Power Calculations for a General Class of Family-Based Association Tests: Dichotomous Traits

Christoph Lange and Nan M. Laird

Department of Biostatistics, Harvard School of Public Health, Boston

Using large-sample theory, we present a unified approach to power calculations for family-based association tests. Currently available methods for power calculations are restricted to special designs or require approximations or simulations. Our analytical approach to power calculations is broadly applicable in many settings. We discuss power calculations for two scenarios that have high practical relevance and in which power previously could only be assessed by simulation studies or by approximations: (1) studies using both affected and unaffected offspring and (2) studies with missing parental information. When the population prevalence is high, it can be worthwhile to genotype unaffected offspring. For many scenarios, high power can be achieved with reasonable sample sizes, even when no parental information is available.

Introduction

In this article, we address power calculations for generalized family-based association tests (FBATs) (Laird et al. 2000; Rabinowitz and Laird 2000). We use the term "FBAT" to denote a genetic-association test that uses genetic data on family members to compute the distribution of a suitable test statistic under the null hypothesis, conditioning on the phenotypes. Examples are given below. FBATs are powerful tests for detecting linkage between a marker and a disease-susceptibility locus in the presence of linkage disequilibrium between the two loci. The best known FBAT is the transmission/disequilibrium test (TDT) (Ott 1989; Spielman et al. 1993); it was designed for the special setting of sampling affected individuals and their parents, where parents' genotypes are available. However, many genetic studies have missing parents, affected and unaffected offspring, continuous phenotypes, and/or multiple phenotypes. A variety of approaches have been proposed to deal with these issues; for reviews, see Zhao (2000) or Schulze and McMahon (2002). Most research has focused on the distribution of the proposed test statistics under the null hypothesis and has assessed the achieved power of the proposed tests either by simulation studies (Risch 2000; Horvath et al. 2001; Q. Yang, X. Xu, and N. M. Laird, unpublished data) or by approximations (Whittaker and Lewis 1998).

For "simple" scenarios (i.e., trios or trios with one additional offspring), Knapp (1999*a*) and Chen and Deng (2001) computed the power by deriving the expected value of the test statistic under the alternative hypothesis and then computing the power of the expected statistic. In contrast with that, we compute the expected power of the actual test statistic. First, we obtain the power of the test statistic, conditional on the phenotypes and mating types. Then, we integrate the conditional power over the phenotypes and mating types, to obtain the expected power. Since we are able to compute the conditional power for virtually any scenario (e.g., multiple offspring, missing parental information, etc.), and since integrating over the data can always be solved numerically, our approach to power calculation can be applied much more generally. The approach is broadly applicable to multiallelic loci, continuous traits, and/or multivariate phenotypes (Lange et al., in press). Here, we will discuss only the situation in which we observe one dichotomous trait; continuous traits will be discussed in a separate article.

To illustrate the generality of our new approach, we compare power results obtained by our approach with those of both Knapp (1999*a*) and Whittaker and Lewis (1998). Our approach and Knapp's agree well in the case in which Knapp's has been applied. Since Whittaker and Lewis (1998) derived their results under the assumption that the alternative hypothesis would be very close to the null hypothesis, the two approaches demonstrate perfect agreement when the alternative hypotheses are indeed close to the null hypothesis. However, when the alternative hypotheses are far away from the null hypothesis, the Whittaker and Lewis (1998) method becomes less reliable. In addition, we will present power calculations for situations in which the pa-

Received February 21, 2002; accepted for publication June 17, 2002; electronically published August 12, 2002.

Address for correspondence and reprints: Dr. Christoph Lange, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115. E-mail: clange@hsph.harvard.edu

0002-9297/2002/7103-0012$15.00

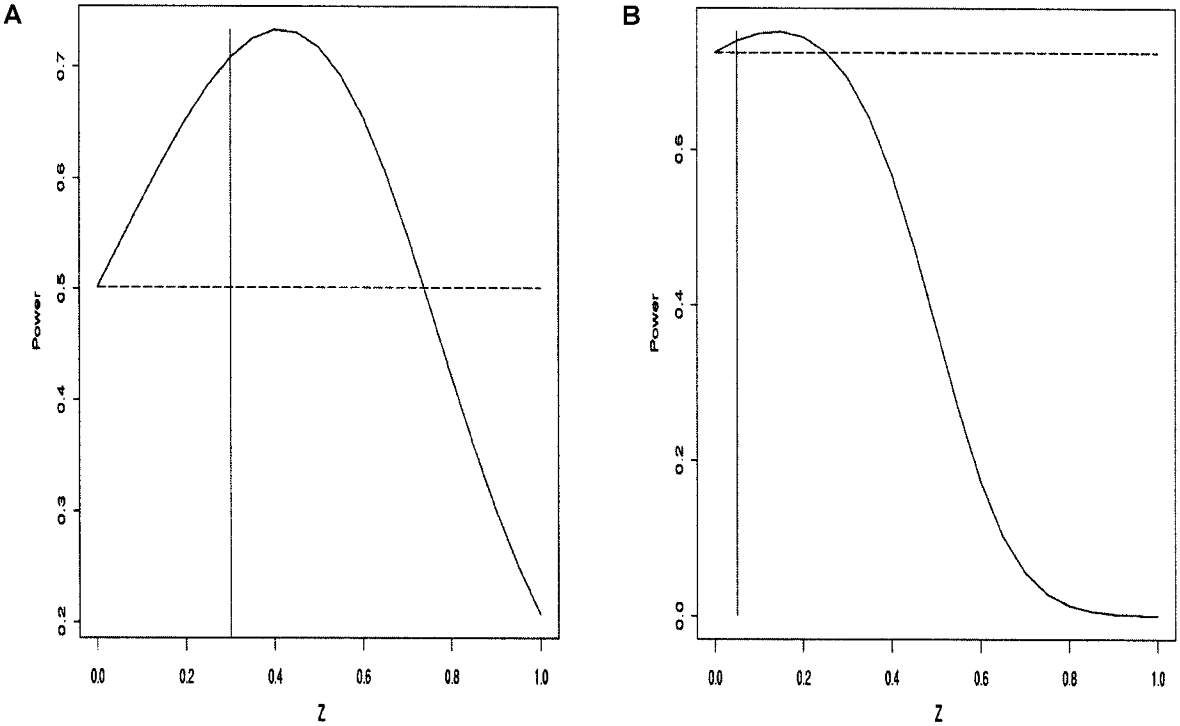

Figure 1 Power of FBAT test for trios with one additional unaffected offspring and both parents' genotypes observed. The dotted line shows the power of the standard TDT, including only the affected offspring. The vertical line shows the location of the disease prevalence K. *a,* Multiplicative model for a common disease: disease prevalence $K = 0.3$, allele frequency of the disease gene $p = 0.143$, fraction of the disease attributable to carrying at least one disease gene AF $= 0.25$, significance level $\alpha = 0.01$, and sample size 100. Optimal offset choice $z = 0.4$. Power gain by optimal choice of z over $z = 0$ is 25%. *b,* Multiplicative model for a rare disease: disease prevalence $K = 0.05$, allele frequency of the disease gene $p = 0.05$, fraction of the disease attributable to carrying at least one disease gene AF $= 0.3$, significance level $\alpha = 10^{-4}$, and sample sizes 100. Optimal choice of $z = 0.15$. Power gain by optimal choice of z over $z = 0$ is 5%.

rental genotypes are missing, but additional offspring are available.

In our section on "Notation and Power Calculations," we discuss how the unconditional/expected power can be computed. In the following section, on "Computation of the Conditional Marker Distribution and the Conditional Family-Type Distribution," we derive the conditional probabilities required for the computation of the unconditional power. Our approach is generally applicable and can handle a variety of different scenarios—for example, multiple offspring, multiple phenotypes, missing parental information, and environmental effects. Our "Results" section shows power calculations for scenarios in which additional offspring are given and/or parental genotypic information is missing. Finally, in our section on "Application to Study Design," we use our approach to power calculations to design a study for bipolar disorder.

Notation and Power Calculations

For simplicity of exposition, we assume that we observe a biallelic marker with alleles A and B. The disease penetrances for 0, 1, or 2 disease alleles are f_0, f_1, and f_2, respectively. We denote the allele frequency of the disease gene by p, the population prevalence of the disease by K, and the fraction of the disease attributable to carrying at least one copy of the disease gene by AF—that is, AF $= (K - f_0)/K$. Like Risch and Merikangas (1996), Camp (1997), Knapp (1999*a*) and Whittaker and Lewis (1998), we assume the best-case scenario for the marker locus—that is, that the marker locus is the disease locus and that, hence, the A allele frequency is p.

Furthermore, there are n independent families, and the ith family has m_i offspring. We denote the marker score for jth offspring in the ith family by X_{ij}. The total number of marker scores $X_{ij}, j = 1,\ldots,m_i; i = 1,\ldots,n$ is N. Note that the actual coding of the marker score depends on the assumed genetic model. The corresponding trait information is given by Y_{ij}, where affected offspring are coded by $Y_{ij} = 1$, unaffected offspring by $Y_{ij} = 0$, and offspring with unknown phenotype by $Y_{ij} = NA$. When parental genotypic information for the ith family is recorded, it is denoted by P_{i1} and P_{i2}. For biallelic markers, the possible values of P_{i1} and P_{i2} can be characterized as 0, 1, or 2 for the number of target alleles. Laird et al. (2000) then defined the generalized FBAT statistic by

$$GT = \frac{\left(\sum_i U_i\right)^2}{\sum_i \mathrm{Var}\,(U_i)} \;.$$

In this equation, $U_i = \sum_j T_{ij}[x_{ij} - E_0\,(X_{ij})]$, $\mathrm{Var}\,(U_i) = \sum_{j,j'} T_{ij}T_{ij'}\,\mathrm{Cov}_0\,(X_{ij},X_{ij'})$, and T_{ij} an appropriate coding of the phenotype Y_{ij}. $E_0(X_{ij})$ is the expected marker score under the null hypothesis and $\mathrm{Cov}_0\,(X_{ij},X_{ij'})$ is the marker covariance under the null hypothesis. When the phenotype is missing (i.e., when $Y_{ij} = NA$), we set $T_{ij} = 0$. When both parents are observed, setting $T_{ij} = y_{ij}$ gives the TDT discussed by Spielman et al. (1993), which is based on affected ($y_{ij} = 1$) offspring only. Setting $T_{ij} = y_{ij} - z$, with a constant offset z,$0 < z < 1$, defines the TDT proposed by Whittaker and Lewis (1998), which includes unaffected offspring in the computation of the test statistic. When both parents are observed, the marker means and covariances under the null hypothesis, $E_0(X_{ij})$ and $\mathrm{Cov}_0\,(X_{ij},X_{ij'})$, are computed conditional on the parental genotypes, P_{i1} and P_{i2}, where the transmission probabilities are defined by Mendelian laws.

When parents are missing, the RC-TDT (Knapp 1999*b*), S-TDT (Spielman and Ewens 1998), and FBAT (Laird et al. 2000; Rabinowitz and Laird 2000) use alternative conditions for the computation of the marker means and covariances under the null hypothesis (e.g., the minimal sibship condition for S-TDT or the R-condition for RC-TDT). These conditions are based on suitable sets of available genetic information within one family and the observed phenotypes. Loosely speaking and without loss of generality, any of those conditions can be understood as a function of offspring genotypes and available parental genotypes that is held constant when $E_0(X_{ij})$ and $\mathrm{Cov}_0\,(X_{ij},X_{ij'})$ are computed. We will denote these conditions by S_i. Although the methodology proposed here is valid for all conditions addressed by RC-TDT, S-TDT, and FBAT, we will use the FBAT condition in all our examples. For FBAT, S_i is the minimal sufficient statistic for the parental genotypes (Rabinowitz and Laird 2000).

Standard asymptotic theory implies that, under the null hypothesis and given the phenotypes $\mathbf{Y} = (Y_{11},\dots,Y_{nm_n})$ and the condition $\mathbf{S} = (S_1,\dots,S_n)$, GT is χ^2 distributed with 1 df—that is,

$$GT \sim \chi^2_1 \;.$$

When the marker means $E_A(X_{ij})$ and (co)variances $\mathrm{Cov}_A\,(X_{ij},X_{ij'})$ are given under the alternative hypothesis, the distribution of GT under the alternative hypothesis can be computed by a scaled, noncentral χ^2 distribution: $\omega GT \sim \chi^2_{1,\gamma}$, with

$$\gamma = \frac{\left\{\sum_{ij} T_{ij}[E_A(X_{ij}) - E_0(X_{ij})]\right\}^2}{\sum_i \sum_{j,j'} T_{ij}T_{ij'}\,\mathrm{Cov}_A\,(X_{ij},X_{ij'})}$$

and

$$\omega = \frac{\sum_i \sum_{j,j'} T_{ij}T_{ij'}\,\mathrm{Cov}_A\,(X_{ij},X_{ij'})}{\sum_i \sum_{j,j'} T_{ij}T_{ij'}\,\mathrm{Cov}_0\,(X_{ij},X_{ij'})} \;. \qquad (1)$$

The proof of this result is shown in appendix A. The extension of the conditional power formula (1) to multiallelic loci can be found in the work of Lange and Laird (in press). In this setup, the conditional power of GT for the significance level α is given by

$$\mathcal{P}^{\mathrm{cond}}_{GT\,|\,\mathbf{Y},\mathbf{S}} = P\left[\chi^2_{1,\gamma} \geqslant \omega q_{\chi^2_1}(1-\alpha)\right] \;. \qquad (2)$$

It is important to note that, since $E_0(.)$, $E_A(.)$, $\mathrm{Cov}_0\,(.,.)$ and $\mathrm{Cov}_A\,(.,.)$ are computed conditional on $\mathbf{Y}$ and $\mathbf{S}$, formula (2) can not be used directly when the phenotypes $\mathbf{Y}$ and the data defining $\mathbf{S}$ are not observed. $\mathbf{Y}$ and $\mathbf{S}$ have to be integrated out to obtain the unconditional/expected power—that is,

$$\mathcal{P}_{GT\,|\,\mathcal{A}} = E(\mathcal{P}^{\mathrm{cond}}_{GT\,|\,\mathbf{Y},\mathbf{S}} \mid \mathcal{A}) \;, \qquad (3)$$

where $\mathcal{A}$ is the ascertainment condition for the phenotype $\mathbf{Y}$. We make the assumption that the ascertainment condition depends only on the phenotype $\mathbf{Y}$ and the expectation is over the distribution of $\mathbf{Y},\mathbf{S} \mid \mathcal{A}$, but the approach can be extended so that $\mathcal{A}$ depends on the phenotypes of the parents. Thus, the second stage of computing the unconditional/expected power involves the conditional distribution of the family types $p\,(\mathbf{y}_i = (y_{i1},\dots,y_{im_i}),s_i \mid \mathcal{A})$, under the alternative hypothesis. Since $\mathbf{y}_i,\mathbf{s}_i$ are discrete and bounded random variables, (3) can be written as a finite sum:

$$\mathcal{P}_{GT\,|\,\mathcal{A}} = \sum_{\mathbf{y},\mathbf{s}} \mathcal{P}^{\mathrm{cond}}_{GT\,|\,\mathbf{y},\mathbf{s}}\,p\,(\mathbf{y},\mathbf{s} \mid \mathcal{A}) \;. \qquad (4)$$

Hence, $\mathcal{P}_{GT\,|\,\mathcal{A}}$ can always be computed by direct evaluation of each possible term in the summation. In appendix B, we describe technical details that accelerate the computation of (4).

When the power function in the unconditional/expected power formula (3) is approximated by a first-order Taylor expansion in $\mathbf{Y}$ and $\mathbf{S}$—that is, by E[Power(FBAT)] $\hat{=}$ Power[E(FBAT)]—the approach of Knapp (1999*a*) is obtained. In full generality, the Lagrange term in the Taylor approximation—and, therefore, the accuracy of Knapp's approach (1999*a*)—will depend on the second moment expressions of **S** and the phenotypes **Y**, which are not fixed by the ascertainment condition. Since all phenotypes are fixed at 1 in the scenarios considered by Knapp (1999*a*), and, since the parental genotypes are observed, $S = (\mathbf{P}_1,\mathbf{P}_2)$, the approximation error is minimal. However, when not all phenotypes are defined by the ascertainment condition (e.g., only the first offspring must be affected) and **S** becomes more complex (e.g., parental genotypic information is missing), the approximation error can become considerably larger.

Although the numerical differences we found between our approach and Knapp's were usually not noteworthy for the scenarios considered by Knapp (1999*a*) and Chen and Deng (2001), the theoretical advantages of our methodology are of practical relevance. Since the direct computation of the unconditional mean and variance of $GT \mid \mathbf{Y},\mathbf{S}$ is already rather difficult for the scenarios considered in Knapp (1999*a*), and since approximations have to be utilized (Knapp 1999*a*), potential extensions to more-realistic scenarios become even more complex (Chen and Deng 2001). On the other hand, the methodology proposed here can easily be applied to complex scenarios (e.g., missing parental genotypes and multiple continuous phenotypes per offspring). Result (1) for conditional power calculations allows us to compute the conditional power for any scenario, as long as we are able to derive the conditional marker distribution under the alternative hypothesis. In the next section (formula [5]), we show that this can be done for virtually any scenario. Then the second step, "unconditioning" the conditional power, can be achieved numerically at all times, either by numerical summation, numerical integration, Monte Carlo simulation, or Markov chain–Monte Carlo (MCMC) methods.

Note that, when numerical integration/summation is not feasible, computing the sum by Monte Carlo simulations or MCMC is more efficient than assessing the power by simulations. In a pure simulation experiment, the test result is either significant or not significant, which means that we are looking at a discrete variable that can either be 1 or 0. However, when Monte Carlo simulation is used for the computation of the sum in (4), the variable of interest is the conditional power, which is a continuous variable between 0 and 1. It is obvious that a continuous variable contains more information and has less variance than a discrete variable. A pure simulation study will therefore require far more replicates than computing the sum of the conditional power by Monte Carlo simulation. Finally, we note that Knapp (1999*a*) and Chen and Deng (2001) compute an approximation to the power of the "expected" test statistic rather than the expected power of the actual test statistic.

Computation of the Conditional Marker Distribution and the Conditional Family-Type Distribution

In this section, we discuss the computation of the conditional marker distribution under both hypotheses and the conditional distribution of the family types (i.e., $\mathbf{y}_i$ and s_i) under the alternative hypothesis. Although the conditional marker distribution is required in the conditional power calculation for the scaling parameter ω and the noncentrality parameter γ, the distribution of the family types conditional on the ascertainment condition is needed to integrate out these variables in (3) to and obtain the unconditional/expected power.

Since we assume independence of the families, it is sufficient to discuss the marker distribution within one family. The conditional marker distribution $p[\mathbf{x}_i = (x_{i1},\dots,x_{im_n}) \mid \mathbf{y}_i,s_i]$ can be derived by repeated application of Bayes's theorem and is given, under the null and alternative hypotheses, by

$$P(\mathbf{x}_i \mid \mathbf{y}_i,s_i) = \frac{p(\mathbf{y}_i \mid \mathbf{x}_i)p(\mathbf{x}_i \mid s_i)}{\sum\limits_{\mathbf{x}} p(\mathbf{y}_i \mid \mathbf{x})p(\mathbf{x} \mid s_i)} \ . \qquad (5)$$

When the marker locus and the disease locus are not the same, the probability of the disease locus $\mathbf{g}_i$, given the marker locus $\mathbf{x}_i$, has to be computed, by $p(\mathbf{g}_i \mid \mathbf{x}_i)$. Then $p(\mathbf{y}_i \mid \mathbf{x}_i)$ in (5) is replaced by $\sum_{\mathbf{g}_i} p(\mathbf{y}_i \mid \mathbf{g}_i)p(\mathbf{g}_i \mid \mathbf{x}_i)$.

The probability $p(\mathbf{x}_\mathrm{i} \mid s_i)$ can be computed under both hypotheses by repeated application of Bayes's theorem. Under the null hypothesis, note that $p\,(\mathbf{y}_i \mid \mathbf{x}_i)$ does not depend on the marker score and (5) simplifies to $P\,(\mathbf{x}_i \mid s_i)$, which will not depend upon model assumptions and population parameters (appendix I in Rabinowitz and Laird 2000). Nevertheless formula (5) is of practical relevance also under the null hypothesis, since it provides a standardized way for the computation of the conditional marker distribution that can easily be implemented in a software package, e.g. when either one or both parents are missing it can be used as an alternative algorithm to compute the conditional marker distribution given in Rabinowitz and Laird (2000).

When trios with one offspring are ascertained, $m_i = 1$, and the parental genotypes are observed, the sufficient statistic S_i is given by p_{i1},p_{i2} and the conditional marker distribution for the TDT by Spielman and Ewens (1998) can be obtained by direct application of formula (5),

$$P(x_i \mid y_i = 1,s_i) = \frac{P(y_i = 1 \mid x_i)P(x_i \mid p_{i1},p_{i2})}{\sum\limits_{x} P(y_i = 1 \mid x)P(x \mid p_{i1},p_{i2})} \ ,$$

where $\sum_x$ is over all possible values of x_i, given p_{i1} and p_{i2}; it depends on the marker coding, as well as on p_{i1}

and p_{i2}. The probability $P(x \mid p_{i1},p_{i2})$ is given by Mendelian law under both H_0 and H_A. The conditional distribution $P(y_i \mid x_i)$ is given, under the null hypothesis, by

$$P(y_i \mid x_i) = K^{y_i}(1-K)^{1-y_i} \ , \qquad (6)$$

where K is the disease prevalence and under the alternative hypothesis by

$$P(y_i \mid x_i) = f_{x_i}^{y_i}(1-f_{x_i})^{1-y_i} \ , \qquad (7)$$

where f_0, f_1, and f_2 are the penetrances of the underlying disease model. The application of formula (5) to more complex scenarios (e.g., multiple offspring and missing parental information) will be discussed in our "Results" section.

The computation of the distribution of the family type $(\mathbf{y}_i,s_i)$, given $\mathcal{A}$, is done in a similar way. The ascertainment condition $\mathcal{A}$ describes how offspring are sampled from the total population on the basis of their phenotypes. For example, for the TDT proposed by Spielman and Ewens (1998), the ascertainment condition is given by $\mathcal{A} = \{y{:}y = 1\}$, which means that only affected offspring are used in the test. Given the ascertainment condition $\mathcal{A}$, the conditional distribution of $(\mathbf{y}_i,s_i)$ can be computed by

$$P(\mathbf{y}_i,s_i \mid \mathcal{A}) = \frac{\mathbf{I}\{\mathbf{y}_\mathrm{i} \in \mathcal{A}\} P(\mathcal{A} \mid \mathbf{s}_\mathrm{i})\mathbf{p}(\mathbf{s}_\mathrm{i})}{p(\mathcal{A})} \ , \qquad (8)$$

with $I\{\mathbf{y}_i \in \mathcal{A}\} = 1$ for $\mathbf{y}_i \in \mathcal{A}$ and $I\{\mathbf{y}_i \in \mathcal{A}\} = 0$ otherwise. As for the conditional marker distribution, probabilities $p(\mathcal{A})$, $p(\mathcal{A} \mid s_i)$, and $p(s_i)$ can be computed, under the alternative hypothesis, by repeated application of Bayes's theorem. All technical details for the computation of the conditional probability are included in a technical report that is available on our Web page.

Results

In this section, we discuss two power-calculation scenarios, additional offspring and missing parental information; both scenarios are highly relevant to association studies. Many diseases considered in association studies have late onset (e.g., Alzheimer disease), and so genotyping the parents is not feasible. However, it might be relatively easy to sample siblings.

Application I: Power When Both Parents Are Available and One Additional Unaffected Offspring Is Included

In this section, we compute the power of the FBAT for affected trios with one additional unaffected sibling ($m_{ij} = 2$) and examine the influence of weighting scheme on the power. We will also compare our results with those of Whittaker and Lewis (1998). Whittaker and Lewis (1998) suggested including the unaffected offspring by use of the coding $T_{ij} = y_{ij} - z$, where $z = K$ is the disease prevalence. Since K is not known, in general, we examine the power of FBAT as a function of z.

We use formula (5) to derive the joint conditional distribution of the marker scores x_{i1} and x_{i2}. All phenotypes are fixed by the ascertainment condition $\mathcal{A}$—that is, $\mathcal{A} = \{y_{i1} = 1, y_{i2} = 0\}$—and the parental genotypes are known. Hence, the minimal sufficient statistic of FBAT depends only on the parental genotypes—that is, $s_i = (p_{i1},p_{i2})$. The conditional marker distribution is given by

$$P(x_{i1},x_{i2} \mid y_{i1},y_{i2},s_i) = \frac{\prod_{j=1}^{2} P(y_{ij} \mid x_j)P(x_j \mid p_{i1},p_{i2})}{\prod_{j=1}^{2} \left\{\sum_{x} P(y_{ij} \mid x)P(x \mid p_{i1},p_{i2})\right\}} \ . \qquad (9)$$

Note that we assume independence of $Y_{ij} \mid X_j$ and of $X_j \mid (P_{i1},P_{i2})$ in (9). In principle, a more general model can be used that allows for nonindependence of the values of $Y_{ij} \mid X_j$, as in the work of Q. Yang, X. Xu, and N. M. Laird (unpublished data). Whereas $P(x \mid p_{i1},p_{i2})$ depends only on Mendelian transmission, $P(y_{ij} \mid \tilde{x})$ depends on the hypothesis and the underlying genetic model. Under the null hypothesis, it is given by (6) and under the alternative hypothesis by (7).

Under the assumption of an additive coding for the marker scores (i.e., $x_{ij} = 0,1,2$), there are $3^2 = 9$ possible combinations for $S_i = (p_{i1},p_{i2})$, which we will denote here by $S^{(1)}, \ldots, S^{(9)}$. Further, the number of observed $S^{(k)}$ is given by n_k (i.e., $n = \sum_{k=1}^{9} n_k$). Then, the scaling parameter ω for the conditional power formula (2) can be computed by

$$\omega = \frac{\sum_{k=1}^{9} n_k[\mathrm{Var}_0(X \mid y = 1,S^{(k)})(1-z)^2 + \mathrm{Var}_0(X \mid y = 0,S^{(k)})z^2]}{\sum_{k=1}^{9} n_k[\mathrm{Var}_{H_1}(X \mid y = 1,S^{(k)})(1-z)^2 + \mathrm{Var}_{H_1}(X \mid y = 0,S^{(k)})z^2]} \ ,$$

and the noncentrality parameter γ can be computed by

$$\gamma = \left\{\sum_{k=1}^{9} n_k[E_0(X \mid y = 1,S^{(k)}) - E_1(X \mid y = 1,S^{(k)})](1-z) - \sum_{k=1}^{9} n_k[E_0(X \mid y = 0,S^{(k)}) - E_1(X \mid y = 0,S^{(k)})]z\right\}^2$$

$$/\sum_{k=1}^{9} n_k[\mathrm{Var}_1(X \mid y = 1,S^{(k)})(1-z)^2 + \mathrm{Var}_1(X \mid y = 0,S^{(k)})z^2] \ .$$

Further, we use formula (8) to compute the distribution of $S^{(k)}$ conditional on the ascertainment condition $\mathcal{A}$, which is given here by $\mathcal{A} = \{Y_1 = 1, Y_2 = 0\}$. Having these conditional distributions, it is straightforward to compute the unconditional/expected power.

To compare our results with those of Whittaker and Lewis (1998), we assume a multiplicative model, $f_1 = \sqrt{f_0 f_2}$. Under the assumption that the alternative hypothesis would be close to the null hypothesis (low gene effect)—that is,

$$\sqrt{f_0/f_2} \approx 1 \ .$$

Whittaker and Lewis (1998) suggested that inclusion of unaffected siblings gives maximum power when the offset z is chosen to be the disease prevalence and that the power gained by inclusion of unaffected offspring does not outweigh the cost of the additional genotyping. For alternative hypotheses satisfying the assumption of low gene effects, our power calculations, as well as methods discussed elsewhere (Q. Yang, X. Xu, and N. M. Laird, unpublished data), confirm the finding of Whittaker and Lewis (1998).

However, when we consider scenarios like those discussed by Boehnke and Langefeld (1998) and Knapp (1999*a*), in which the authors assume that $\sqrt{f_0/f_2}$ is in the range 1.5–4, and in which unaffected siblings are included, we observe, for common diseases, a substantial gain in power over the standard TDT, which uses only affected. For $\sqrt{f_0/f_2} = 4$ these results are illustrated for a common and a rare disease in figure 1. Figure 1 strongly suggests that genotyping unaffected offspring can be worthwhile for common diseases, as shown elsewhere (Q. Yang, X. Xu, and N. M. Laird, unpublished data). Furthermore, it is important to note that, in any case, $z > 0$ is a better choice than the standard TDT. However, the power is relatively insensitive to the choice of the offset z in a limited range around the prevalence. When $\sqrt{f_0/f_2}$ is in the range 1.5–4, our analytical power calculations show that the optimal FBAT is obtained when the offset z is chosen to be greater than the disease prevalence, although the amount of power gained over use of z as "disease prevalence" is very small.

Power of FBATs When Both Parental Genotypes Are Missing

Since the standard TDT discussed by Spielman et al. (1993) requires the parental genotypes, a variety of extensions have been proposed when either one or both parents are missing: SDT by Horvath and Laird (1998), S-TDT by Spielman and Ewens (1998), RC-TDT by Knapp (1999*b*), and FBAT by Rabinowitz and Laird (2000). The advantage of the FBAT approach is its flexibility; it can handle scenarios with one missing parent, arbitrary numbers of offspring within a family, etc. We will therefore concentrate here on power calculations for FBATs. However, for many scenarios, RD-TDT and FBAT give similar results (Horvath et al. 2001).

For simplicity, we examine only scenarios where both parents are missing. We assume that we observe at least two offspring per family ($m_i \geqslant 2$). The sufficient statistic $S_i(\mathbf{x}_i)$ is given in Rabinowitz and Laird (2000) and can be used directly in formula (5) to derive the conditional marker distribution for FBATs under the null and alternative hypothesis.

Having these conditional distributions, the conditional marker mean and variance can be calculated under the null and alternative hypothesis (formula [5]). Then, the scaling parameter ω and the noncentrality parameter γ in the conditional power formula (2) can be computed as in our section on "Application I." In the "unconditioning" step, we assume that families are ascertained with at least one affected offspring—that is, $\mathcal{A} = \{Y_1 = 1\}$. The probability $p(\mathbf{y}_i, s_i \mid \mathcal{A})$ is obtained by application of formula (8).

We compute the power of GT and its dependence on the offset z for a variety of sampling plans assuming parental genotype data are always missing. We assume that the penetrances under the alternative model are given by a multiplicative model. Figure 2 shows the results of the asymptotic power calculations outlined above for common and rare disease scenarios. To study the effect of missing parental information on the power, we also give the power of GT for the same number of offspring when the parental genotypes are known.

As expected, the loss of power caused by missing parental information decreases with increasing family size. While the loss of power for two siblings with no parental information is relatively large compared to the power when both parents are given, the loss of power is moderate for the families with 3 siblings. In fact, for $K \geqslant 0.2$, the power of the three sibling design is indistinguishable from the "two-parent and two-offspring" design, and should be preferred as it requires less genotyping and has less sensitivity to z.

For all tests, choosing z to be the disease prevalence seems to be a reasonable choice. However, the choice of z has only a minor affect on the power of all tests as long as z is in a sensible range around the disease prevalence. For rare diseases, $z = 0$ is a reasonable choice, which corresponds to treating the phenotypes of the unaffected siblings as unknown. When trios with one affected offspring or two offspring without parental information are given, there is no dependence on the offset z. Although this observation is trivial for affected trios, it is unexpected for two offspring with no parental information. In this case, it can be explained by the definition of the sufficient statistic. Since the sufficient statistic conditions on the observed marker scores, the correlation between the two marker scores is always -1 when only two offspring without parental information

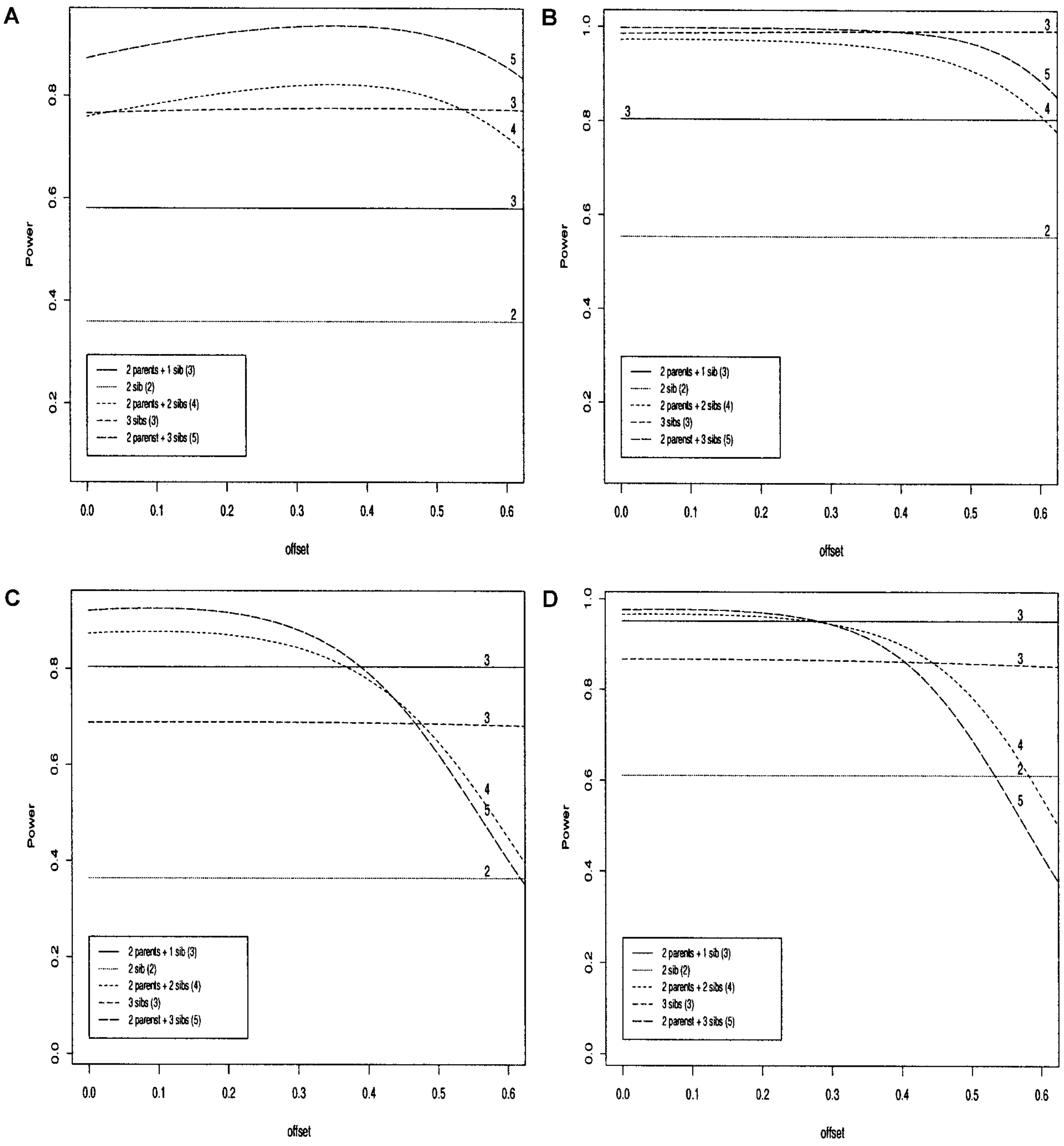

Figure 2 Power of FBAT tests for multiplicative models: Significance level $\alpha = 0.01$. The numbers shown above the graphs correspond to the numbers of genotyped subjects in each family. *a*, $n = 200, p = 0.2, K = 0.3$, AF = 0.2. *b*, $n = 100, p = 0.05, K = 0.2$, AF = 0.2. *c*, $n = 100, p = 0.05, K = 0.05$, AF = 0.2. *d*, $n = 200, p = 0.2, K = 0.05$, AF = 0.3.

are given. This simplifies the formulas for the scaling parameter ω and the noncentrality parameter γ, so that they do not depend upon the offset. One might also get the impression that the power does not depend on the offset for three offspring and no parental information. However, this impression is due to the selected offset range between 0.0 and 0.6 in figure 2. For offsets between 0.6 and 1.0, the power also depends on the offset choice for this family type.

Application to Study Design

An ancillary genetic study of bipolar disorder is being planned that builds on patients enrolled in a large ongoing clinical trial. For illustration, we assume that 1,003 families with one affected proband will participate in the study: 213 probands with both parents, 175 probands with one parent and one sibling, 175 probands with one parent and two siblings, 220 probands with one sibling, and 220 probands with two siblings. Because of cost and recruitment considerations, siblings will not be phenotyped. With low disease prevalence ($K \triangleq 0.01$), this should not have a substantial impact on the power, but this can be tested using our method. We compare power under the assumptions that siblings are phenotyped or not, for a fixed set of penetrance functions and a range of allele frequencies that give $K \triangleq 0.01$. For the case when the siblings are phenotyped, we set the offset to 0.01. Furthermore, we assume that the significance level is $\alpha = 0.00001$. The achieved power for a range of potential allele frequencies is shown in table 1 when the sibling are phenotyped or not. Table 1 shows that phenotyping additional offspring has virtually no effect on the achieved power and therefore is not worthwhile in this study.

Discussion

In this study, we have presented an approach to power calculations for FBATs. Our approach differs from the approach taken by Knapp (1999*a*) and its extension in the sense that it computes the expected power of the actual test statistic, whereas Knapp's approach gives the power of the expected statistic. Although the results of the two approaches do not differ greatly for the examples considered in the literature, the difference becomes relevant when the family size becomes larger, when parental information is missing, or when extensions to continuous traits are considered. For all these scenarios, the power has so far been assessed by simulation experiments. Our approach allows the computation of the unconditional/expected power for these scenarios and thereby becomes an important tool for the design of genetic studies (e.g., comparisons of sample designs in terms of genotyping and prevalence). For discordant sibships, we applied our approach to verify the results obtained by the approximation proposed by Whittaker and Lewis (1998). Although we found that offset choices close to the population prevalence are not always optimal, they seem to be a reasonable rule of thumb. Further, genotyping of unaffected offspring can be beneficial when the population prevalence is high.

Table 1

Power for Bipolar Disorder Study

	Power	
p	Additional Offspring Phenotyped	Additional Offspring Not Phenotyped
.5	.989	.986
.4	.997	.995
.3	.998	.997
.2	.994	.993
.1	.911	.902
.05	.488	.474
.01	.007	.007

NOTE.—Significance level $\alpha = 0.00001$, and penetrance function $f_{AA} = 0.03$, $f_{AB} = 0.02$, and $f_{BB} = 0.01$.

We discussed design issues and powerful offset choices for situations in which no parental information is available. For many scenarios, high power can be achieved with a reasonable sample size in the absence of parental information. In these situations, the influence of the offset on the power is negligible.

The methodology proposed here is fully general; hence, extensions to sampling designs and power calculations for multiallelic loci and continuous phenotypes are straightforward. We have implemented our approach to power calculations in a software package called "PBAT," which is available on our Web page. In addition to the scenarios discussed here, PBAT can also be used for power calculations for continuous traits and when marker locus and disease locus are not identical.

Acknowledgments

We would like to thank Dr. Ed Silverman for his helpful comments on an earlier version of the paper. This research was supported in part by grants MH59532 and HL66383 from the National Institutes of Health. The Childhood Asthma Management Program (CAMP) is supported by contracts from the National Heart, Lung and Blood Institute, including N01 HR16040. We acknowledge the CAMP investigators and research team for assistance in collection of the CAMP Genetics Ancillary Study data. Comments from three referees were very helpful in preparing this version of the paper.

Appendix A

Asymptotic Distribution of *GT* under the Alternative Hypothesis

We denote the vector containing all marker information by

$$\mathbf{X} = (X_{11}, X_{12}, \ldots, X_{nm_n})^t ,$$

and the corresponding vector of the coded trait information by $\mathbf{T} = (T_{11}, T_{12}, \dots, T_{nm_n})^t$. Then, the statistic GT can be written as

$$GT = \left\{\frac{\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]}{\sqrt{\mathbf{T}^t \mathrm{Var}_0(\mathbf{X})\mathbf{T}}}\right\}^2$$

$$= \frac{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}{\mathbf{T}^t \mathrm{Var}_0(\mathbf{X})\mathbf{T}} \left\{\frac{\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]}{\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}}\right\}^2$$

Under the alternative hypothesis, note that

$$E\left\{\frac{\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]}{\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}}\right\}^2 = \frac{\mathbf{T}^t[E_1(\mathbf{X}) - E_0(\mathbf{X})]}{\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}}$$

$$\mathrm{Var}\left(\frac{\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]}{\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}}\right) = 1$$

Furthermore,

$$\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]/\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}} = \sum\nolimits_{i,j} \frac{T_{ij}}{\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}}[X_{ij} - E_0(X_{ij})]$$

is a weighted sum of potentially dependent random variables. However, since we assume that the families are independent and that the family size is bounded, we can rewrite $\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]$ as a sum of independent subsums—that is,

$$\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})] = \sum_{i=1}^{n} Z_i \quad \text{with} \quad Z_i = \sum_j T_{ij}[X_{ij} - E_0(X_{ij})] \ , \tag{A1}$$

where Z_i is computed on the basis of the data of the ith family. The values of Z_i are therefore independent. We standardize the weighted sum (A2) by its variance $\mathrm{Var}\{\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]\} = \sum_i \mathrm{Var}(Z_i) = \mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}$ and can then apply standard asymptotic theory to $\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]/\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}$. The asymptotic distribution of $\mathbf{T}^t[\mathbf{X} - E_0(\mathbf{X})]/\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}$ under the alternative hypothesis is, therefore, given by a noncentral χ^2 distribution with 1 df and noncentrality parameter $\gamma = \{\mathbf{T}[\mathbf{E}_1(\mathbf{X}) - \mathbf{E}_0(\mathbf{X})]^t/\sqrt{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}\}^2$. Thus, under the alternative hypothesis,

$$\frac{\mathbf{T}^t \mathrm{Var}_1(\mathbf{X})\mathbf{T}}{\mathbf{T}^t \mathrm{Var}_0(\mathbf{X})\mathbf{T}} GT \sim \chi^2_{1,\gamma}$$

Appendix B

We refer to each combination of possible values for $\mathbf{y}_i$ and s_i as a "family type," denoted by the vector $\mathbf{q}_k = (\mathbf{y}, s), k = 1, \dots, FT$, where FT is the number of possible family types. If we define an n-dimensional unit vector by $1_n = (1, \dots, 1)^t$, it is easy to see that

$$\mathcal{P}_{GT} = \sum_{n_1,\dots,n_{FT} \in \mathbb{N}: n_1 + \dots + n_{FT} = n} \mathcal{P}_{GT} \mid_{(1_{n_1}\mathbf{q}_1, \dots, 1_{n_{FT}}\mathbf{q}_{FT})} p_{\mathrm{Multinomial}(\pi_1, \dots, \pi_{FT})}(n_1, \dots, n_{FT}) \tag{B1}$$

where $p_{\mathrm{Multinomial}(\pi_1, \dots, \pi_{FT})}(n_1, \dots, n_{FT})$ is the density of the multinomial distribution with probabilities $\pi_1, \dots, \pi_{FT}$. The $\pi_1, \dots, \pi_{FT}$ are defined by $\pi_k = P(\mathbf{q}_k \mid \mathcal{A}), k = 1 \dots, FT$. Therefore, it is always possible to compute the exact unconditional power of GT by numerical integration of (B1). In situations with many potential family types $\mathbf{q}_k$ (e.g., either families with many offspring or offspring with many traits), this numerical integration may be very time consuming and can be replaced either by Monte Carlo simulations or by MCMC methods. However, for the situations considered in this study, the numerical computation of the sum is feasible.

For many scenarios, the number of family types can be reduced. For example, when the phenotypes are fixed by the ascertainment condition (e.g., $y = 1$ in the work of Spielman et al. [1993] or $y_1 = 1$ and $y_2 = 0$ in the work of Whittaker and Lewis [1998]), and when the parental genotypes are observed, the family types are defined by the observed parental information. For simplicity of exposition, assume an additive marker coding. When the scaling parameter ω and the noncentrality parameter δ are computed for the conditional power formula (2), it is easy to see that $\mathbf{q} = (p_1,p_2) = (0,0),(0,2),(2,0)$ and $(2,2)$ are noninformative and do not contribute to the scaling parameter ω and the noncentrality parameter δ. Further, it is important to note that $\mathbf{q} = (0,1)$ and $\mathbf{q} = (1,0)$ make the same contribution to ω and δ. The same is true for $\mathbf{q} = (0,2)$ and $\mathbf{q} = (2,0)$. In this setup, therefore, it is possible to reduce the number of family types to four distinct types—(0,1), (1,1), (1,2) and "noninformative"—and to change the probabilities π_i appropriately. This reduction of the number of family types accelerates the computation of the unconditional power substantially.

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

FBAT Web page, http://www.biostat.harvard.edu/~fbat/default.html (for PBAT software package)

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

Camp NJ (1997) Genomewide transmission/disequilibrium testing—consideration of the genotypic relative risks at disease loci. Am J Hum Genet 61:1424–1430

Chen W-M, Deng H-W (2001) A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. Genet Epidemiol 21:53–67

Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–1897

Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype-phenotype associations. Eur J Hum Genet 9:301–306

Knapp M (1999*a*) A note on power approximations for the transmission/disequilibrium test. Am J Hum Genet 64: 1177–1185

Knapp M (1999*b*) Using exact *P* values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib transmission/disequilibrium test. Am J Hum Genet 65:1208–1210

Laird NM, Horvath S, Xu X (2000) Implementing a unified approach to family based tests of associaton. Genet Epidemiol Suppl 19:S36–S42

Lange C, Laird NM (2002) On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power and optimality considerations. Genet Epidemiol 23:1–16

Lange C, Silverman EK, Xu X, Weiss ST, Laird NM. A multivariate transmission disequilibrium test: FBAT-GEE. Biostatistics (in press)

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

Risch N (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856

Rabinowitz D, Laird NM (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50:211–223

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

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

Schulze TG, McMahon FJ (2002) Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines. Am J Med Genet 114:1–11

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

Whittaker JC, Lewis CM (1998) Power comparisons of the transmission/disequilibrium test and sib–transmission/disequilibrium-test statistics. Am J Hum Genet 65:578–580

Zhao H (2000) Family-based association studies. Stat Methods Med Res 9:563–587