# Use of Parents, Sibs, and Unrelated Controls for Detection of Associations between Genetic Markers and Disease

Daniel J. Schaid[1,2] and Charles Rowland[1]

Departments of [1]Health Sciences Research and [2]Medical Genetics, Mayo Clinic/Mayo Foundation, Rochester, MN

## Summary

**Detecting the association between genetic markers and complex diseases can be a critical first step toward identification of the genetic basis of disease. Misleading associations can be avoided by choosing as controls the parents of diseased cases, but the availability of parents often limits this design to early-onset disease. Alternatively, sib controls offer a valid design. A general multivariate score statistic is presented, to detect the association between a multiallelic genetic marker locus and affection status; this general approach is applicable to designs that use parents as controls, sibs as controls, or even unrelated controls whose genotypes do not fit Hardy-Weinberg proportions or that pool any combination of these different designs. The benefit of this multivariate score statistic is that it will tend to be the most powerful method when multiple marker alleles are associated with affection status. To plan these types of studies, we present methods to compute sample size and power, allowing for varying sibship sizes, ascertainment criteria, and genetic models of risk. The results indicate that sib controls have less power than parental controls and that the power of sib controls can be increased by increasing either the number of affected sibs per sibship or the number of unaffected control sibs. The sample-size results indicate that the use of sib controls to test for associations, by use of either a single-marker locus or a genomewide screen, will be feasible for markers that have a dominant effect and for common alleles having a recessive effect. The results presented will be useful for investigators planning studies using sibs as controls.**

## Introduction

Detecting the association between genetic markers and complex diseases can be a critical first step toward identification of the genetic basis of disease. Associations can result from genetic linkage and/or linkage disequilibrium. Genetic linkage between disease and marker loci causes the marker alleles to be associated with disease within families (i.e., cosegregation). Linkage disequilibrium implies that the occurrence of particular haplotypes, composed of specific alleles at the disease and marker loci, produces disease-marker associations between families (i.e., at the population level). In family studies, allowing for linkage disequilibrium in linkage analyses can increase the power to detect causative genes (Clerget-Darpoux et al. 1986). Traditionally, case-control studies have been used to study associations at the population level. However, case-control studies with unrelated controls can be prone to biases due to population stratification. Sampling the parents of diseased cases as controls has proved to be a powerful design (Spielman et al. 1993; Risch and Merikangas 1996) that detects true genetic associations—that is, those that are caused by both linkage and linkage disequilibrium of the disease and marker loci. Note that parents are not used as controls in the traditional manner but, rather, to assess whether their marker alleles are transmitted to their diseased child according to Mendelian probabilities, with any distortion suggesting genetic associations that are due to both linkage and linkage disequilibrium. But the utility of this design is limited to the availability of the genetic markers for the parents. Unless DNA can be extracted from archival specimens, this means that the use of parents as controls will be limited to early-onset diseases. On occasion, sibs can be used to infer the genotypes of missing parents, but, because inference of the missing parents' genotypes depends on their genotypes and on their offspring's genotypes, this inference of missing data can lead to biased results (Curtis and Sham 1995; Curtis 1997). To adequately account for the missing parental genotypes, a model can be used (Schaid and Li 1997), but this requires assumptions about the population, such as random mating of parents and a homogeneous population. These model-based approaches

Received March 9, 1998; accepted for publication August 21, 1998; electronically published October 23, 1998.

Address for correspondence and reprints: Dr. Daniel J. Schaid, Harwick 7, Mayo Clinic, 200 First Street S.W., Rochester, MN 55905. E-mail: schaid@mayo.edu

0002-9297/98/6305-0028$02.00

require further evaluations, because of the required assumptions.

As an alternative to the use of parental controls, sib controls can be used to avoid the biases that occur in a stratified population, with the obvious advantage of the availability of sibs, but not of parents, for diseases that occur at older ages. Although for >40 years this type of design has been considered as a valid way to avoid population stratification (Manuila 1958), only recently have scientists explored the utility of the use of sib controls (Curtis 1997; Ewens and Spielman 1997; Langefeld et al. 1997; Monks et al. 1997; Spielman and Ewens 1998). These scientists have tackled the analytic issues by using seemingly different approaches. For example, Langefeld et al. (1997) and Boehnke and Langefeld (1998) proposed analytic methods when using only one discordant (affected/unaffected) pair of sibs from each sibship and proposed the use of simulations to compute a probability value, because of the complexity of the dependencies among the alleles within a sibship. Curtis (1997) proposed a similar method of analysis, allowing for only one affected subject per sibship and choosing as control the unaffected sib whose genotype is maximally different from that of the case. Spielman and Ewens (1998) proposed a statistic that they called the "sib-TDT" (S-TDT), which is based on the use of affected and unaffected sibs within a sibship. This method is based on comparison of the number of occurrences of a particular marker allele among the diseased cases with its expected value when there is no association, with the variance taken into account. The mean and variance are computed within each sibship, on the basis of the hypergeometric distribution. Monks et al. (1997) performed simulations to evaluate the power of the S-TDT method for two strategies when there are multiple marker alleles: (1) test the association of each allele individually and correct for the multiple tests by the Bonferroni correction and (2) test all alleles simultaneously. Their simulation results indicated that testing all alleles simultaneously can be more powerful than testing each individually.

Although these novel designs and analytic methods will prove useful, a general statistical method is needed that allows for multiple marker alleles and combination of different types of controls, such as cases and parental controls with cases and sib controls and, perhaps, even with cases and unrelated controls. For example, at times it may be advantageous to sample unrelated controls, in addition to sib or parental controls, to assess the potential impact of population stratification, and to pool when this is deemed appropriate. We present a general score statistic that allows pooling of different types of controls and that is similar to some of the recently proposed analytic methods. An important advantage of our proposed methods of analysis is that they can be performed by use of software that is available in most statistical-analysis software packages. To plan these types of studies, we present methods to compute sample size and power when sib controls, parental controls, and unrelated controls are sampled. These computations are then used to contrast the power of using different types of controls and different ascertainment schemes. Further extensions are discussed, such as allowance for censored data—which is particularly needed for complex diseases, to control for the confounding effects of age—and inclusion of covariates.

## Methods

### *Score Statistic for Sib Controls*

Spielman and Ewens (1998) have presented a method to compare the frequency of a particular marker allele in affected versus unaffected sibs. Their method is implicitly based on construction of a 2 × 3 table for each sibship, in which the first row is for affected sibs, the second row is for unaffected sibs, and the three columns are for the three genotypes when there are two marker alleles. The method that we propose for multiple marker alleles is an extension of this approach.

To define notation and the general setup for this methodology, we first consider the comparison of genotype frequencies in affected versus unaffected sibs and then consider the comparison of allele frequencies. Let $N_s$ denote the total number of sibships, such that each sibship has at least one affected and at least one unaffected sib. Let $G$ denote the number of observed genotypes among all subjects in the total sample, and let $K$ denote the number of alleles. For each sibship, create a 2 × $G$ table in which the first row is the count of genotypes for the affected sibs, denoted as vector $\mathbf{x}_{gi}$ for the $i$th sibship, and in which the second row is the count of genotypes for the unaffected sibs, denoted as vector $\mathbf{y}_{gi}$ ; the subscript $g$ is used to indicate reference to genotypes (we shall later use subscript $a$ to refer to alleles). An example 2 × 6 table for three alleles and six genotypes is given in table 1. The marginal row totals for the $i$th sibship are the numbers of affected ($N_i^d$) and unaffected ($N_i^c$) sibs, with a total of $N_i$ sibs in the $i$th sibship. The marginal column totals of the genotype counts is the vector $\mathbf{t}_{gi} = \mathbf{x}_{gi} + \mathbf{y}_{gi}$. Under the null hypothesis of no

**Table 1**

**Genotype Frequencies**

| Affection Status | Frequency of Genotype | | | | | | |
|---|---|---|---|---|---|---|---|
| | *AA* | *AB* | *AC* | *BB* | *BC* | *CC* | Total |
| Affected: $\mathbf{x}'_{gi}$ = | $x_{gi1}$ | $x_{gi2}$ | $x_{gi3}$ | $x_{gi4}$ | $x_{gi5}$ | $x_{gi6}$ | $N_i^d$ |
| Unaffected: $\mathbf{y}'_{gi}$ = | $y_{gi1}$ | $y_{gi2}$ | $y_{gi3}$ | $y_{gi4}$ | $y_{gi5}$ | $y_{gi6}$ | $N_i^c$ |
| Total $\mathbf{t}'_{gi}$ = | $t_{gi1}$ | $t_{gi2}$ | $t_{gi3}$ | $t_{gi4}$ | $t_{gi5}$ | $t_{gi6}$ | $N_i$ |

association of the markers with affection status, and conditional on the marginal totals, the 2 × $G$ table has a hypergeometric distribution, which allows computation of the mean and covariance matrix for the vector of genotype counts for the affected sibs, $\mathbf{x}_{gi}$. For the $i$th sibship, the vector of expected counts of genotypes among the affected sibs is $\mathbf{e}_{gi} = \mathbf{t}_{gi}N_i^d/N_i$, and the covariance matrix for the $\mathbf{x}_{gi}$ vector is $\mathbf{V}_{gi} = [\mathrm{diag}(\mathbf{p}_{gi}) - \mathbf{p}_{gi}\mathbf{p}'_{gi}]N_i^dN_i^c/(N_i - 1)$, where $\mathbf{p}_{gi} = \mathbf{t}_{gi}/N_i$, and $\mathrm{diag}(\mathbf{p}_{gi})$ is a diagonal matrix with the elements of $\mathbf{p}_{gi}$ running down the diagonal and with the off-diagonal elements being zero. The dimension of $\mathbf{V}_{gi}$ is $G \times G$. To compare the total observed genotype counts for the affected sibs across all families, $\mathbf{x}_{g\cdot} = \Sigma\mathbf{x}_{gi}$, with its expected value when there is no association, $\mathbf{e}_{g\cdot} = \Sigma\mathbf{e}_{gi}$, the following statistic can be used:

$$S_g = (\mathbf{x}_{g\cdot} - \mathbf{e}_{g\cdot})'\mathbf{V}_{g\cdot}^{-}(\mathbf{x}_{g\cdot} - \mathbf{e}_{g\cdot}) \ , \tag{1}$$

where the subscript dot indicates summation over all sibships. Because genotype frequencies sum to one, $\mathbf{V}_{g\cdot}$ is not of full rank, so a generalized inverse, $\mathbf{V}_{g\cdot}^{-}$, is used. Alternatively, an arbitrary element can be eliminated from the vector of differences, $(\mathbf{x}_{g\cdot} - \mathbf{e}_{g\cdot})$, as can the corresponding row and column of $\mathbf{V}_{g\cdot}$, and then these reduced vectors and inverted matrix can be substituted into expression (1), to compute $S_g$. For large sample sizes, $S_g$ has an approximate $\chi^2$ distribution with $(G - 1)$ df. Note that the statistic $S_g$ is the stratified statistic that Mantel and Haenszel (1959) proposed for case-control studies, in order to adjust for a confounding factor by stratification. In our application, the confounding factor is the varying genotype frequencies from sibship to sibship. This approach does assume that the strata (i.e., sibships) are independent of each other. When they are not, such as when sibships originate from the same pedigree, this approach can still be motivated by considering the conditional moments of the stratified tables as originating from the score statistic for a partial likelihood (Cox 1975). The problem with this approach is that the number of genotypes is often large, resulting in both a large number of df and low power.

An alternative approach is to compare the observed counts of alleles, not genotypes, with their expected values, as has been proposed by Spielman and Ewens (1998). Because alleles are correlated among sibs, this correlation must be taken into account. To do so, note that allele counts are linear combinations of genotype counts; it is easy to compute a covariance matrix for linear combinations of random variables. To illustrate the counting of alleles among affected sibs, note that the count of the $k$th allele can be obtained by letting $k/k$ homozygotes contribute a count of 2, $k/j$ heterozygotes contribute a count of 1, and all other genotypes contribute a count of 0. This can be represented in vector notation, as $\mathbf{b}'_k\mathbf{x}_{gi}$, where the $j$th element of the vector $\mathbf{b}_k$ has the value 2, 1, or 0, depending on whether the corresponding $j$th genotype has 2, 1, or 0 alleles of type $k$. After these $\mathbf{b}_k$ vectors are bound into a matrix, $\mathbf{B}' = (\mathbf{b}_1,\mathbf{b}_2, \ldots ,\mathbf{b}_K)$, where $\mathbf{B}$ has dimension $K \times G$, the vector of observed allele counts among the affected sibs in the $i$th sibship can be represented as $\mathbf{x}_{ai} = \mathbf{B}\mathbf{x}_{gi}$. Under the null hypothesis, the vector of expected counts of alleles is $\mathbf{e}_{ai} = \mathbf{B}\mathbf{e}_{gi}$, and the covariance matrix of $\mathbf{x}_{ai}$ is $\mathbf{V}_{ai} = \mathbf{B}\mathbf{V}_{gi}\mathbf{B}'$. A valid statistical comparison of the total observed allele counts among the diseased sibs across all families, versus the total expected allele counts, can be made with the multivariate statistic

$$S_{\mathrm{sib}} = (\mathbf{x}_{a\cdot} - \mathbf{e}_{a\cdot})'\mathbf{V}_{a\cdot}^{-}(\mathbf{x}_{a\cdot} - \mathbf{e}_{a\cdot}) \ , \tag{2}$$

where, once again, the dot subscript indicates summation over all families. The statistic $S_{\mathrm{sib}}$ has an asymptotic $\chi^2$ distribution with, at most, $(K - 1)$ df. It is worthwhile to note that the $S_{\mathrm{sib}}$ statistic is the score statistic for a conditional logistic regression model with stratification on sibships. That is, the genotypes of subjects are coded into a genotype covariate vector that has length $(K - 1)$ and elements of 0, 1, or 2, which are simply counts of the number of alleles of each type that a person possesses; one of the $K$ alleles is arbitrarily chosen as a baseline allele for computation of relative risks and hence is ignored in this covariate vector. After the strata for sibships have been set up and the covariate vectors have been created, standard software for conditional logistic regression can be used to calculate the score statistics for association. This type of coding forces additive effects of alleles onto the log odds ratio. The advantage of using the score statistic is that it can be rapidly computed, which is appealing when one is either calculating simulated $P$ values for small samples or evaluating many marker loci. Alternatively, the likelihood ratio statistic could be used. When sib controls are used, it may be particularly important to control for known environmental risk factors, by including them within the regression model. This also allows evaluation of interaction between environmental risk factors and genetic markers, on the relative risk of disease.

The $S_{\mathrm{sib}}$ statistic in expression (2) is a generalization of the $Z$ statistic, called "S-TDT," which has been proposed by Spielman and Ewens (1998) for the case of $K = 2$ alleles. Spielman and Ewens have suggested that, when there are more than two alleles, a $Z$ statistic be computed for each of the $K$ alleles, with each $Z$ having an approximate standard normal distribution, and that the maximum of these $K$ statistics, $Z_{\max}$, be used to perform a global test of association. The Bonferroni correction is needed when $Z_{\max}$ is used. In fact, the terms used in expression (2) can be used to compute the $Z$ values. For the $k$th allele, the $Z$ is $Z_k = (\mathbf{x}_{a\cdot} -$

$\mathbf{e}_{a\bullet})_k/\sqrt{\mathbf{V}_{a\bullet kk}}$, where $(\mathbf{x}_{a\bullet} - \mathbf{e}_{a\bullet})_k$ is the $k$th element of the vector of differences and $\mathbf{V}_{a\bullet kk}$ is the $k$th diagonal element from the covariance matrix. Hence, by computation of the multivariate statistic in expression (2), the elements are also available for computation of the statistic $Z_{\max}$. The statistic $Z_{\max}$ is likely to be more powerful than the multivariate statistic $S_{\text{sib}}$, when a single marker allele is associated with disease. However, if multiple alleles are associated with disease, as is expected when linkage disequilibrium is not complete, the approach using $Z_{\max}$ can have less power than the multivariate statistic (Schaid 1996).

The statistic $S_{\text{sib}}$ in expression (2) is also similar to the approach given by Boehnke and Langefeld (1998), who proposed that, when there is a single discordant pair per sibship, the statistic $AC_2 = \Sigma_{k=1}^{K}(n_k^d - n_k^c)^2/(n_k^d + n_k^c)$, be used, where $n_k^d$ is the count, across all sibships, of the $k$th marker allele among cases and where $n_k^c$ is that among controls, with the rule that only those marker alleles that differ within a discordant sib pair contribute to these counts. It can be shown that the $k$th element of the vector of differences, $(\mathbf{x}_{a\bullet} - \mathbf{e}_{a\bullet})_k$, is equal to $(n_k^d - n_k^c)/2$. If $\mathbf{V}_{a\bullet}$ in expression (2) were replaced by a matrix having diagonal elements $(n_k^d + n_k^c)/4$ and off-diagonal elements 0, then the statistics $S_{\text{sib}}$ and $AC_2$ would be equivalent. Boehnke and Langefeld (1998) have proposed that $P$ values for the $AC_2$ statistic be computed by simulations, because the $AC_2$ statistic does not account for covariances of alleles among sibs. Our proposed statistic in expression (2) considers the same contrasts as are evaluated by the $AC_2$ statistic yet appropriately accounts for covariances, thereby eliminating the need for simulations.

### *Score Statistic for Parental Controls*

Schaid (1996) showed that, when trios (a case and parental controls) are used, a score statistic can be used to test for disease-marker associations. The parents are used as controls, but only to assess the status of transmission of each of their alleles to their affected child. The form of the score statistic is similar to that for the sib controls. For a given trio, denote the two alleles of the mother as "$m_1$" and "$m_2$," the two alleles of the father as "$f_1$" and "$f_2$," and the genotype of the case (affected child) as "$g_i^d$." The genotypes of the potential offspring that these parents can produce are in the set $G_i = \{m_1f_1, m_1f_2, m_2f_1, m_2f_2\}$; one of these four genotypes corresponds to the case's genotype, $g_i^d$, and the other three can be considered pseudo–sib controls (Self et al. 1991). Let the $K \times 1$ vector $\mathbf{x}_{ai}$ denote the observed count of each of the $K$ alleles in the case's genotype, $g_i^d$. The vector of expected allele counts under the null hypothesis is $\bar{\mathbf{x}}_{ai}$, where $\bar{\mathbf{x}}_{ai}$ is the average over the four $\mathbf{x}$-coded genotypes in the set $G_i$. The covariance matrix of the $\mathbf{x}_{ai}$ vector, $\mathbf{V}_{ai}$, is the covariance matrix of the four $\mathbf{x}$-coded vectors in the set $G_i$. With this notation and computation of $\mathbf{x}_{ai}$, $\bar{\mathbf{x}}_{ai}$, and $\mathbf{V}_{ai}$ for each trio stratum, the score statistic for parental controls is $S_{\text{par}} = (\mathbf{x}_{a\bullet} - \bar{\mathbf{x}}_{a\bullet})'\ V_{a\bullet}^{-}(\mathbf{x}_{a\bullet} - \bar{\mathbf{x}}_{a\bullet})$, where the dot notation indicates summation over all trio strata. The $S_{\text{par}}$ score statistic has an asymptotic $\chi^2$ distribution with, at most, $(K-1)$ df. Note that the score statistic for parental controls, $S_{\text{par}}$, is the same as a score statistic for conditional logistic regression. In this case, each trio is a stratum, and within each stratum is the case and the three pseudo–sib controls; if there are multiple affected sibs, then each affected sib would have its own stratum, requiring replication of parental genotypes, to create the pseudo–sib controls. Then, the genotypes of the case and its pseudo–sib controls can be coded in the manner outlined for sib controls. In contrast to the situation with sib controls, the main effects of environmental risk factors cannot be evaluated when parental controls are used; only interaction between the environmental risk factors and the genetic markers can be assessed, such that the marker-genotype relative risks vary according to the environmental risk factors (Self et al. 1991; Schaid 1995).

### *Score Statistic for Unrelated Controls*

When cases and controls are not genetically related, a common method to compare allele frequencies between cases and controls is to create a 2 × $K$ table of allele counts for cases (row 1) and controls (row 2) and then to compute Pearson's $\chi^2$ statistic. However, the validity of this method requires that alleles within genotypes be statistically independent under the null hypothesis. Departures from independence (i.e., departure from Hardy-Weinberg genotype proportions) can lead to inflated type I error rates (Schaid and Jacobsen 1998). However, the method outlined above for computation of the covariance matrix of allele counts for sib controls can also be used to compute a covariance matrix for unrelated controls that is robust to departures from Hardy-Weinberg proportions (authors' unpublished data). In other words, with a single stratum for all unrelated cases and controls, create the 2 × $G$ table as in table 1 and then compute the observed and expected vectors of allele counts for diseased cases, the matrix of covariances, and the $\chi^2$ statistic, in the manner outlined for sib controls. If there is a strong confounder (associated with both disease and marker alleles), such as ethnic background, then one can stratify on the confounder and then apply the statistical methods outlined above ("Score Statistic for Sib Controls").

### *Pooling across Different Control Groups*

An appeal of the proposed $\chi^2$ statistics is the ability to pool data while adjusting for the types of controls used. The strategy is to first compute expectations and covariances for each type of control used and then to create a summary statistic across all strata. Because expectations and covariances are computed in a similar manner for sib controls and unrelated controls, we only need to expand on the similarities of the sib-controls and parental-controls statistics, to illustrate appropriate methods to pool the data. Consider parental controls. For an affected child in the *i*th trio, create a $G \times 1$ vector, denoted $\mathbf{x}_{gi}$, that indicates which genotype the case possesses (i.e., elements of $\mathbf{x}_{gi}$ are 0 or 1); this $\mathbf{x}_{gi}$ vector is similar to the first row of table 1. In fact, multiple affected sibs can be included ($N_i^d > 1$), such that $\mathbf{x}_{gi}$ is a vector of genotype counts among all $N_i^d$ sibs. Under the null hypothesis, all of the genotypes in the set $G_i$ have equal probabilities, $\frac{1}{4}$ (i.e., Mendelian probabilities). Indistinguishable genotypes in the set $G_i$ can be collapsed, and their probabilities can be summed, to compute the probabilities of the different genotypes as arranged for the $\mathbf{x}_{gi}$ vector (see column heading of table 1). Let $\mathbf{p}_{gi}$ denote the $G \times 1$ vector of these conditional (on parental genotypes) probabilities. Then the expected value of $\mathbf{x}_{gi}$ under the null hypothesis is $\mu_{gi} = N_i^d \mathbf{p}_{gi}$, and the multinomial covariance matrix of $\mathbf{x}_{gi}$ is $\mathbf{V}_{gi} = N_i^d[\text{diag}(\mathbf{p}_{gi}) - \mathbf{p}_{gi}\mathbf{p}'_{gi}]$. Now, when the matrix $\mathbf{B}$ is used to transform the data from genotype counts to allele counts, the vector of observed allele counts among affected cases is $\mathbf{x}_{ai} = \mathbf{B}\mathbf{x}_{gi}$, the vector of expected allele counts is $\bar{\mathbf{x}}_{ai} = \mathbf{B}\mu_{gi}$, and the covariance matrix of $\mathbf{x}_{ai}$ is $\mathbf{V}_{ai} = \mathbf{B}\mathbf{V}_{gi}\mathbf{B}'$. Note that this covariance matrix is identical to simple computation of the covariance matrix of the four $\mathbf{x}$-coded vectors within the set $G_i$; but this latter formulation illustrates that we are taking linear combinations of genotype covariances, as we had done when considering sib controls. The main distinction between parental controls and sib controls is the vector $\mathbf{p}_{gi}$ that defines the joint distribution of genotypes within a stratum; for parental controls, $\mathbf{p}_{gi}$ is determined by Mendelian probabilities, whereas, for sib controls, $\mathbf{p}_{gi}$ is estimated on the basis of the distribution of genotypes within a sibship.

Now, suppose that there are $J_{\text{par}}$ strata for parental controls, $J_{\text{sib}}$ strata for sib controls, and $J_{\text{unr}}$ strata for unrelated controls, giving a total of $J = J_{\text{par}} + J_{\text{sib}} + J_{\text{unr}}$ strata. For each stratum, compute the vector of allele counts for the cases, $\mathbf{x}_{ai}$, as well as its expected value, $\mathbf{e}_{ai}$, and the covariance matrix for these allele counts, $\mathbf{V}_{ai}$, where expectations and covariances depend on the type of controls, as outlined above. Then, with use of the dot notation to indicate summation over all $J$ strata, the pooled $\chi^2$ statistic takes the same form as that used for sib controls and parental controls, $S_{\text{pool}} = (\mathbf{x}_{a\bullet} - \mathbf{e}_{a\bullet})' V_{a\bullet}^{-}(\mathbf{x}_{a\bullet} - \mathbf{e}_{a\bullet})$. Alternatively, one could use the elements of $(\mathbf{x}_{a\bullet} - \mathbf{e}_{a\bullet})$ and the diagonal elements from $V_{a\bullet}$ to compute univariate standard normal $Z$ statistics and then use the maximum of these, $Z_{\max}$, as a global test of association, again using the Bonferroni correction.

### Sample Size and Power

For a given alternative hypothesis and a large sample size, each of the score statistics for parental, sib, and unrelated controls is distributed as a noncentral $\chi^2$ distribution with a noncentrality parameter that depends on the marker-allele frequencies and the genotype relative risks. Details for parental controls are presented by Schaid (1996). For practical planning of a study, it is simplest to consider sample size and power calculations when there are only two marker alleles: *A*, the high-risk allele, and *B*, the low-risk allele. Let $p$ and $q = 1 - p$ denote the frequencies of alleles *A* and *B*, respectively. Under the assumption of Hardy-Weinberg equilibrium, the genotype probabilities are $P(AA) = p^2$, $P(AB) = 2pq$, and $P(BB) = q^2$.

We consider power as it depends on the marker-genotype relative risks, because these parameters summarize the associations between the marker genotypes and affection status. However, even though we refer to dominant, recessive, or other patterns of marker-genotype relative risks, it is important to recognize that these relative risks are only for the marker genotypes and will not necessarily correspond to the relative risks for the disease-causing genotypes—unless the marker is a causative gene. The magnitudes of the marker-genotype relative risks depend on the frequencies of the disease-causing allele(s), the strength of linkage disequilibrium between the disease and marker loci, and the relative risks for the disease-causing genotypes (Schaid 1996). Incomplete linkage disequilibrium will cause the marker-genotype relative risks to be less than the disease-causing genotype relative risks. To write the penetrances as functions of relative risks, let the genotype *BB* be the baseline genotype (i.e., that having relative risk $r_{BB} = 1$) and let $r_{AA}$ and $r_{AB}$ denote the genotype relative risks for the genotypes *AA* and *AB*, respectively. If $f_g$ is the penetrance for genotype $g$, then the penetrances can be written as $f_{AA} = r_{AA} f_{BB}$ and $f_{AB} = r_{AB} f_{BB}$.

*Sib controls.*—When sib controls are used, sample size and power will depend on the number of affected ($N^d$) and unaffected ($N^c$) sibs per sibship ($N = N^d + N^c$). The power calculations that have been given by Spielman and Ewens (1998), are useful for evaluation of the relative power of parental controls versus sib controls, but only for particular genotype configurations. When planning a study, we must take into account that the genotype configurations in a nuclear family will depend on the genetic parameters, such as marker-allele frequencies and marker-genotype relative risks. So, instead of the conditioning on the marginal totals of each stratified table, as is done in the computation of the score statistics, these marginal totals are considered as random variables, because they are not known prior to sampling. This variation is taken into account to compute the power for sib controls, as outlined by the methods in Appendix A.

*Parental controls.*—Methods to compute sample size and power for the TDT method have been given, by Risch and Merikangas (1996), for multiplicative relative risks: $r_{AA} = r_{AB}^2$. They also considered sampling either one or two affected sibs per sibship and have pointed out that it is more powerful to sample two, instead of one, affected sibs per sibship, because this increases the chance that at least one of the parents will be heterozygous for the marker alleles and, hence, informative for determination of transmission status. Schaid (1998) has derived methods to determine sample size and power of the TDT method for general genotype relative risks and has shown that, when marker alleles are common—and when multiplicative genotype relative risks are assumed when, in fact, the true pattern of relative risks is additive (i.e., $r_{AA} = 2r_{AB} - 1$), dominant (i.e., $r_{AA} = r_{AB}$) or recessive ($r_{AB} = 1$)—the required sample size for the TDT statistic can be grossly underestimated. However, the calculations by Schaid (1998) were for $N^d = 1$ affected child per sibship. In Appendix B we extend these ideas to allow for an arbitrary number of affected sibs ($N^d \geqslant 1$) when computing power.

*Unrelated controls.*—To compute power for unrelated controls, we need to specify the marker-allele frequencies, the marker-genotype relative risks, and the lifetime risk of disease, denoted as "$P_d$;" in this case, the baseline penetrance, $f_{BB}$, can be calculated, on the basis of the genetic relative risks, allele frequencies, and $P_d$, as $f_{BB} = P_d/(p^2r_{AA} + 2pqr_{AB} + q^2)$. Also, both the total number of cases and the ratio of controls to cases, denoted as "$R_c$," are specified. With these parameters, power can be calculated, as detailed in Appendix C.

*Power comparisons.*—The power of the parental-, sib-, and unrelated-controls designs were compared, to evaluate how well sib controls perform relative to these other designs. To do so, power was computed for sampling of 100 nuclear families, with a population lifetime risk of disease, $P_d = .05$, and by a two-sided statistical test having a type I error rate of 5% (i.e., $Z_\alpha = 1.96$). Two marker alleles were assumed, such that the high-risk allele was either rare ($p = .01$) or common ($p = .2$). The genetic models considered were those for dominant ($r_{AA} = r_{AB}$) and recessive ($r_{AB} = 1$) effects; rare recessive effects are not presented because all designs had weak power to detect these. We also considered the influence of the size of the sibship, $N$, on power, allowing $N$ to be 2–4. For both parental- and sib-control designs, each sibship must have at least one affected sib ($N^d \geqslant 1$), in order to be informative. In contrast, the sib-control design is not informative when all sibs are affected ($N^d = N$), whereas this type of sibship is highly informative for the parental-control design. But, to fairly compare the power of these two types of designs, we allowed $N^d$ to vary from a minimum ascertainment criterion (either one or two affected sibs) to a maximum value of $N - 1$, so that each sibship has at least one unaffected sib. In practice, including sibships with all sibs affected can dramatically increase the power for the parental controls.

The power for unrelated controls requires that we fix the number of affected and unaffected subjects. To contrast the power for unrelated controls versus that for either parental or sib controls, we fixed the number of affected cases for the unrelated-controls design to be the same as the expected number of affected sibs in the total sample of 100 nuclear families, that was used for the parental- and sib-control designs. This expected number of affected sibs, which depends on the sibship size, the marker-allele frequency and genotype relative risks, the ascertainment criteria, and the population risk of disease, was computed by consideration of the probabilities of all possible tables (see expressions [A1] and [A2] in Appendix A) and the number of affected sibs per table. The number of unrelated controls was set equal to the number of affected cases. We chose this 1:1 ratio because it is commonly chosen in practice and because it simplified the comparison of power. Technically, it would be more accurate to allow the ratio of controls per case, $R_c$, to vary according to the sib-control design. But this made some power comparisons potentially misleading. For example, a design with two affected sibs and one unaffected sib would lead to a choice of $R_c = .5$ unrelated controls per case, a design rarely used. So our power comparisons reflect the power for sampling strategies that are straightforward to employ, and they do not necessarily reflect the relative efficiency, in terms of their total sample size requirements, of different designs.

## Results

The power for parental, sib, or unrelated controls, for a rare dominant effect, is presented in figure 1, for two ascertainment schemes: (1) sampling at least one affected in sibships of size 2 (fig. 1*A*), size 3 (fig. 1*B*), or size 4 (fig. 1*C*) and (2) sampling at least two affecteds in sibships of size 3 (fig. 1*D*) or size 4 (fig. 1*E*). These power curves illustrate several key points. First, when the ascertainment criterion is at least one affected (fig. 1*A*–*C*), unrelated controls and parental controls have similar power, which is greater than that for sib controls. Second, there is a gain in power for sib controls as the sibship size increases from two sibs (fig. 1*A*) to three sibs (fig. 1*B*) to four sibs (fig. 1*C*). This is essentially due to the increase in the number of controls per sibship,

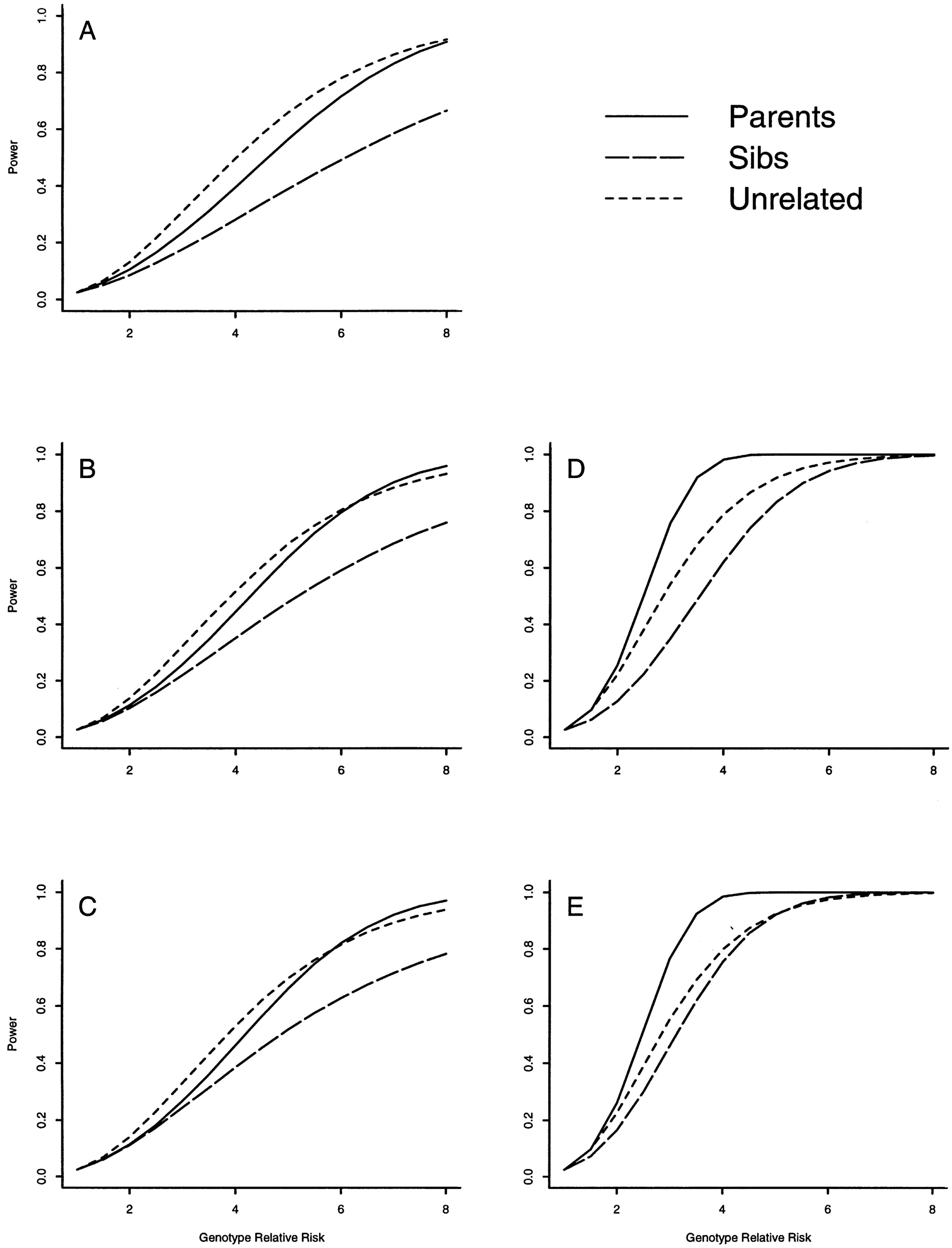

**Figure 1** Power to detect associations with a rare ($p$ = .01) marker allele having dominant effects, when parental controls from 100 nuclear families, sib controls from 100 nuclear families, and 100 unrelated controls are used. Ascertainment criteria are as follows: at least one affected sib in sibships of size 2 (*A*), size 3 (*B*), or size 4 (*C*) or at least two affected sibs in sibships of size 3 (*D*) or size 4 (*E*).

with little increase in the number of affecteds per sibship. For example, for a relative risk of 4, the frequency of more than one affected sib per sibship increased from 3%, for a sibship size of 2, to 8%, for a sibship of size 4. Third, power is greater when the ascertainment criterion is at least two affecteds per sibship (fig. 1*D* and *E*), compared with that when there is at least one affected per sibship (fig. 1*B*–*C*); again, sib controls had the least amount of power, compared with parental controls and unrelated controls. Interestingly, when at least two affected sibs are required, the gain in power for parental controls can be substantial, giving greater power than that given by unrelated controls (fig. 1*D* and *E*). The power increased for all three types of controls when the frequency of the marker allele increased from .01 (fig. 1) to .20 (fig. 2), although, for this dominant effect, the contrast in power for the three different types of controls was qualitatively similar in figures 1 and 2.

The power for a common marker having a recessive effect is illustrated in figure 3. When at least one affected sib is required (fig. 3*A*–*C*), the power of parental controls and unrelated controls is similar, and both of these designs have greater power than does the sib-controls design. Increasing the number of sibs increases the power of sib controls (fig. 3*A*–*C*). Notice, however, that, when the ascertainment criterion increases from at least one affected (fig. 3*A*–*C*) to at least two affected (fig. 3*D* and *E*), there is a substantial gain in power for the parental-controls design, so that this design has greater power than either that for unrelated controls or that for sib controls. This is similar to the power comparisons for the rare dominant effect, portrayed in figure 1. That is, both figures 1 and 3 highlight the fact that the parental-controls design can, by increasing the minimum number of affected sibs per sibship, achieve a substantial gain in power to detect a rare high-risk genotype.

The sample size and power for a study using sib controls is determined by the difference between the expected count of *A* alleles among cases and that predicted by the marginal genotype distribution among sibs ($\mu_{\mathrm{alt}}$), as well as by the SD of this measure, under the specified alternative hypothesis ($\sigma_{\mathrm{alt}}$); see expression (A5) in Appendix A. To facilitate sample-size and power computations when planning a study using sib controls, we computed the values of $\mu_{\mathrm{alt}}$ and $\sigma_{\mathrm{alt}}$, allowing sibship size to vary. The negative binomial distribution was used to predict the distribution of sibship size, with a mean of 2 and a variance of 4 (slightly less than the mean of 2.6 and variance of 5.1 that were reported, by Brass [1958], for the United States population in 1950). For these computations, at least one unaffected sib was required, as well as either at least one affected sib (i.e., a truncated distribution of sibship size 2–8), or at least two affected sibs (i.e., a truncated distribution with sibship size 3–8). The population lifetime risk of disease was assumed to be 5%. The values of $\mu_{\mathrm{alt}}$ and $\sigma_{\mathrm{alt}}$ are presented in table 2, for a variety of genetic parameters related to a dominant effect (i.e., $r_{AA} = r_{AB}$), and in table 3, for parameters related to a recessive effect (i.e., $r_{AB} = 1$). Also presented in tables 2 and 3 are the number of sibships required to detect an association either when a single genetic marker is evaluated or when a genomewide screen is performed, both designs having 80% power. When evaluating a single marker locus, we used a two-sided test having a type I error of 5% (i.e., $Z_\alpha = 1.96$). When performing a genomewide screen, we used the Bonferroni correction to control the false-positive rate when evaluating multiple marker loci (Risch and Merikangas 1996; Spielman and Ewens 1998). We assumed that there were 500,000 independent tests (five variants in each of 100,000 genes; Lander 1996), which required $Z_\alpha = 5.33$. Although this method will be adequate for independent tests of association, it will tend to be conservative when alleles from different marker loci are in linkage disequilibrium, such as when these loci are physically close. To illustrate how to compute sample size, consider a single-locus test (requiring $Z_\alpha = 1.96$), with 80% power ($Z_\beta = .84$) to detect a high-risk allele having population frequency $p = .1$, and a dominant effect with $r_{AA} = r_{AB} = 2$. From the data in table 2, it can be inferred that, if at least one affected sib is required, then $\mu_{\mathrm{alt}} = .04769$ and $\sigma_{alt} = .27254$. On the basis of expression (A5), $N_s = (Z_\alpha + Z_\beta)^2\sigma^2_{\mathrm{alt}}/\mu^2_{\mathrm{alt}}$, resulting in 257 sibships required, such that each sibship has at least one affected and at least one unaffected sib.

Tables 2 and 3 illustrate several key points regarding sample-size requirements. The sample sizes required for a genomewide screen are four to five times the sample size required for a single-locus test. For a marker with dominant genotype relative risks, sample sizes are likely to be feasible for marker relative risks as low as 2—and perhaps even lower, if the high-risk allele is common. In contrast, for recessive effects, sample sizes are likely to be feasible when the high-risk allele is not rare and the genotype relative risk is large. For both dominant and recessive effects, the number of sibships required can be reduced by ascertainment of at least two affected sibs per sibship, with the greatest reduction occurring when the high-risk allele is rare. However, requiring a larger number of affecteds per sibship increases both (*a*) the size of each sibship if the presence of unaffected control sibs is to be guaranteed and (*b*) the amount of effort required to identify these perhaps unusual sibships.

## Discussion

A general multivariate score statistic, $S_{\mathrm{sib}}$, has been presented in order to assess the association between a multiallelic genetic-marker locus and affection status,

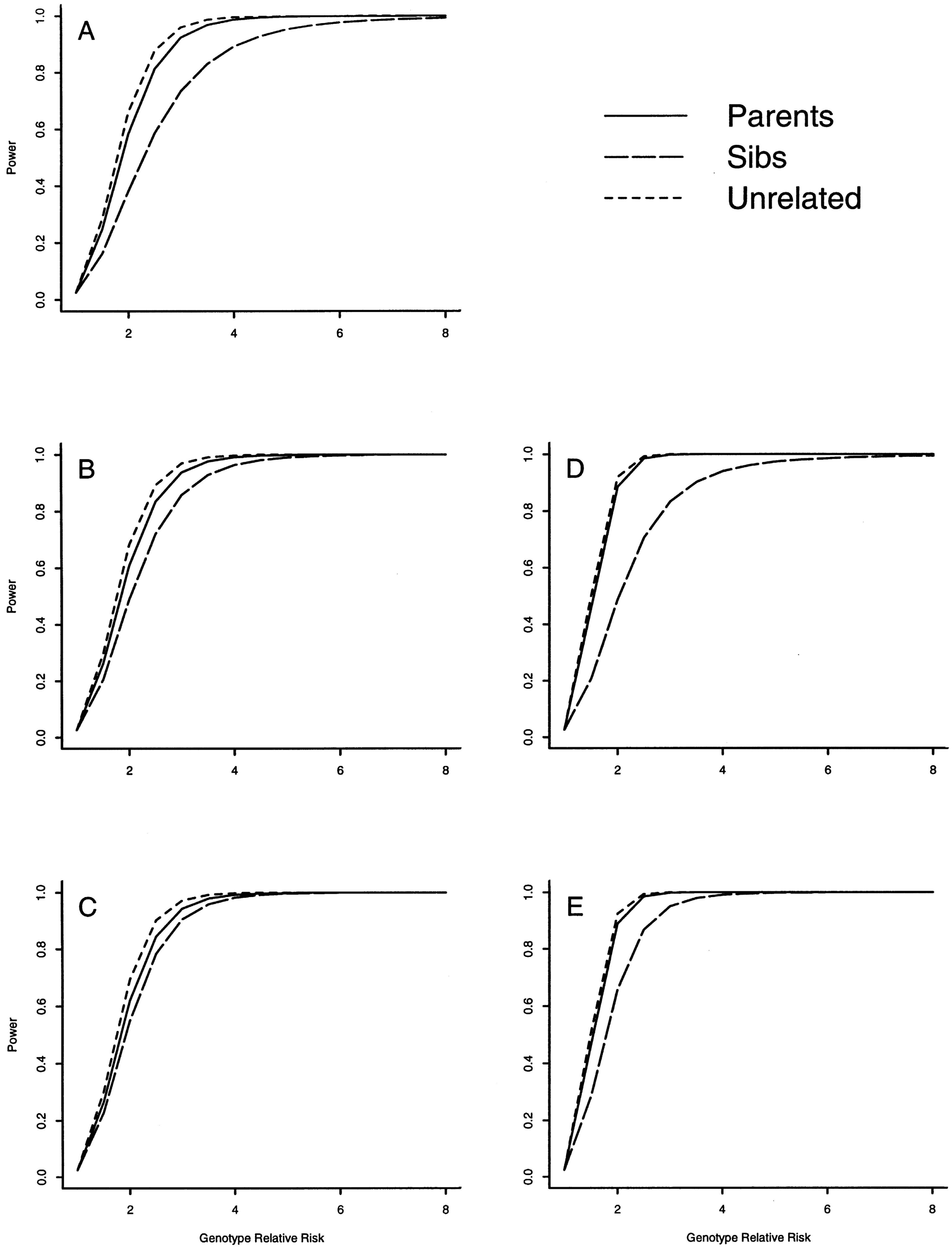

**Figure 2** Power to detect associations with a common ($p$ = .20) marker allele having dominant effects, when parental controls from 100 nuclear families, sib controls from 100 nuclear families, and 100 unrelated controls are used. Ascertainment criteria/panel differentiation are as in fig. 1.

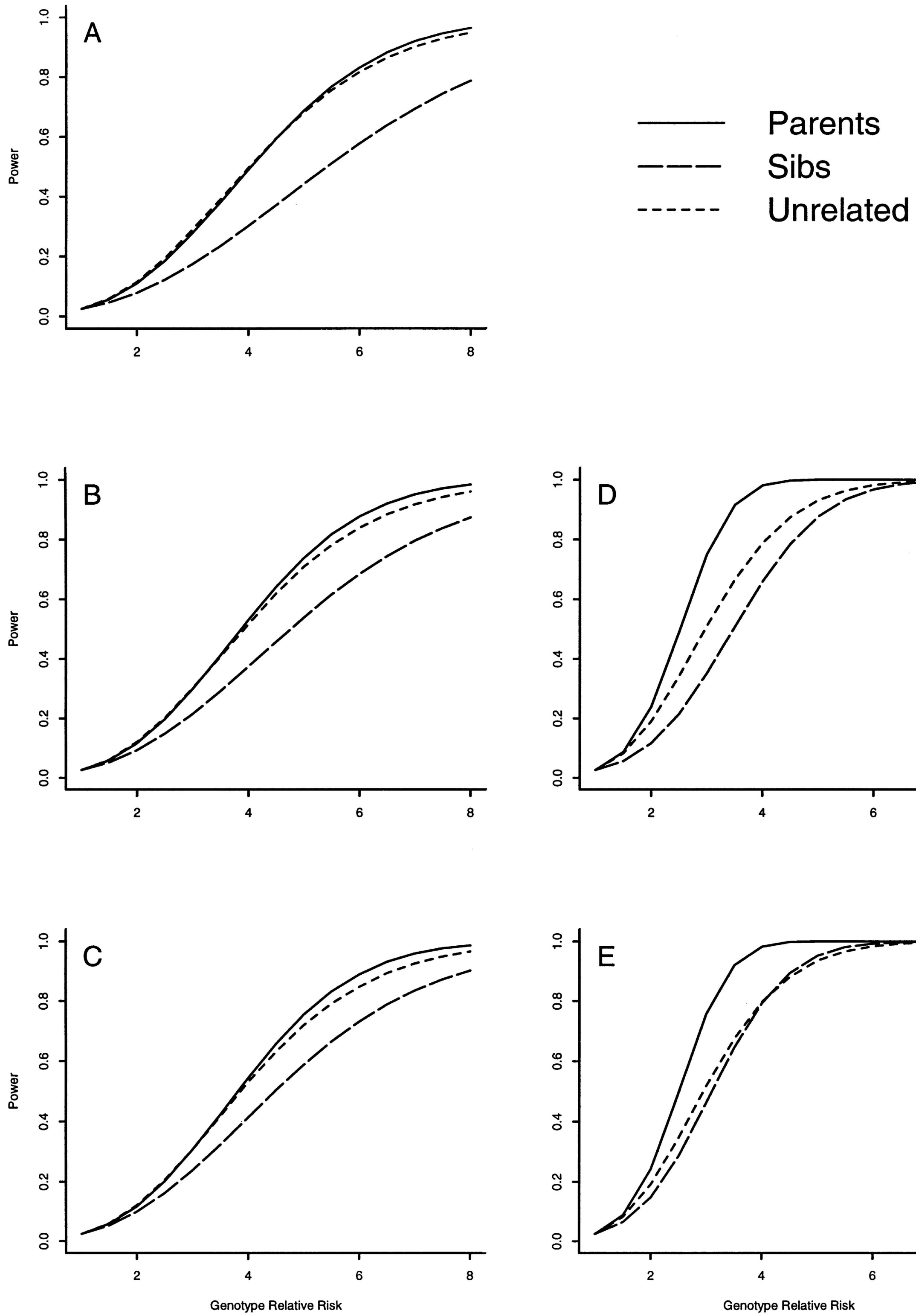

**Figure 3** Power to detect associations with a common ($p$ = .20) marker allele having recessive effects, when parental controls from 100 nuclear families, sib controls from 100 nuclear families, and 100 unrelated controls are used. Ascertainment criteria/panel differentiation are as in fig. 1.

## Table 2

**Number of Sibships Required to Have 80% Power, under the Assumption of 5% False-Positive Rate, to Detect Association with a Genetic Marker Having Autosomal Dominant Effects on Disease Risk**

| | | | No. of Sibships | |
|---|---|---|---|---|
| $p$, $r_{AA} = r_{AB}$, and min $N^d$ [a] | $\mu_{alt}$ [b] | $\sigma_{alt}$ [b] | Single-Locus Test | Genomewide Screen |
| .01: | | | | |
| 2: | | | | |
| 1 | .00671 | .09787 | 1,669 | 8,098 |
| 2 | .01395 | .14482 | 846 | 4,104 |
| 4: | | | | |
| 1 | .01947 | .12534 | 326 | 1,579 |
| 2 | .05964 | .23248 | 120 | 579 |
| 8: | | | | |
| 1 | .04308 | .16844 | 121 | 583 |
| 2 | .17635 | .34886 | 31 | 150 |
| .1: | | | | |
| 2: | | | | |
| 1 | .04769 | .27254 | 257 | 1,244 |
| 2 | .08371 | .37019 | 154 | 745 |
| 4: | | | | |
| 1 | .11019 | .29990 | 59 | 283 |
| 2 | .20649 | .43029 | 35 | 166 |
| 8: | | | | |
| 1 | .17845 | .31973 | 26 | 123 |
| 2 | .31443 | .45238 | 17 | 79 |
| .2: | | | | |
| 2: | | | | |
| 1 | .06598 | .34249 | 212 | 1,027 |
| 2 | .10297 | .44233 | 145 | 703 |
| 4: | | | | |
| 1 | .13102 | .35098 | 57 | 274 |
| 2 | .19885 | .46689 | 44 | 210 |
| 8: | | | | |
| 1 | .18341 | .35227 | 29 | 141 |
| 2 | .25930 | .47359 | 27 | 128 |

[a] $p$ = frequency of high-risk allele; min $N^d$ = minimum no. of affected sibs per sibship. $r_{AA} = r_{AB}$ denotes that these two values are assumed to be equal for a dominant model.

[b] Parameters used in expression (A5), to determine sample size.

when sib controls are used. This method generalizes both the approach suggested by Spielman and Ewens (1998) for two marker alleles and the approach suggested by Boehnke and Langefeld (1998). One advantage of our proposed statistic is that, for large sample sizes, it has an approximately $\chi^2$ distribution and so avoids the longer time required by the methods given by Boehnke and Langefeld (1998) for computation of simulated $P$ values. This time savings can be significant when one is evaluating multiple marker loci, such as in a genomewide screen. Limited simulations (not shown) have indicated that, when sample size is large, the $AC_2$ statistic and the $S_{sib}$ statistic have similar power and that the $\chi^2$ distribution is adequate for computation of $P$ values for the $S_{sib}$ statistic. However, for small sample sizes, the asymptotic $\chi^2$ distribution may not be adequate, requiring simulated $P$ values; the permutation method that has been given by Spielman and Ewens (1998) can be used to compute simulated $P$ values for $S_{sib}$. Although the minimum sample size required for the $\chi^2$ distribution to be adequate is not known, guidelines for the validity of the Mantel-Haenszel stratified $\chi^2$ statistic may prove useful (Mantel and Fleiss 1980).

Another advantage of our general approach is that the ages of unaffected sibs can be accounted for in a refined stratified analysis. This may be critical when one is evaluating complex diseases with late ages at onset, because unaffected sibs could later develop the disease. This type of analysis can be performed by further stratifying on the age at onset of affected sibs, in much the same manner as the log-rank statistic is used in survival analysis. That is, for each affected sib, create a 2 × $G$ table, where the sib controls are chosen such that the age at which they are known to be free of disease is at least the age

## Table 3

**Number of Sibships Required to Have 80% Power, under the Assumption of 5% False-Positive Rate, to Detect Association with a Genetic Marker Having Autosomal Recessive Effects on Disease Risk**

| | Frequency of | | No. of Sibships | |
|---|---|---|---|---|
| $p$, $r_{AA}$, and min $N^d$ | $\mu_{alt}$ | $\sigma_{alt}$ | Single-Locus Test | Genomewide Screen |
| .01: | | | | |
| 2: | | | | |
| 1 | .00007 | .08135 | 10,878,713 | 52,790,672 |
| 2 | .00012 | .09826 | 4,916,616 | 23,858,654 |
| 4: | | | | |
| 1 | .00021 | .08202 | 1,229,407 | 5,965,891 |
| 2 | .00051 | .10057 | 302,786 | 1,469,315 |
| 8: | | | | |
| 1 | .00048 | .08350 | 234,192 | 1,136,452 |
| 2 | .00181 | .10821 | 28,169 | 136,693 |
| .1: | | | | |
| 2: | | | | |
| 1 | .00622 | .25238 | 12,918 | 62,683 |
| 2 | .01123 | .31174 | 6,044 | 29,329 |
| 4: | | | | |
| 1 | .01833 | .26813 | 1,680 | 8,151 |
| 2 | .04471 | .35970 | 509 | 2,466 |
| 8: | | | | |
| 1 | .04145 | .29822 | 407 | 1,972 |
| 2 | .14020 | .46479 | 87 | 419 |
| .2: | | | | |
| 2: | | | | |
| 1 | .02150 | .34224 | 1,990 | 9,656 |
| 2 | .03816 | .42613 | 979 | 4,751 |
| 4: | | | | |
| 1 | .06022 | .37004 | 297 | 1,439 |
| 2 | .13382 | .49563 | 108 | 523 |
| 8: | | | | |
| 1 | .12542 | .41203 | 85 | 412 |
| 2 | .32203 | .58088 | 26 | 124 |

Note.—Data are as described in the footnotes to table 2, except that $r_{AB}$ = 1.

at which the affected sib is diagnosed. After this refined stratification is setup, the approach outlined for sib controls can be used to compute the multivariate score statistic. The conditional logistic-regression models that we propose to account for nongenetic covariates may also be critical, in order to adjust for differences in environmental risk factors among sibs. Although the score statistic and conditional logistic regression are valid methods to test the null hypothesis of no association between the genetic markers and affection status, a potential difficulty with conditional logistic regression is that the variance estimates for the maximum-likelihood relative-risk estimates may not be accurate when there are dependent data, such as when multiple affected sibs and parental controls are used or when residual sib correlation exists. Further work, such as use of generalized estimating equations, is needed to allow for this type of data.

A further advantage of our proposed approach is that it is applicable to designs using parents as controls, sibs as controls, unrelated controls whose genotypes do not fit Hardy-Weinberg proportions, or any combination of these different designs. However, if a family has both parents and sib controls available, then only the parents should be used. The benefit of this general approach is that a multivariate $\chi^2$ statistic can be used, which will tend to be the most powerful method when multiple marker alleles are associated with affection status. Alternatively, the components of this statistic can be used to compute the maximum of the univariate statistics, $Z_{\max}$, which will tend to be the most powerful method when a single marker allele is associated with affection status (Schaid 1996). If there is evidence that a candidate gene will likely have a dominant or recessive effect, then use of marker-genotype scores different from the allele-counting vectors (i.e., allele-counting vector $\mathbf{b}' = (2,1,0)$) can result in greater power, especially for recessive effects (Schaid 1996).

The methods that have been presented for computation of sample size and power, using either sib controls or parental controls, are general enough to allow one to consider varying the sibship sizes, ascertainment criteria, and genetic models of risk. Although we have presented power calculations for only two marker alleles, it is possible to compute power for $K > 2$ marker alleles, by use of the noncentral $\chi^2$ distribution, with a noncentrality parameter that depends on allele frequencies and genotype relative risks. Because it can be difficult to specify all the genotype relative risks in a plausible manner, a simpler, conservative approach is to assume that only one marker allele is positively associated with disease, then modify the $Z_\alpha$ critical value to the Bonferonni corrected value, $Z_{\alpha/(K-1)}$, and then use this corrected value in the formulas for power that are presented in the appendices. If it is anticipated that a pooled analysis will be required, because only sib controls will be available for some families, and other families will have parental controls, then the methods that have been presented for computation of sample size and power for each of these types of controls can be adapted to computation of sample size and power for a pooled analysis.

The results indicate that parental controls offer greater power than sib controls, for both dominant and recessive effects, which agrees with the findings by Spielman and Ewens (1998). The power of sib controls can be increased by either increasing the number of affected sibs per sibship or increasing the number of unaffected control sibs, with the former offering the greatest gain. The power gain produced by increasing the number of unaffected sibs follows the pattern of relative efficiency for matched case-control studies; for one affected subject matched with $M$ unaffected subjects, the relative efficiency, $M/(M + 1)$ (Ury 1975), suggests little gain in power when the number of unaffected controls per each affected subject exceeds four.

The sample-size results suggest that the use of sib controls to test for associations, by either a single-marker-locus test or a genomewide screen, will be feasible either for markers that have a dominant effect and for common alleles that have a recessive effect. Although we have assumed that the lifetime risk of disease is 5%, the sample sizes presented in tables 2 and 3 are only 2%–22% larger than those required when the lifetime risk is 10%. Hence, the results in tables 2 and 3 can be useful guidelines for investigators planning studies using sibs as controls.

## Acknowledgment

This research was supported by the U.S. Public Health Services, National Institutes of Health, grant GM51256.

## Appendix A

### Power Calculations for Sib Controls

To illustrate computation of power for sib controls, denote the number of affected sibs with genotypes *AA*, *AB*, and *BB* as $x_1$, $x_2$, and $x_3$, respectively, with the vector $\mathbf{x}_g = (x_1,x_2,x_3)$; analogous counts for unaffected sibs are denoted as $y_1$, $y_2$, and $y_3$, respectively, with vector $\mathbf{y}_g = (y_1,y_2,y_3)$. The total number of affected sibs in a sibship is $\mathrm{N}^{\mathrm{d}}$, the total number of unaffected sibs is $\mathrm{N}^{\mathrm{c}}$, and $\mathrm{N} = \mathrm{N}^{\mathrm{d}} + \mathrm{N}^{\mathrm{c}}$. The possible values of $x_i$ and $y_i$ are illustrated in table A1 for each of the six parental mating types. Note that sibships are not informative for associations when all offspring have the same genotype, as is the case with mating types 1, 3, and 6.

**Table A1**

**Parental Mating Types and Possible Sib Genotype Counts, with Mendelian Genotype Probabilities**

| Parental Mating Type | Possible Value of Sib Genotype Counts[a] | | |
|---|---|---|---|
| | *AA* | *AB* | *BB* |
| 1. *AA* × *AA* | $N^d$ | 0 | 0 |
| | $N^c$ | 0 | 0 |
| | (1) | (0) | (0) |
| 2. *AA* × *AB* | $x_1 = 0,\ldots,N^d$ | $x_2 = N^d - x_1$ | 0 |
| | $y_1 = 0,\ldots,N^c$ | $y_2 = N^c - y_1$ | 0 |
| | $(\frac{1}{2})$ | $(\frac{1}{2})$ | (0) |
| 3. *AA* × *BB* | 0 | $N^d$ | 0 |
| | 0 | $N^c$ | 0 |
| | (0) | (1) | (0) |
| 4. *AB* × *AB* | $x_1 = 0,\ldots,N^d$ | $x_2 = 0,\ldots,N^d - x_1$ | $x_3 = N^d - x_1 - x_2$ |
| | $y_1 = 0,\ldots,N^c$ | $y_2 = 0,\ldots,N^c - y_1$ | $y_3 = N^c - y_1 - y_2$ |
| | $(\frac{1}{4})$ | $(\frac{1}{2})$ | $(\frac{1}{4})$ |
| 5. *AB* × *BB* | 0 | $x_2 = 0,\ldots,N^d$ | $x_3 = N^d - x_2$ |
| | 0 | $y_2 = 0,\ldots,N^c$ | $y_3 = N^c - y_2$ |
| | (0) | $(\frac{1}{2})$ | $(\frac{1}{2})$ |
| 6. *BB* × *BB* | 0 | 0 | $N^d$ |
| | 0 | 0 | $N^c$ |
| | (0) | (0) | (1) |

[a] Values in parentheses are Mendelian genotype probabilities.

Conditional on the mating type, the probabilities of the offspring's genotypes are given by Mendelian segregation probabilities, denoted as $p_1$, $p_2$, and $p_3$ for genotypes *AA, AB,* and *BB,* respectively; these probabilities are also illustrated in table A1. Furthermore, conditional on a mating type but not on the ascertainment criteria, the joint probability of the offspring's affection status and genotypes in a particular table, say the *j*th table, denoted as $T_j$, is

$$P(T_j|m) = \binom{N}{x_1 x_2 x_3 y_1 y_2 y_3} \prod_{i=1}^{3} (p_i f_i)^{x_i} [p_i(1-f_i)]^{y_i} \ . \qquad \text{(A1)}$$

Now, conditional on the ascertainment criteria but not on parental mating types, the probability of table $T_j$ is

$$P(T_j) = \frac{P(T_j|m)P(m)}{\sum\limits_{T_s \epsilon S} P(T_s|m)P(m)} \ , \qquad \text{(A2)}$$

where $P(m)$ is the probability of mating type *m,* determined by Hardy-Weinberg proportions and allele frequencies, and the sum in the denominator is over the set *S* of all possible 2 × 3 tables that are consistent with the ascertainment scheme. For example, the sum in the denominator of expression (A2) would equal 1 if $N^d$ were allowed to vary from 0 to *N*. However, sibships with no affected sibs are not informative, so, in practice, $N^d$ will be $\geqslant 1$. For particular values of $N^d = i$ and $N^c = N - i$, the number of 2 × 3 tables to enumerate is $M_i = 3 + 2(i+1)(N-i+1) + (i+1)(i+2)(N-i+1)(N-i+2)/4$. So, the total number of tables in the set *S, M,* is the sum of the values of $M_i$ over the range of values of $N^d$ that are consistent with the ascertainment scheme.

Now, the difference between the observed count of *A* alleles among cases and that predicted by the marginal genotype distribution among *N* sibs, the difference denoted as $\delta$, is the random variable of interest, and its mean and variance will determine power. To determine these moments, collect the first row (i.e., $\mathbf{x}_g$ counts for cases) from all *M* tables and bind these into the *M* × 3 matrix **X**. A matrix **Y** is similarly created for the $\mathbf{y}_g$ counts for controls from all tables. The marginal genotype counts of the tables is the matrix $\mathbf{T} = \mathbf{X} + \mathbf{Y}$. The random variable $\delta$ can be computed for all *M* tables by means of the matrix notation $\delta = (\mathbf{X} - \mathbf{T}N^d/N)\mathbf{b}$, where $\mathbf{b}'$ is the vector (2,1,0). The expected value of $\delta$ under the alternative hypothesis is $\mu_{\text{alt}} = \Sigma_{T_j \epsilon S} \delta_j P(T_j)_{\text{alt}}$, where $P(T_j)_{\text{alt}}$ is computed on the basis of expressions (A1) and (A2), with $f_i$ values that depend on the genotype relative risks specified by the alternative hypothesis. Under the null hypothesis, $\mu_{\text{nul}} = 0$. The variance of $\delta$ under the alternative hypothesis is

$$\sigma^2_{\text{alt}} = \sum_{T_j \epsilon S} (\delta_i - \mu_{\text{alt}})^2 P(T_j)_{\text{alt}} \ . \qquad \text{(A3)}$$

The variance of $\delta$ under the null hypothesis, $\sigma^2_{\text{nul}}$, can be calculated on the basis of expression (A3) but with substitution of $\mu_{\text{nul}} = 0$ for $\mu_{\text{alt}}$ and of $P(T_j)_{\text{nul}}$ for $P(T_j)_{\text{alt}}$, where $P(T_j)_{\text{nul}}$ is computed with all $f_i$ values equal, as under the null hypothesis. Sibships that are not informative, because all sibs have the same genotype, have $\delta = 0$ and hence reduce power by reduction of the effective sample size. Power or sample size (no. of sibships, $N_s$) can be determined by solving the following expression for either $Z_\beta$ or $N_s$, respectively,

$$\sqrt{N_s}\,|\mu_{\text{alt}}| = Z_\alpha \sigma_{\text{nul}} + Z_\beta \sigma_{\text{alt}} \ , \qquad \text{(A4)}$$

where $Z_\alpha$ and $Z_\beta$ are the $(1-\alpha)$th and $(1-\beta)$th percentiles of a standard normal distribution, giving type I error of $\alpha$ and power of $1-\beta$. However, the accuracy of this asymptotic method depends on having a large amount of statistical information, and the magnitude of statistical information decreases with decreasing allele frequencies and marker-genotype relative risks. Simulations (not shown) have indicated that expression (A4) overestimates power when $\sigma_{\text{nul}}$ dramatically differs from $\sigma_{\text{alt}}$, which occurs when marker alleles are rare. A closer approximation, which slightly underestimates power, is given when $\sigma_{\text{nul}}$ is replaced by $\sigma_{\text{alt}}$ in expression (A4), which results in

$$\sqrt{N_s}\,|\mu_{\text{alt}}| = (Z_\alpha + Z_\beta)\sigma_{\text{alt}} \ . \tag{A5}$$

The power computations given above apply when the number of sibs within a sibship, $N$, is constant. Since $N$ will likely vary, the anticipated distribution of $N$ can be used to more accurately estimate power or sample size. For example, the distribution of $N$ could be estimated by either pilot data or an assumed negative binomial distribution (Cavalli-Sforza and Bodmer 1971, p. 313). With this distribution assumed, the values of $\mu_{\text{alt}}$ and $\sigma_{\text{alt}}$ in expression (A5) can be replaced with weighted averages, in which each of these parameters is computed for each value of $N$ and the results are combined by use of the probabilities of the different values of $N$ as weights.

## Appendix B

### Power Calculations for Parental Controls

When there are two marker alleles, the TDT statistic can be used for parental controls. For this, one needs to count the total number of times that heterozygous parents transmit the $A$ allele, denoted "$n_A$," and the total number of times that heterozygous parents transmit the $B$ allele, denoted "$n_B$." The total number of informative transmissions is $n_T = n_A + n_B$. The TDT statistic can be written as

$$\text{TDT} = \frac{(n_A - n_B)^2}{n_A + n_B} = \frac{(\hat{\pi} - .5)^2}{1/4n_T} \ ,$$

where $\hat{\pi} = n_A/n_T$, the estimated frequency of transmission of an $A$ allele from a heterozygous parent to an affected child. To compute sample size and power, we need to determine the expected values of $n_A$ and $n_T$ for a given alternative hypothesis and sampling scheme.

First consider computing the expected value of $n_T$, denoted "E[$n_T$]." For the six mating types in table A1, only the three with at least one heterozygous parent (i.e., mating types 2, 4, and 5) are informative for determination of transmission status of alleles. Because mating types 2 and 5 have only one heterozygous parent, $N^d$ affected sibs from each of these mating types contribute a count of $N^d$ to the value of $N^T$, and, because mating type 4 has two heterozygous parents, $N^d$ affected sibs from this mating type contribute a count of $2N^d$ to the value of $n_T$. So the expected value of $n_T$ is

$$\text{E}[n_T] = \sum_{j\varepsilon S_2} P(T_j)N_i^d + \sum_{j\varepsilon S_4} P(T_j)2N_i^d + \sum_{j\varepsilon S_5} P(T_j)N_i^d \ , \tag{B1}$$

where $S_m$ is the set of 2 × 3 tables that originate from mating type $m$.

The expected value of $n_A$, the number of times that $A$ is preferentially transmitted over $B$, can be determined by calculation of the number of $A$ transmissions, as follows. Let $x_{ji}$ denote the number of affected sibs from table $T_j$ who are in the $i$th genotype category ($i$ = 1,2,3 for genotypes $AA$, $AB$, and $BB$, respectively). For tables that originate from mating type 2, the number of $A$ transmissions is $x_{j1}$; from mating type 4, $(2x_{j1} + x_{j2})$; from mating type 5, $x_{j2}$. So the expected value of $n_A$ can be represented as

$$\text{E}[n_A] = \sum_{j\varepsilon S_2} P(T_j)x_{ji} + \sum_{j\varepsilon S_4} P(T_j)[2x_{j1} + x_{j2}] + \sum_{j\varepsilon S_5} P(T_j)x_{j2} \ , \tag{B2}$$

where $P(T_j)$ is given by expression (A2) and each of the summations is restricted to the indicated set of tables $S_m$ that originate from mating type $m$. Using expressions (B1) and (B2), we can calculate $\pi = \text{E}[n_A]/\text{E}[n_T]$ and, consequently, can determine power or sample size ($N_s$ = no. of sibships), by solving for $Z_\beta$ or $N_s$ in the expression

$$\sqrt{N_s}\sqrt{\text{E}[n_T]}\ |\pi - .5| = Z_\alpha/2 + Z_\beta\sqrt{\pi(1 - \pi)} \ .$$

If the size of the sibship varies, then E[$n_A$] and E[$n_T$] should be computed for each different sibship size and then averaged, much as in the method described for sib controls.

## Appendix C

### Power Calculations for Unrelated Controls

When unrelated controls are sampled, the probability of genotype $g$ among the controls is $P(g|c) = [(1 - f_g)P(g)]/\Sigma_g(1 - f_g)P(g)$, which can be enumerated, for all genotypes, in the vector $\mathbf{p}_c' = [P(AA \mid c), P(AB \mid c), P(BB \mid c)]$. For the affected cases, the probability of genotype $g$ is $P(g|d) = f_gP(g)/\Sigma_g f_gP(g)$, enumerated in the vector $\mathbf{p}_d' = [P(AA \mid d), P(AB \mid d), P(BB \mid d)]$.

For the 2 × 3 table of genotype counts, the expected marginal distribution of genotypes is $\mathbf{p}_t = (\mathbf{p}_d + \mathbf{p}_cR_c)/(R_c + 1)$, where $R_c$ denotes the ratio of the number of controls to the number of cases. The difference between the number of $A$ alleles among cases that are expected to occur under a specified hypothesis and the number that is predicted by the marginal genotype distribution is $\mu_{\text{alt}} = N^d\mathbf{b}'(\mathbf{p}_d - \mathbf{p}_t)$. Under the null hypothesis, $\mu = 0$, but, as the genotype relative risks deviate from 1, $\mu_{\text{alt}}$ deviates from 0. Under the alternative hypothesis, the covariance matrix of genotype counts among cases is approximated by

$$\mathbf{V}_{\text{alt}} = N^d[\text{diag}(\mathbf{p}_t) - \mathbf{p}_t\mathbf{p}_t']R^c/(R^c + 1)\ . \qquad \text{(C1)}$$

So the variance of the number of $A$ alleles among cases is $\sigma^2_{\text{alt}} = \mathbf{b}'\mathbf{V}_{\text{alt}}\mathbf{b}$. The variance under the null hypothesis, $\sigma^2_{\text{nul}}$, can be computed by substitution of the vector of genotype probabilities determined by Hardy-Weinberg proportions, $\mathbf{p}_g$, for the vector $\mathbf{p}_t$ in expression (C1). On the basis of these results, power or sample size ($N^d$ = no. of cases) can be calculated by solving the following expression for either $Z_\beta$ or $N^d$, respectively,

$$\sqrt{N^d}|\mu_{\text{alt}}| = Z_\alpha\sigma_{\text{nul}} + Z_\beta\sigma_{\text{alt}}\ . \qquad \text{(C2)}$$

Then the total sample size is $N^d(1 + R^c)$. Similar to the asymptotic accuracy regarding sib controls in expression (A4), simulations (not shown) indicate that a more accurate method than that of expression (C2), albeit conservative, is to use the following expression to determine sample size and power: $\sqrt{N^d}|\mu_{\text{alt}}| = (Z_\alpha + Z_\beta)\sigma_{\text{alt}}$.

## References

Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961

Brass W (1958) Models of birth distributions in human populations. Bull Inst Int Stat 36:165–178

Cavalli-Sforza LL, Bodmer WF (1971) The genetics of human populations. WH Freeman, San Francisco

Clerget-Darpoux F, Bonaiti-Pellie C, Honchez J (1986) Effects of misspecifying genetic parameters in lod score ananlysis. Biometrics 42:393–399

Cox DR (1975) Partial likelihood. Biometrika 62:269–276

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

Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56:811–812

Ewens WJ, Spielman RS (1997) The sib-TDT (S-TDT): a TDT (transmission/disequilibrium test) without parents. Am J Hum Genet Suppl 61:A275

Lander ES (1996) The new genomics: global views of biology. Science 274:536–539

Langefeld CD, Pericak-Vance MA, Saunders AM, Boehnke M (1997) Family-based tests for association using discordant sib pairs. Am J Hum Genet Suppl 61:A282

Mantel N, Fleiss JL (1980) Minimum requirements for the Mantel-Haenszel one-degree of freedom chi-square test and a related rapid procedure. Am J Epidemiol 112:129–134

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from the retrospective study of disease. J Natl Cancer Inst 22:719–748

Manuila A (1958) Blood groups and disease—hard facts and delusions. JAMA 167:2047–2053

Monks SA, Martin ER, Weir BS, Kaplan NL (1997) A sibship test of linkage in the absence of parental information. Am J Hum Genet Suppl 61:A286

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Schaid DJ (1995) Relative-risk regression models using cases and their parents. Genet Epidemiol 12:813–818

——— (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

——— (1998) Likelihoods and *TDT* for the case-parents design. Genet Epidemiol (in press)

Schaid DJ, Jacobsen SJ (1998) Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions. Am J Epidemiol (in press)

Schaid DJ, Li H (1997) Genotype relative-risks and association tests for nuclear families with missing parents. Genet Epidemiol 14:1113–1118

Self SG, Longton G, Kopecky KJ, Liang KY (1991) On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–61

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

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

Ury HK (1975) Efficiency of case control studies with multiple controls per case: continuous or dichotomous data. Biometrics 31:643–649