)}$. Because there is an equal number of parameters and independent observations, maximum likelihood estimates of the parental mating type frequencies $m_{(ij)}$ can be calculated by equating the sample frequency of each sib-pair outcome with its respective probability, namely

$$n_{22}/n = \hat{m}_{22} + \hat{m}_{(21)}/4 + \hat{m}_{11}/16$$
$$n_{00}/n = \hat{m}_{00} + \hat{m}_{(10)}/4 + \hat{m}_{11}/16$$
$$n_{11}/n = \hat{m}_{(20)} + \hat{m}_{(21)}/4 + \hat{m}_{(10)}/4 + \hat{m}_{11}/4$$
$$n_{21}/n + n_{12}/n = \hat{m}_{(21)}/2 + \hat{m}_{11}/4$$
$$n_{10}/n + n_{01}/n = \hat{m}_{(10)}/2 + \hat{m}_{11}/4$$
$$n_{20}/n + n_{02}/n = \hat{m}_{11}/8$$

Solving these equations, we get the unbiased maximum likelihood estimators $\hat{m}_{ij}$. These are given by

$$\hat{m}_{11} = 8(n_{20} + n_{02})/n$$
$$\hat{m}_{(10)} = [2(n_{10} + n_{01}) - 4(n_{20} + n_{02})]/n$$
$$\hat{m}_{(21)} = [2(n_{21} + n_{12}) - 4(n_{20} + n_{02})]/n$$
$$\hat{m}_{00} = [2n_{00} - (n_{10} + n_{01}) + (n_{20} + n_{02})]/2n$$
$$\hat{m}_{22} = [2n_{22} - (n_{21} + n_{12}) + (n_{10} + n_{01})]/2n$$
$$\hat{m}_{(20)} = [2n_{11} - (n_{21} + n_{12}) + (n_{10} + n_{01})]/2n$$

Table 1. Genotype Outcomes, Scores, and Probabilities for Affected–Unaffected Sib Pair

		Sample		
Affected sib	Unaffected sib	frequency	score (S)	probability
AA	*AA*	n_{22}	0	$m_{22} + m_{(21)}f_2 / 2(f_2 + f_1) + m_{11}f_2 / 4(f_2 + 2f_1 + 1)$
AA	*Aa*	n_{21}	1/4	$m_{(21)}f_2 / 2(f_2 + f_1) + m_{11}f_2 / 2(f_2 + 2f_1 + 1)$
AA	*aa*	n_{20}	1/2	$m_{11}f_2 / 4(f_2 + 2f_1 + 1)$
Aa	*AA*	n_{12}	−1/4	$m_{(21)}f_1 / 2(f_2 + f_1) + m_{11}f_1 / 2(f_2 + 2f_1 + 1)$
Aa	*Aa*	n_{11}	0	$m_{(21)}f_1 / 2(f_2 + f_1) + m_{(20)} + m_{11}f_1 / 2(f_2 + 2f_1 + 1) + m_{(10)}f_1 / 2(f_1 + 1)$
Aa	*aa*	n_{10}	1/4	$m_{(10)}f_1 / 2(f_1 + 1) + m_{11}f_1 / (f_2 + 2f_1 + 1)$
aa	*AA*	n_{02}	−1/2	$m_{11} / 4(f_2 + 2f_1 + 1)$
aa	*Aa*	n_{01}	−1/4	$m_{(10)} / 2(f_1 + 1) + m_{11} / 2(f_2 + 2f_1 + 1)$
aa	*aa*	n_{00}	0	$m_{(10)} / 2(f_1 + 1) + m_{11} / 4(f_2 + 2f_1 + 1) + m_{00}$

Then the frequency of A in the parents can be estimated by

$$\hat{p}_2 = \hat{m}_{22} + \tfrac{3}{4}\hat{m}_{(21)} + \tfrac{1}{2}\hat{m}_{11} + \tfrac{1}{2}\hat{m}_{(20)} + \tfrac{1}{4}\hat{m}_{(10)}$$
$$= [n_{22} + \tfrac{3}{4}(n_{21} + n_{12}) + \tfrac{1}{2}(n_{11} + n_{20} + n_{02}) + \tfrac{1}{4}(n_{10} + n_{01})]/n$$

which, in this case, is the same as the A allele frequency in the combined sibling sample. Because

$$\hat{p}_1 = [n_{22} + n_{21} + n_{20} + \tfrac{1}{2}(n_{12} + n_{11} + n_{10})]/n$$

we have

$$\hat{p}_1 - \hat{p}_2 = [(n_{21} - n_{12}) + (n_{10} - n_{01}) + 2(n_{20} - n_{02})]/4n.$$

The variance of $\hat{p}_1 - \hat{p}_2$ is a function of h, the frequency of heterozygosity in the parents. Whereas DNA pooling required us to use the Hardy-Weinberg assumption in the estimation of h (formula 5 of Risch and Teng 1998), individual genotyping allows us to obtain a more direct estimate, robust to nonrandom mating. Specifically,

$$\hat{h} = \hat{m}_{11} + \tfrac{1}{2}\hat{m}_{(21)} + \tfrac{1}{2}\hat{m}_{(10)}$$
$$= [n_{21} + n_{12} + n_{10} + n_{01} + 4(n_{20} + n_{02})]/n$$

In this case, under the null hypothesis, $\mathrm{var}(\hat{p}_1 - \hat{p}_2) = h/16n$ (e.g., this can be calculated from the variance of S in Table 1 using $f_2 = f_1 = 1$). Therefore, we can construct the statistic

$$T_{DS} = \frac{(n_{21} - n_{12} + n_{10} - n_{01} + 2n_{20} - 2n_{02})/4n}{\sqrt{(n_{21} + n_{12} + n_{10} + n_{01} + 4n_{20} + 4n_{02})/16n^2}} \quad (1)$$

The subscripts on T denote that we do not assume Hardy-Weinberg equilibrium and that sibs are used to contruct the parent allele frequency.

To calculate the power of statistic 1, we reformat T_{DS} to

$$T_{DS} = \frac{(n_{21} - n_{12} + n_{10} - n_{01} + 2n_{20} - 2n_{02})/\sqrt{16n}}{\sqrt{(n_{21} + n_{12} + n_{10} + n_{01} + 4n_{20} + 4n_{02})/16n}} \quad (2)$$

We assume the denominator converges to its expected value (by the Law of Large Numbers), and thus, we need only calculate this expectation along with the mean and variance of the numerator under the alternative hypothesis. We denote the expectation of the square of the denominator as $E(\sigma_0^2)$ and the mean and variance of the numerator as $\sqrt{n}\,\nu$ and σ_a^2. From Table 1,

$$E(\sigma_0^2) = \tfrac{1}{32}[m_{(21)} + m_{(10)} + m_{11}(3f_2 + 2f_1 + 3)/(f_2 + 2f_1 + 1)]$$
$$\nu = \tfrac{1}{2}(m_{(21)}\pi_{21} + m_{(10)}\pi_{10} + m_{11}\pi_{11})$$

and $\quad \sigma_a^2 = E(\sigma_0^2) - \nu^2$

Then, the power is given by

$$\Phi\left(\frac{\sqrt{E(\sigma_0^2)}\,Z_\alpha + \nu\sqrt{n}}{\sigma_a}\right) \quad (3)$$

r Affected and *s* Unaffected Sibs

By using the same logic described above for one affected and one unaffected sib, we can construct a sibship-based disequilibrium test statistic for the general case of r affected and s unaffected sibs. We classify the various outcomes into six groups based on the possible matings that could have produced them: (I) All sibs are *AA*; (II) all sibs are *aa*; (III) all sibs are *Aa*; (IV) all sibs are either *AA* or *Aa*; (V) all sibs are either *Aa* or *aa*; and (VI) the genotypes *AA* and *aa* (and possibly *Aa*) appear among the sibs. These categories are meant to be mutually exclusive, so that, for example, group IV excludes the case of all sibs being *AA*. In theory, it may be possible to obtain additional information by subdividing groups IV and V by the number of *Aa* individuals; however, by the above grouping scheme, we are able to obtain analytic formulas for power and sample size, as described below. We can characterize each possible outcome as a vector with the six elements (j_2, j_1, j_0, k_2, k_1, k_0) where j_i is the number of affected sibs with i A alleles, and k_i is the number of unaffected sibs with i A alleles. Note that $j_2 + j_1 + j_0 = r$, and $k_2 + k_1 + k_0 = s$, and we define $t = r + s$. The possible outcomes, by group, are listed in Table 2, along with their probabilities under the alternative hypothesis. Under the null hypothesis, the corresponding probabilities can be obtained by using the population mating-type frequencies instead of the conditional (on having r affected children) mating-type frequencies and substituting in $f_2 = f_1 = 1$.

To derive the T_{DS} statistic, we first sum up the probabilities across all possible outcomes within each group under the null hypothesis. We obtain the following totals:

$$\begin{aligned}
&\text{I: } m_{22} + m_{(21)}(\tfrac{1}{2})^t + m_{11}(\tfrac{1}{4})^t \\
&\text{II: } m_{11}(\tfrac{1}{4})^t + m_{(10)}(\tfrac{1}{2})^t + m_{00} \\
&\text{III: } m_{(21)}(\tfrac{1}{2})^t + m_{(20)} + m_{11}(\tfrac{1}{2})^t + m_{(10)}(\tfrac{1}{2})^t \\
&\text{IV: } m_{(21)}[1 - (\tfrac{1}{2})^{t-1}] + m_{11}[(\tfrac{3}{4})^t - (\tfrac{1}{2})^t - (\tfrac{1}{4})^t] \\
&\text{V: } m_{11}[(\tfrac{3}{4})^t - (\tfrac{1}{2})^t - (\tfrac{1}{4})^t] + m_{(10)}[1 - (\tfrac{1}{2})^{t-1}] \\
&\text{VI: } m_{11}[1 + (\tfrac{1}{2})^t - 2(\tfrac{3}{4})^t]
\end{aligned} \quad (4)$$

We denote by n_{I} the number of observations that fall into group I and similarly for the other groups. By equating the sample frequencies of each group, that is, n_{I}/n, n_{II}/n, etc., with their respective probabilities, and solving the six equations, we can get unbiased maximum likelihood estimates of the $m_{(ij)}$'s under the null

hypothesis, which are denoted by $\hat{m}_{(ij)}$. Recalling that $p_2 = m_{22} + \frac{3}{4}\, m_{(21)} + \frac{1}{2}\, m_{(20)} + \frac{1}{2}\, m_{11} + \frac{1}{4}\, m_{(10)}$, and using the maximum likelihood estimates of the m_{ij} based on the simplified classification scheme given above, we can estimate p_2 by

$$\hat{p}_2 = [n_{\mathrm{I}} + \tfrac{1}{2}n_{\mathrm{III}} + \tfrac{3}{4}n_{\mathrm{IV}} + \tfrac{1}{4}n_{\mathrm{V}} + \tfrac{1}{2}n_{\mathrm{VI}}]/n \qquad (5)$$

This formula can be easily derived by taking the linear combination in equation 5 applied to the formulas in equation 4. Then, to obtain $\hat{p}_2$, we can simply assign a score $S(p_2)$ of 1, 3/4, 1/2, 1/4, or 0 depending on the group membership of the outcome; these scores are given in Table 2.

This derivation is similar to the approach we took for the simple case of one affected and one unaffected sib. However, in this general case, collapsing all possible sibship outcomes (ignoring affection status) into the six groups defined above, although unbiased, does not use all of the information available. Specifically, within group IV there is additional information about parental mating type based on the frequency of sibships defined by the number of *AA* and *Aa* sibs. For example, in sibships of size 3, this would correspond to the relative frequency of sibships with two *AA* and one *Aa* sib versus those with one *AA* and two *Aa* sibs, which provides some information on the relative frequency of the parental mating type $AA \times Aa$ versus $Aa \times Aa$. A similar comment applies to group V (for matings $Aa \times aa$ and $Aa \times Aa$). For the four other sibship groups, further subdivision is either not possible (groups I, II, and III) or provides no additional information about mating type (group VI, in which the parental mating type is automatically $Aa \times Aa$). By not further subgrouping groups IV and V, we are able to derive formulas for the estimate of p_2 and $\mathrm{Var}(\hat{p}_1 - \hat{p}_2)$ that are simple and robust and can therefore also perform all power calculations and sample estimates analytically. Presumably, there is also some loss of efficiency in doing so, although much of the information about parental-mating type frequencies is contained in the relative frequency of groups I to VI. A maximum likelihood solution to estimate the parental mating type frequencies allowing for subgrouping of groups IV and V may be possible by numerical means; however, no simple formulas for parameter estimation, power calculations, or sample size estimates are possible in this case. Furthermore, we demonstrate below in numerical examples that our simple statistic is more efficient than one based on comparing the frequency of allele A in affected versus unaffected sibs, for sibships of size 3 or greater.

Scores can also be assigned for the estimate of p_1. To do so, we simply take $(j_2 + 1/2j_1) / r$, independent of which group contains the outcome. These scores $[S(p_1)]$ are also given in Table 2. To estimate $p_1 - p_2$, we can then assign scores based on the difference in the scores $S(p_1)$ and $S(p_2)$; these scores, $S(p_1 - p_2)$, are also given in Table 2. As can be seen there, the score is $(j_2 - j_1) / 4r$ in sibships with only *AA* and *Aa* sibs, $(j_1 - j_0) / 4r$ in sibships with only *Aa* and *aa* sibs, and $(j_2 - j_0) / 2r$ in sibships with *AA* and *aa* sibs.

In some sense, some of the scoring of sibships, as given in Table 2, may seem counterintuitive. Consider a sibship of two affected and one unaffected. For groups I to III, the uniform scoring of 0 is straightforward, as all sibs (affected and unaffected) have the same genotype. Now, suppose the two affected sibs have genotypes *AA* and *Aa*. This sibship will be scored the same (0) if the unaffected sib has genotype *AA* or *Aa*. This is because, in either case, the sibship belongs to group IV, and the unaffected child does not change the possible mating types of the parents. On the other hand, if the unaffected sib is genotype *aa*, the sibship now belongs to group VI and gets a score of +1/2 because the parental mating type is $Aa \times Aa$. As another example, suppose the two affected sibs have genotypes *AA* and *aa*. Then the sibship will be scored 0 whatever the genotype of the unaffected sib (i.e., *AA*, *Aa*, or *aa*) because the sibship automatically belongs to group VI. A scoring routine based on the frequency of the *A* allele in the affected sibs versus the unaffected sib would score this family differently based on whether the unaffected sib was *AA*, *Aa*, or *aa* (e.g., −1/2 if the unaffected sib is *AA*, 0 if *Aa*, and +1/2 if *aa*). However, it is clear that in the creation of a TDT-type statistic (comparing offspring with parents' allele frequency), in this case the unaffected child provides no additional information.

Under the null hypothesis, $E(\hat{p}_1 - \hat{p}_2) = 0$. To calculate $\mathrm{Var}(\hat{p}_1 - \hat{p}_2)$, we note that $\hat{p}_1 - \hat{p}_2 = [\Sigma S_i(p_1 = p_2)] / n$ is the average of n independent, identically distributed scores, so that $\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{1}{n}\mathrm{Var}[S(p_1 - p_2)]$, where the subscript i has been suppressed. Because $E[S(p_1 - p_2)] = 0$, we simply calculate $\mathrm{Var}[S(p_1 - p_2)] = E\{[S(p_1 - p_2)]^2\}$. After some lengthy algebra, we obtain

$$\begin{aligned}\mathrm{Var}[S(p_1 - p_2)] = {} & (m_{(21)} + m_{(10)})[\tfrac{1}{r} - (\tfrac{1}{2})^{t-1}]/16 \\ & + m_{11}[\tfrac{1}{r} - \tfrac{1}{3}(\tfrac{3}{4})^t - (\tfrac{1}{2})^t - (\tfrac{1}{4})^t]/8\end{aligned}$$

By using logic similar to that used in the derivation of $\hat{p}_2$ and using the maximum likelihood estimates of the m_{ij}, we can estimate this variance by

$$\begin{aligned}\hat{V}[S(p_1 - p_2)] = \hat{\sigma}_0^2 = {} & \tfrac{1}{16n}(n_{\mathrm{IV}} + n_{\mathrm{V}})\frac{[\frac{1}{r} - (\frac{1}{2})^{t-1}]}{[1 - (\frac{1}{2})^{t-1}]} \\ & \{\tfrac{1}{r}[1 - (\tfrac{1}{2})^t - (\tfrac{3}{4})^t + (\tfrac{1}{4})^t] \\ & + \frac{n_{\mathrm{VI}}}{8n}\frac{+ (\frac{3}{8})^{t-1} - \frac{1}{3}(\frac{3}{4})^t - (\frac{1}{2})^t - (\frac{1}{4})^t\}}{[1 - (\frac{1}{2})^{t-1}][1 + (\frac{1}{2})^t - 2(\frac{3}{4})^t]}\end{aligned} \qquad (6)$$

Thus, the T_{DS} statistic, for the general case of r affected and s unaffected sibs, is given by

$$T_{DS} = \frac{\Sigma S_i(p_1 - p_2)}{\hat{\sigma}_0 \sqrt{n}}$$

in which the scores are given in Table 2 and $\hat{\sigma}_0$ by the square root of formula 6. Under the null hypothesis, the T_{DS} statistic is approximately normally distributed with mean 0 and variance 1.

To calculate the power of this test, we need to determine $\nu = E[S(p_1 - p_2)]$, $E(\hat{\sigma}_0^2)$, and $\mathrm{Var}[S(p_1 - p_2)]$ under the alternative hypothesis. Then, using the formulas in Table 2, and after some tedious algebra, we obtain the following results:

$$\begin{aligned}\nu = E[S(p_1 - p_2)] &= \tfrac{1}{4} m^{(r)}_{(21)}[c_{21} - c_{20} - (\tfrac{1}{2})^s(c^r_{21} - c^r_{20})] \\ &+ \tfrac{1}{4} m^{(r)}_{(10)}[c_{01} - c_{00} - (\tfrac{1}{2})^s(c^r_{01} - c^r_{00})] \\ &+ \tfrac{1}{4} m^{(r)}_{11}[2(c_{12} - c_{10}) - (\tfrac{3}{4})^s(c_{12} + c_{11})^r \\ &+ (\tfrac{3}{4})^s(c_{11} + c_{10})^r - (\tfrac{1}{4})^s c^r_{12} + (\tfrac{1}{4})^s c^r_{10}] \end{aligned} \quad (7)$$

$$\begin{aligned}E(\hat{\sigma}_0^2) = &\frac{[\tfrac{1}{r} - (\tfrac{1}{2})^{t-1}]}{16[1 - (\tfrac{1}{2})^{t-1}]} \{m^{(r)}_{(21)}[1 - (\tfrac{1}{2})^s(c^r_{21} + c^r_{20})] \\ &+ m^{(r)}_{(10)}[1 - (\tfrac{1}{2})^s(c^r_{01} + c^r_{00})] + m^{(r)}_{11}[(\tfrac{3}{4})^s(c_{12} + c_{11})^r \\ &+ (\tfrac{3}{4})^s(c_{11} + c_{10})^r - (\tfrac{1}{4})^s c^r_{12} - (\tfrac{1}{4})^s c^r_{10} - 2(\tfrac{1}{2})^s c^r_{11}\} \\ &+ m^{(r)}_{11}\{\tfrac{1}{r}[1 - (\tfrac{1}{2})^t - (\tfrac{3}{4})^t + (\tfrac{1}{4})^t] + (\tfrac{3}{8})^{t-1} \\ &- \tfrac{1}{3}(\tfrac{3}{4})^t - (\tfrac{1}{2})^t - (\tfrac{1}{4})^t\} \\ &\times \frac{[1 - (\tfrac{3}{4})^s(c_{11} + c_{10})^r - (\tfrac{3}{4})^s(c_{12} + c_{11})^r + (\tfrac{1}{2})^s c^r_{11}]}{8[1 - (\tfrac{1}{2})^{t-1}][1 + (\tfrac{1}{2})^t - 2(\tfrac{3}{4})^t]} \end{aligned} \quad (8)$$

and

$$\begin{aligned}\sigma_a^2 = \mathrm{Var}[S(p_1 - p_2)] &= \tfrac{1}{16r} m^{(r)}_{(21)}[r - 4(r-1)c_{21}c_{20} - r(\tfrac{1}{2})^s(c^r_{21} + c^r_{20})] \\ &+ \tfrac{1}{16r} m^{(r)}_{(10)}[r - 4(r-1)c_{01}c_{00} - r(\tfrac{1}{2})^s(c^r_{01} + c^r_{00})] \\ &+ \tfrac{1}{16r} m^{(r)}_{11}\{(\tfrac{3}{4})^s[r(c_{21} + c_{11})^r - 4(r-1)\, c_{12}c_{11} \\ &(c_{12} + c_{11})^{r-2} + r(c_{11} + c_{10})^r - 4(r-1)c_{11}c_{10} \\ &(c_{11} + c_{10})^{r-2} - 4(2f_1 + r)c^2_{10}(c_{11} + c_{10})^{r-2} - 4(rf_2 \\ &+ 2f_1)c_{12}c_{10}(c_{12} + c_{11})^{r-2}] - r(\tfrac{1}{2})^{s-1}c^r_{11} - r(\tfrac{1}{4})^s \\ &(c^r_{12} + c^r_{10}) + 4(r-1)(c_{12} - c_{10})^2 + 4(c_{12} + c_{10})\} \end{aligned} \quad (9)$$

The power can then be calculated using formula 3, substituting formulas 7, 8, and 9 for ν, $E(\hat{\sigma}^2{}_0)$, and $\sigma^2{}_a$, respectively.

Numerical Results—Individual Genotyping vs. Pooling

Using the power formulas described above, we can calculate required sample sizes to detect linkage disequilibrium. The logic is the same as described in Risch and Teng (1998) for sample pooling; again, we use a significance level of 5×10^{-8} and 80% power. The required sample sizes are given in Table 3. Using the T_{DS} test for sibships without parents with individual genotyping can produce a significant advantage over the pooled statistic (T_{HS}), depending on the family structure (compare with Table 4 in Risch and Teng 1998). For families with one affected sib, the sample sizes are roughly comparable, with low allele frequencies slightly favoring the T_{DS} statistic but high allele frequencies slightly favoring the T_{HS} statistic. As the number of affected sibs increases, however, the advantage of the T_{DS} statistic increases. For two affected sibs, on average (across genetic models), 25% fewer families are required; for three affected sibs, 35% fewer are needed, whereas for four affected sibs, nearly half as many families are necessary using individual genotyping and the T_{DS} statistic. As in the case for one affected child, the ratios are highest at low allele frequencies. The only exception is the high frequency dominant situation, in which the T_{HS} test may retain a slight advantage. We note also that these conclusions are reasonably independent of the number of unaffected sibs used.

From Table 2 and Table 3 of Risch and Teng (1998), we can also contrast the number of families required under individual genotyping when both parents are available versus using two unaffected sibs when they are not (giving an identical number of family members). Using two unaffected sibs requires ~50% more families, roughly independent of the number of affected sibs and genetic model. This number can be substantially higher, however, for a very common dominant allele.

Combining Families of Different Structure

As described previously in Risch and Teng (1998), it is typical that an investigator will have families of different structure, including different numbers of affected sibs and possibly unaffected sibs. As in the case for pooled samples, we suggest taking a weighted sum of allele frequency differences $(\hat{p}_1 - \hat{p}_2)$ for the various family structures, in which the weight is according to the number affected in the family and the number of families of that structure. Thus, for families with r affected sibs, we multiply $(\hat{p}_1 - \hat{p}_2)$ by rn_{ri} before summing, in which n_{ri} is the number of families with r affected of structure i, and then divide the total by $N = \Sigma r n_{ri}$. To obtain the denominator, we simply sum $r^2 n^2{}_{ri}\mathrm{Var}(\hat{p}_1 - \hat{p}_2)$, in which the variance of $\hat{p}_1 - \hat{p}_2$ for a given family structure under the null hypothesis is given in the formulas above, divide by N^2, and then take the square root.

DISCUSSION

We have considered test statistics that can be created when individual genotyping is performed in nuclear families containing affected and unaffected sibs without parents. We have shown previously that to calcu-

Table 2. Probabilities of Different Outcomes for r Affected and s Unaffected Sibs and Scores for the T_{DS} Statistic

Group	Outcome	Score			Probability
		$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_1 - \hat{p}_2$	
I	$(r,0,0,s,0,0)$	+1	+1	0	$m^{(r)}_{22} + 2^{-s} m^{(r)}_{(21)} c^r_{21} + 4^{-s} m^{(r)}_{11} c^r_{12}$
II	$(0,0,r,0,0,s)$	0	0	0	$4^{-s} m^{(r)}_{11} c^r_{10} + 2^{-s} m^{(r)}_{(10)} c^r_{00} + m^{(r)}_{00}$
III	$(0,r,0,0,s,0)$	1/2	1/2	0	$2^{-s} m^{(r)}_{(21)} c^r_{(20)} + m^{(r)}_{(20)} + 2^{-s} m^{(r)}_{11} c^r_{11} + 2^{-s} m^{(r)}_{(10)} c^r_{01}$
IV	$(j_2,j_1,0,k_2,k_1,0)$	$(j_2 + \frac{1}{2}j_1)/r$	3/4	$(j_2 - j_1)/4r$	$\binom{r}{j_2}\binom{s}{k_2} 2^{-s} [m^{(r)}_{(21)} c^{j_2}_{21} c^{j_1}_{20} + 2^{-k_2} m^{(r)}_{11} c^{j_2}_{12} c^{j_1}_{11}]$
V	$(0,j_1,j_0,0,k_1,k_0)$	$j_1/2r$	1/4	$(j_1 - j_0)/4r$	$\binom{r}{j_0}\binom{s}{k_0} 2^{-s} [m^{(r)}_{(10)} c^{j_1}_{01} c^{j_0}_{00} + 2^{-k_0} m^{(r)}_{11} c^{j_1}_{11} c^{j_0}_{10}]$
VI	$(j_2,j_1,j_0,k_2,k_1,k_0)$	$(j_2 + \frac{1}{2}j_1)/r$	1/2	$(j_2 - j_0)/2r$	$\binom{r}{j_2,j_1,j_0}\binom{s}{k_2,k_1,k_0} 2^{k_1 - 2t} m^{(r)}_{11} c^{j_2}_{12} c^{j_1}_{11} c^{j_0}_{10}$

late the TDT for families with parents, individual genotyping is only required for the parents, to obtain a direct estimate of h. The child allele frequencies can still be obtained by DNA pooling, which could lead to a significant reduction in genotyping effort, especially for larger sibships.

Because it is possible to estimate the variance in the allele frequency difference between the affected and unaffected sibs without the Hardy–Weinberg assumption in families without parents, estimators that are immune to population stratification artifacts can be constructed. The statistic we have described, the T_{DS} test, is analogous to the TDT because it contrasts allele frequencies between parents and affected offspring, as in the TDT, and uses a variance estimate independent of the Hardy–Weinberg assumption. In this case, the parent allele frequencies are estimated from the total offspring sibship, including both the affected and unaffected offspring.

When the tested sibship contains only a single affected, the power of the T_{DS} statistic is quite close to the pooled T_{HS} statistic, so the primary advantage of the T_{DS} statistic is its robustness. However, as the number of affected in the sibship increases, the power of the T_{DS} test increases relative to the T_{HS} test, providing an additional advantage. We also note that the T_{DS} statistic is easily calculated using the scores given in Table 2 and its variance by formula 6 above.

When families with multiple affected sibs are used, neither the pooled statistic T_{HS} described in Risch and Teng (1998) nor the T_{DS} test described here compare favorably in terms of power with tests based on using unrelated controls instead of unaffected sibs. Thus, strategies involving both family-based as well as unrelated controls may be preferable.

It may be tempting to use the same group of affecteds in a two-stage process—that is, first comparing them to unrelated controls to increase power to identify candidate loci and then comparing these same affected individuals to family-based controls (parents or unaffected sibs) for robustness. However, in this approach, the two tests will be positively correlated under

Table 3. Number of Sibships Without Parents Required to Detect LD Using Individual Genotyping

	$r = 1$		$r = 2$		$r = 3$	$r = 4$
	$s = 1$	$s = 2$	$s = 1$	$s = 2$	$s = 2$	$s = 2$
Dominant						
$P = 0.05$	642	430	196	137	72	52
$P = 0.20$	455	312	250	173	147	155
$P = 0.70$	5,659	4,254	6,890	4,510	6,414	9,546
Recessive						
$P = 0.05$	79,709	61,544	18,376	13,202	3,358	965
$P = 0.20$	2,097	1,528	582	410	152	76
$P = 0.70$	443	297	254	178	152	160
Multiplic						
$P = 0.05$	2,538	1,715	946	642	314	180
$P = 0.20$	870	604	421	286	177	125
$P = 0.70$	938	649	595	404	311	269
Additive						
$P = 0.05$	1,490	1,002	520	356	177	110
$P = 0.20$	688	475	359	244	173	142
$P = 0.70$	1,393	980	959	647	505	42

(r) Number of affected sibs; (s) Number of unaffected sibs.

the null hypothesis, and so the threshold for significance for the second test needs to be constructed taking this correlation into account.

Other tests of linkage disequilibrium based on sibships without parents and individual genotyping have been proposed. Penrose first suggested the use of unaffected sibs as controls in association studies to protect against artifactual results owing to population stratification (Clarke et al. 1956). The method of C.A.B. Smith (Smith 1961), as also described in Clarke et al. (1956), is essentially based on a comparison of genotypes in affected children with their unaffected sibs. The proposal of Curtis (1997) is similar in this regard. Since our paper was submitted, two additional papers (Boehnke and Langefeld 1998; Spielman and Ewens 1998) have appeared describing sibship-based statistics. These tests are also based on allele (or genotype) frequency difference between affected and unaffected sibs, similar to the original Smith test. For sibships with one affected and one unaffected sib, all of these tests (including ours) are equivalent. However, for larger sibships the tests diverge.

We have chosen to focus on a TDT-like statistic, estimating parental allele and heterozygosity frequency, as this approach yields a more efficient test for sibships with multiple affecteds. However, a critical assumption underlying this advantage is that unaffected sibs reflect a random distribution of parental alleles. This will certainly be nearly true whenever the "locus-specific" penetrance for the tested locus is low and the unaffected sibs are selected randomly. However, this statistic would not necessarily be more efficient than a statistic based on comparison of allele frequencies in affected versus unaffected sibs, when the locus-specific penetrance is high or when the unaffected sibs are chosen from the opposite extreme of a continuous distribution from which the affecteds are chosen (e.g., lean sibs of obese sib pairs) (Eaves and Meyer 1994; Risch and Zhang 1995). In this case, the allele frequency in unaffected sibs is also expected to deviate from the parental allele frequency. The relative efficiency of the two types of tests, in this case, will depend on the degree to which the allele frequency in affected sibs is expected to deviate from that in unaffected sibs relative to that in the parents, and on the number of unaffected sibs.

At first glance, it may seem mysterious as to why the T_{DS} statistic has increased efficiency over other sibship-based statistics that compare affected and unaffected sibs. These latter statistics are based solely on comparisons of genotypes *within* sibships. However, there is additional information available in the sample that our statistic incorporates, namely, the relative frequency of the different sibship genotype constellations (ignoring affection status in the sibship). For example, for sibships of size 3, we also use the frequency of sibships with three *AA* sibs, two *AA* and one *Aa* sib, two *AA* and one *aa* sib, and so on (for all possible genotype combinations). This distribution of sibship genotypes provides information regarding the frequency of the six possible parent mating types. Because the mating-type frequencies are estimated without assuming random mating, the estimation procedure is robust to any deviation from random mating including population stratification. For example, in the extreme stratification case in which half the sibships have three *AA* sibs and the other half three *aa* sibs, our procedure estimates half the parent mating types to be $AA \times AA$ and the other half to be $aa \times aa$, a complete deviation from random mating and Hardy–Weinberg genotype frequencies.

The analogy of the T_{DS} statistic to the TDT statistic may also seem mysterious if the latter is viewed as a statistic derivable only from intact nuclear families. As we showed in Risch and Teng (1998), however, the TDT is calculated from three components: (1) the frequency of allele *A* in the offspring (p_1); (2) the frequency of allele *A* in the parents (p_2); and (3) the frequency of heterozygous parents (h). It is entirely unnecessary to have intact families to derive these statistics. For example, p_1 and p_2 can be obtained, in theory, by DNA pooling, whereby all children are pooled together and all parents are pooled together. Even if parent DNA samples are separated from their offspring's, a TDT can still be calculated. All that is required is knowing that a sample is from a child or a parent. Thus, it is obviously unnecessary to know which child genotypes are associated with which parent genotypes to construct a TDT.

In the T_{DS} statistic, we are effectively recreating a TDT-type statistic. In this case, however, parental allele frequencies and heterozygosity are not estimated directly from the parents, who are missing, but from the offspring. That this can be done without bias derives from the fact that there are at least as many different possible sibship genotype constellations as parent mating types.

ACKNOWLEDGMENTS

This work was supported, in part, by grants from the National Human Genome Research Institute (HG00348) and the Nancy Pritzker Foundation. We are grateful to Dr. Michael Boehnke for many helpful comments and suggestions on this manuscript and to Drs. David Curtis and Cedric Clarke for pointing out the Clarke et al. reference.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Boehnke, M. and C.D. Langefeld. 1998. Genetic association mapping based on discordant sib pairs: The discordant-alleles test. *Am. J. Hum. Genet.* **62:** 950–961.

Clarke, C.A., J. Wyn Edwards, D.R.W. Haddock, A.W. Howel-Evans, R.B. McConnell, and P.M. Sheppard. 1956. ABO blood groups and secretor character in duodenal ulcer. *Br. Med. J.* **2:** 725–731.

Curtis, D. 1997. Use of siblings as controls in case-control association studies. *Ann. Hum. Genet.* **61:** 319–333.

Eaves, L. and J. Meyer. 1994. Locating human quantitative trait loci: Guidelines for the selection of sibling pairs for genotyping. *Behav. Genet.* **24:** 443–455.

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

Risch, N. and H. Zhang. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268:** 1584–1589.

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

Smith, C.A.B. 1961. Statistical methods and theory. In *Recent advances in human genetics* (ed. L.S. Penrose), pp. 148–149. J.&A. Churchill, Ltd., London, UK.

Spielman, R.S., R.E. McGinnis, and W.J. Ewens. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52:** 506–516.

Spielman, R.S. and W.J. Ewens. 1998. A sibship based test for linkage in the presence of association: The sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62:** 450–458.

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

Received November 9, 1998; accepted in revised form January 20, 1999.