Tests for Genetic Association Using Family Data

Mei-Chiung Shih and Alice S. Whittemore*

Department of Health Research and Policy, Stanford University, Stanford, California

We use likelihood-based score statistics to test for association between a disease and a diallelic polymorphism, based on data from arbitrary types of nuclear families. The Nonfounder statistic extends the transmission disequilibrium test (TDT) to accommodate affected and unaffected offspring, missing parental genotypes, phenotypes more general than qualitative traits, such as censored survival data and quantitative traits, and residual correlation of phenotypes within families. The Founder statistic compares observed or inferred parental genotypes to those expected in the general population. Here the genotypes of affected parents and those with many affected offspring are weighted more heavily than unaffected parents and those with few affected offspring. We illustrate the tests by applying them to data on a polymorphism of the SRSB2 gene in nuclear families with multiple cases of prostate cancer. We also use simulations to compare the power of these family-based statistics to that of the score statistic based on Cox's partial likelihood for censored survival data, and find that the family-based statistics have considerably more power when there are many untyped parents. The software program FGAP for computing test statistics is available at <http://www.stanford.edu/dept/HRP/epidemiology/FGAP>. *Genet. Epidemiol.* 22:128–145, 2002. © 2002 Wiley-Liss, Inc.

Key words: genetic association tests; transmission disequilibrium test (TDT); age at onset

INTRODUCTION

This paper concerns the use of data from family studies to test for association between a disease and genotypes at a given chromosomal locus. The objective is to evaluate whether the test locus is near a disease-causing polymorphism. Tests based on data from case-control studies are problematic because they may detect association that is due not to a nearby disease-caused polymorphism but rather to unrecognized ethnic admixture of the underlying population. This possibility has prompted interest in the use of family-based designs. By comparing genotypes of affected individuals to those of their unaffected siblings or to Mendelian expectation based on genotypes of their parents, these designs avoid this problem.

In previous work we and others [Schaid and Sommer, 1994; Clayton, 1999; Whittemore and Tu, 2000; Tu et al., 2000] proposed likelihood-based test statistics applicable to families. The tests are based on the score statistics derived from a specific likelihood for binary traits

Contract grant sponsor: NIH; Contract grant numbers: CA47448, CA49446.

*Correspondence to: Alice S. Whittemore, Department of Health Research and Policy, Stanford University, Stanford, CA 94305. E-mail: alicesw@stanford.edu

Received for publication: 18 May 2001; revision accepted 5 August 2001

© 2002 Wiley-Liss, Inc.

DOI 10.1002/gepi.0151

in families. The likelihood contribution for a family is the probability of its observed genotypes, given its genealogical structure and the observed phenotypes of its members. For nuclear families, the likelihood function factors into a term that represents the probability of parental genotypes, and a term representing the probability of the offspring genotypes given those of their parents. This factorization induces a decomposition of the score statistic into two components. The first, which we call the Founder statistic, is based on the observed or inferred parental genotypes. A parent's contribution to this statistic depends not only on his or her phenotype (if known) but also on the phenotypes of his or her offspring. Parents with many affected and few unaffected offspring contribute positive terms to the statistic, whereas parents with few affected offspring or many unaffected offspring contribute negative terms.

The second component of the total score statistic, the Nonfounder statistic, is based on the conditional distribution of the offspring genotypes, given the observed or inferred genotypes of their parents. This statistic, which generalizes the transmission-disequilibrium test (TDT) [Spielman et al., 1993] evaluates disequilibrium in the transmission of alleles from parents to affected and unaffected offspring, and it does not use data on parental phenotypes.

Here we focus on nuclear families in which parental genotypes may be missing, and we generalize these test statistics in three ways. First, we extend the underlying likelihood for binary disease outcomes to accommodate other disease phenotypes, such as censored times to failure and quantitative traits like blood pressure. Second, we extend the likelihood to accommodate within-family correlation in phenotypes, which is unrelated to genotypes at the test locus. Third, we relax the assumption that the null distribution of genotypes among the parents is known. This assumption was made in Whittemore and Tu [2000].

The paper is organized as follows. First we describe the family-based test statistics with the three generalizations described above. Then, for censored survival data, we present results from simulations comparing their power to that of score tests based on the partial likelihood function of Cox [1972]. We then apply both approaches to censored survival data in 126 nuclear families with multiple cases of prostate cancer. We conclude by discussing strengths and weaknesses of the genetically-based tests, and some possible generalizations.

TEST STATISTICS

We code an individual's genotype as $g = 0, 1$ or 2 , depending on the number of variant alleles he carries. In a nuclear family, we denote the vector of parental genotypes by $\mathbf{g} = (g_M, g_F)$, where g_M is the maternal genotype and g_F is the paternal one. To simplify the presentation, we shall assume Hardy-Weinberg equilibrium for the parental genotype frequencies, and random parental mating. Thus, we assume that, under the null hypothesis of no association between phenotypes and genotypes, the probability $P(\mathbf{g}; \gamma)$ of parental genotype frequencies depends only on the unknown frequency γ of the deleterious allele:

$$\begin{aligned} P(\mathbf{g}; \gamma) &= P(g_M; \gamma)P(g_F; \gamma) \\ &= \binom{2}{g_M} \gamma^{g_M} (1 - \gamma)^{2 - g_M} \binom{2}{g_F} \gamma^{g_F} (1 - \gamma)^{2 - g_F}. \end{aligned} \quad (1)$$

Let $P(\phi|g, \beta)$ denote the probability of an individual's observed phenotype ϕ , given his genotype g . Here ϕ can be an indicator for disease presence, a (possibly censored) time to disease occurrence, or the value of a quantitative trait, such as height or weight. The unknown parameter β is a measure of association between phenotypes and genotypes, with $\beta =$

0 representing the null hypothesis of no association. Let $\theta = (\beta, \gamma)$ denote the vector of unknown parameters.

We assume that, given an individual's genotype g at the locus in question, his phenotype is independent of both the genotypes and the phenotypes of his relatives. This assumption does not allow for familial phenotype correlations due to shared alleles at other loci or to shared exogenous exposures. However, as we later show using simulations, misspecification of these correlations will not affect the tests' type I error probabilities, although it could impact negatively on their power. We shall describe how one might accommodate such correlations by including family-specific random effects.

We observed genotypes for N unrelated nuclear families, who may have been selected because of their phenotypes. The number of offspring and the phenotypes in a family may vary from family to family. The likelihood contribution $L_i(\theta)$ for the i^{th} family is the conditional probability of the family's observed genotypes, given the number of its offspring and given the phenotypes of its members. See Appendix A for explicit likelihood expressions. The score vector of derivatives of $\log L_i(\theta)$ with respect to θ is

$$\mathbf{u}_i(\theta) = \begin{pmatrix} u_{i1}(\theta) \\ u_{i2}(\theta) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \beta} \log L_i(\theta) \\ \frac{\partial}{\partial \gamma} \log L_i(\theta) \end{pmatrix}.$$

The score has expectation zero and covariance matrix $E[\mathbf{u}_i(\theta)\mathbf{u}_i(\theta)^T]$. Under certain conditions a central limit theorem applies to the Total score statistic

$$\mathbf{U}(\theta) = \sum_{i=1}^N \mathbf{u}_i(\theta).$$

Consequently $\mathbf{U}(\theta)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{J}(\theta) = \sum_i E[\mathbf{u}_i(\theta)\mathbf{u}_i(\theta)^T].$$

(We speak of $\mathbf{U}(\theta)$ as being asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{J}(\theta)$, though it is the standardized version of $\mathbf{U}(\theta)$ that converges.) Appendix A provides expressions for the variances $E[\mathbf{u}_i(\theta)\mathbf{u}_i(\theta)^T]$.

We shall use the asymptotic distribution of the score to test the null hypothesis $\beta = 0$ after maximizing the log-likelihood with respect to γ . Specifically, the test statistics are based on the first component $U_1(\theta)$ of $\mathbf{U}(\theta)$, evaluated at $\beta = 0$ and $\gamma = \hat{\gamma}$, where $\hat{\gamma} = \hat{\gamma}(0)$ maximizes $\sum_i \log P(\mathbf{g}_i; \gamma)$. The asymptotic conditional distribution of $U_1(\theta)$ given $U_2(\theta) = 0$ [that is, given $\gamma = \hat{\gamma}(0)$] is, from the asymptotic distribution of $\mathbf{U}(\theta)$ and multivariate normal theory, normal with mean 0 and variance

$$J_{11} - J_{12}^2 J_{22}^{-1},$$

where J_{ij} is the ij^{th} entry of $\mathbf{J}(\theta)$ evaluated at $\theta = (0, \hat{\gamma})$. The *Total statistic* has the form

$$T^2 = \frac{U_1^2}{J_{11} - J_{12}^2 J_{22}^{-1}}, \quad (2)$$

with all quantities evaluated at $\theta = (0, \hat{\gamma})$. Under the null hypothesis and assuming correct specification of the parental genotype frequencies, the asymptotic distribution of each test statistic is that of a chi-square variable with one degree of freedom.

When the parental genotype vector $\mathbf{g}$ for a family is known, the likelihood contribution for that family can be expressed simply in terms of the null probability of the family genotypes and the conditional probability of the family phenotypes, given the family genotypes. However for many diseases of late onset, some or all parental genotypes are unknown or only partially observed. To handle this possibility, we take the likelihood contribution for the i^{th} family to be $L_i(\theta) = P(\mathbf{x}_i | \Phi_i; \theta)$, where $\mathbf{x}_i$ and Φ_i denote, respectively, the observed genotype data and the observed phenotypes for all family members. We assume that $\mathbf{x}_i$ includes the vector $\mathbf{h}_i = (h_{i1}, \dots, h_{in_i})$ of genotypes for the n_i offspring, but it may lack one or both of the parental genotypes $\mathbf{g}_i = (g_{iM}, g_{iP})$. The probability of $\mathbf{x}_i$ given the family's phenotype Φ_i is

$$P(\mathbf{x}_i | \Phi_i, \theta) = \sum_{\mathbf{g}} L_i(\theta; \mathbf{g}, \mathbf{h}_i). \quad (3)$$

Here $L_i(\theta; \mathbf{g}, \mathbf{h}_i)$ is given by equation (22) in Appendix A. Differentiating the log of (3) with respect to β and γ gives the score for the family as

$$\mathbf{u}_i(\mathbf{x}_i; \theta) = \begin{pmatrix} u_{i1}(\mathbf{x}_i; \theta) \\ u_{i2}(\mathbf{x}_i; \theta) \end{pmatrix} = \begin{pmatrix} \sum_{\mathbf{g}} P(\mathbf{g} | \mathbf{x}_i; \theta) u_{i1}(\theta; \mathbf{g}) \\ \sum_{\mathbf{g}} P(\mathbf{g} | \mathbf{x}_i; \theta) u_{i2}(\theta; \mathbf{g}) \end{pmatrix}. \quad (4)$$

The terms $u_{i1}(\theta; \mathbf{g})$ and $u_{i2}(\theta; \mathbf{g})$ denote the scores for the family if the parental genotype were known to be $\mathbf{g}$, and $P(\mathbf{g} | \mathbf{x}_i; \theta)$ is the null conditional probability of parental genotype $\mathbf{g}$ given the family's observed genotype $\mathbf{x}_i$.

Decomposition of the Score

Clayton [1999] and Whittemore and Tu [2000] noted a factorization of the likelihood contributions from families with known parental genotypes. This factorization induces an additive decomposition of the score $\mathbf{U}(\theta)$, which leads to two types of score statistics, the *Founder statistic* and the *Nonfounder statistic*. The *Founder statistic* evaluates differences between observed or inferred parental genotypes and their null expectations, while the *Nonfounder statistic* evaluates differences between the observed offspring genotypes and those expected, given the available information on the genotypes of their parents.

To describe these statistics, we factor the likelihood function $L_i(\theta; \mathbf{g}_i, \mathbf{h}_i)$ for a family with known parental genotype vector $\mathbf{g}_i$ as the product of two terms:

$$L_i(\theta; \mathbf{g}_i, \mathbf{h}_i) = L_i^F(\theta; \mathbf{g}_i) L_i^{NF}(\beta; \mathbf{h}_i | \mathbf{g}_i). \quad (5)$$

The Founder term $L_i^F(\theta; \mathbf{g}_i)$ is proportional to the conditional probability of the parental genotypes, given the family phenotype. The Nonfounder term $L_i^{NF}(\beta; \mathbf{h}_i | \mathbf{g}_i)$ is proportional to the conditional probability of the offspring genotypes, given the parental genotypes and the family phenotype. Both terms are described in Appendix A. The factorization (5) induces a decomposition of the family-specific score $\mathbf{u}(\theta; \mathbf{g}_i)$ as

$$\begin{aligned}\mathbf{u}(\theta; \mathbf{g}_i) &= \begin{pmatrix} \frac{\partial}{\partial \beta} \log L_i^F(\theta; \mathbf{g}_i) \\ \frac{\partial}{\partial \gamma} \log L_i^F(\theta; \mathbf{g}_i) \end{pmatrix} + \begin{pmatrix} \frac{\partial}{\partial \beta} \log L_i^{NF}(\beta; \mathbf{h}_i | \mathbf{g}_i) \\ 0 \end{pmatrix} \\ &\equiv \mathbf{u}_i^F(\theta; \mathbf{g}_i) + \mathbf{u}_i^{NF}(\theta; \mathbf{g}_i).\end{aligned}\quad (6)$$

When the parental genotypes are uncertain, combining the decomposition (6) with equation (4) yields an analogous decomposition for the family-specific score $\mathbf{u}_i(\mathbf{x}_i; \theta)$:

$$\mathbf{u}_i(\mathbf{x}_i; \theta) = \begin{pmatrix} u_{i1}^F(\mathbf{x}_i; \theta) \\ u_{i2}^F(\mathbf{x}_i; \theta) \end{pmatrix} + \begin{pmatrix} u_{i1}^{NF}(\mathbf{x}_i; \theta) \\ 0 \end{pmatrix}, \quad (7)$$

where

$$u_{ij}^X(\mathbf{x}_i; \theta) = \sum_{\mathbf{g}} P(\mathbf{g} | \mathbf{x}_i; \theta) u_{ij}^X(\theta; \mathbf{g}), \quad j = 1, 2, \quad X = F, NF. \quad (8)$$

Summing equation (7) over families gives

$$\mathbf{U}(\theta) = \sum_i \begin{pmatrix} u_{i1}^F(\mathbf{x}_i; \theta) \\ u_{i2}^F(\mathbf{x}_i; \theta) \end{pmatrix} + \sum_i \begin{pmatrix} u_{i1}^{NF}(\mathbf{x}_i; \theta) \\ 0 \end{pmatrix} \equiv U^F(\theta) + U^{NF}(\theta).$$

The *Founder statistic* is

$$T_F^2 = \frac{(U_1^F)^2}{J_{11}^F - (J_{12}^F)^2 (J_{22}^F)^{-1}},$$

where the entries J_{ij}^F of the covariance matrix $\mathbf{J}^F$ are given in Appendix A. T_F^2 is the standardized difference between the observed or inferred parental genotypes in the families and those expected under the null hypothesis of no association.

Both the Total statistic T^2 and the Founder statistic T_F^2 require estimating γ via a model $P(\mathbf{g}; \gamma)$ for the null parental genotype frequencies. To reduce bias due to misspecification of this model, Clayton proposed basing inferences on a “partial score”

$$\mathbf{U}^*(\theta) = \begin{pmatrix} U_1^{NF}(\theta) \\ U_2^F(\theta) \end{pmatrix}.$$

This approach also was considered by Whittemore and Tu [2000]. The partial score ignores the contribution from u_{i1}^F because it is correlated with u_{i2}^F , and thus misspecification of parental genotype frequencies could bias u_{i1}^F through this correlation. The corresponding score statistic, called the *Nonfounder statistic*, is

$$T_{NF}^2 = \frac{(U_1^*)^2}{J_{11}^* - (J_{12}^*)^2 (J_{22}^*)^{-1}}.$$

The entries J_{ij}^* of the covariance matrix $\mathbf{J}^*$ are described in Appendix A. The Nonfounder statistic evaluates the sum over families of the differences between the observed offspring

genotypes and those expected, given the available information on the parental genotypes. When all parental genotypes are known, T_{NF}^2 does not involve the model $P(\mathbf{g}; \gamma)$. When some parental genotypes are missing, $P(\mathbf{g}; \gamma)$ must be specified in order to infer the null distribution of possible parental genotypes. However, this reliance on $P(\mathbf{g}; \gamma)$ is more robust than that of T^2 and T_F^2 . This is because, although misspecifying $P(\mathbf{g}; \gamma)$ may decrease the power of the test based on T_{NF}^2 , it seems less likely to produce spurious evidence for association, as it could with T^2 and T_F^2 .

Computing the Score Statistics

Explicit expressions for the test statistics require computing, for each of the N families with uncertain parental genotypes, the conditional probability of the parental genotypes $\mathbf{g}$, given the family's observed genotype data, under the null hypothesis:

$$P(\mathbf{g} | \mathbf{x}_i, \hat{\gamma}) = \frac{P(\mathbf{g} | \hat{\gamma})P(\mathbf{x}_i | \mathbf{g})}{\sum_{\mathbf{g}'} P(\mathbf{g}' | \hat{\gamma})P(\mathbf{x}_i | \mathbf{g}')} , \quad i = 1, \dots, K.$$

We also must compute the components $u_{i1}^F(\theta; \mathbf{g})$, $u_{i1}^{NF}(\theta; \mathbf{g})$, and $u_{i2}^F(\theta; \mathbf{g})$ of equation (8). Computing $u_{i2}^F(\theta; \mathbf{g})$ is straightforward: from equation (1),

$$u_{i2}^F(\theta; \mathbf{g}) = \frac{d}{d\gamma} \log P(\mathbf{g}; \gamma) = \frac{g_M + g_F - 4\gamma}{\gamma(1 - \gamma)}. \quad (9)$$

Substituting (9) into (8) gives

$$u_{i2}^F(\mathbf{x}_i; \theta) = \frac{1}{\gamma(1 - \gamma)} \left[\sum_{\mathbf{g}} P(\mathbf{g} | \mathbf{x}_i; \gamma) (g_M + g_F) - 4\gamma \right].$$

To compute $u_{i1}^F(\theta; \mathbf{g})$ and $u_{i1}^{NF}(\theta; \mathbf{g})$, we need to express them in terms of the β -derivatives of the individual phenotype probabilities $P(\phi | g, \beta)$. We shall consider three kinds of phenotype: a dichotomous indicator for disease occurrence; a censored time to failure; and a quantitative trait measurement. For all three phenotypes, we model the genotype effects on phenotype by replacing g by a prespecified function $z(g)$, called the *carrier status* associated with g . For a dominant model, $z(2) = z(1) = 1$ and $z(0) = 0$. For a recessive model, $z(2) = 1$ and $z(1) = z(0) = 0$, and for an additive model, $z(g) = g$. The null expectation of $z(g)$ in the parental population is

$$f_z(\gamma) = \begin{cases} \gamma(2 - \gamma) & \text{for a dominant model,} \\ \gamma^2 & \text{for a recessive model,} \\ 2\gamma & \text{for an additive model.} \end{cases}$$

Appendix B shows that

$$u_{i1}^F(\theta, \mathbf{g}) = \sum_{j \in \{M, F\}} a(\phi_{ij}) [z(g_{ij}) - f_z(\gamma)] + \sum_{j=1}^{n_i} a(\phi_{ij}) \{E[z(h) | \mathbf{g}] - f_z(\gamma)\}, \quad (10)$$

and

$$u_{i1}^{NF}(\theta, \mathbf{g}) = \sum_{j=1}^{n_i} a(\phi_{ij}) \{z(h_{ij}) - E[z(h) | \mathbf{g}]\}. \quad (11)$$

Here $a(\phi)$ is defined by the relation

$$\frac{d}{d\beta} \log P(\phi | g; \beta)_{\beta=0} = a(\phi) z(g). \quad (12)$$

When there are missing parental genotypes, we replace the unobserved $z(g_{iM})$ and/or $z(g_{iF})$ in expressions (10) and (11) with its expected value given the observed family genotype $\mathbf{x}_i$, and replace $E[z(h) | \mathbf{g}]$ with

$$\tilde{E}[z(h) | \mathbf{x}_i] = \sum_{\mathbf{g}} P(\mathbf{g} | \mathbf{x}_i) E[z(h) | \mathbf{g}].$$

$\tilde{E}[z(h) | \mathbf{x}_i]$ is the null expectation of an offspring's carrier status, conditional on the inferred distribution of the genotypes of his parents.

Expressions (10) and (11) show that each family-specific score is a weighted sum over family members of their observed minus expected genotypes. Each family member's contribution to this sum is weighted by the quantity $a(\phi)$. We now show that, for each of the three kinds of phenotype, $a(\phi)$ is the difference between an individual's observed phenotype and its null expectation. Therefore, the score statistics can be interpreted as measuring correlations between genotypes and phenotypes.

The Weights $a(\phi)$ for Different Phenotypes

For qualitative traits, the phenotype ϕ is an indicator for disease occurrence, and we take

$$P(\phi = 1 | g; \beta) = \pi[\rho_0 + \beta z(g)].$$

Here the baseline penetrance parameter ρ_0 is assumed known, and $\pi(x) = e^x / (1 + e^x)$ is the logistic function. Then

$$P(\phi | g; \beta) = e^{\phi[\rho_0 + \beta z(g)]} \left\{ 1 + e^{[\rho_0 + \beta z(g)]} \right\}^{-1}.$$

Differentiating $\log[P(\phi = 1 | g; \beta)]$ with respect to β and evaluating at $\beta = 0$ gives relation (12) with

$$a(\phi) = \phi - \pi(\rho_0), \quad (13)$$

which is the difference between the individual's observed phenotype and its null expectation. For such traits, the Nonfounder statistic is just the TDT (generalized to allow missing parental genotypes and unaffected offspring).

For censored times to failure, the phenotype $\phi = (t, \varepsilon)$, where t is the minimum of a person's time to failure and an independent time to censoring, and $\varepsilon = I(t = \text{time to failure})$. (We use the notion $I(\cdot)$ to represent an indicator that has value of 1 or 0, according to whether an event does or does not occur.) We assume that the distribution of a person's censoring times does not depend on his/her genotype, or on the parameters β and γ . We also assume a proportional hazards model for the effect of genotype on the failure hazard $\lambda(t; g)$:

$$\lambda(t; g) = \exp[\rho_0(t) + \beta z(g)], \quad (14)$$

where the log baseline hazard rate $\rho_0(t)$ is known. Then

$$P(\phi | g; \beta) \propto \exp\{[\rho_0(t) + \beta z(g)]\varepsilon - e^{\beta z(g)} \Lambda_0(t)\},$$

where $\Lambda_0(t) = \int_0^t e^{\rho_0(s)} ds$ is the baseline cumulative hazard function and the proportionality constant depends on the unknown censoring distribution. Differentiating $\log[P(\phi | g; \beta)]$ with respect to β and evaluating at $\beta = 0$ gives relation (12) with

$$a(\phi) = \varepsilon - \Lambda_0(t), \quad (15)$$

which again can be interpreted as the difference between an individual's observed and expected phenotype.

For quantitative traits, we assume that the distribution of the phenotype ϕ given genotype g is Gaussian with mean $\mu = \rho_0 + \beta z(g)$, where ρ_0 is known, and with known variance σ^2 . Without loss of generality we take $\sigma^2 = 1$. Then

$$P(\phi | g; \beta) \propto \exp\left\{-\frac{1}{2}(\phi - \rho_0 - \beta z(g))^2\right\},$$

and again, relation (12) holds with

$$a(\phi) = \phi - \rho_0 \quad (16)$$

representing the difference between the individual's observed phenotype and its null expectation.

Accommodating Residual Family Phenotype Correlation

Within-family correlation under the null hypothesis can be incorporated in the disease distribution or hazard by replacing the i^{th} family's baseline parameter or function ρ_0 with a family-specific $\rho_0 + \rho_i$. We assume that the ρ_i are independent, identically distributed Gaussian variables with mean 0 and variance τ^2 . We show in Appendix B that this extension produces family scores with the same form as (10) and (11), but with different null expected phenotypes in the weights $a(\phi)$. Specifically, an individual's expected phenotype depends on the size and phenotype of his family relative to those of other families. For the qualitative phenotype, the weight for the j^{th} individual in the i^{th} family is

$$a(\phi_{ij}) = \phi_{ij} - E_{\beta=0}[\pi(\rho_0 + \rho_i) | \Phi_i], \quad (17)$$

where $E[\cdot]$ denotes expectation with respect to the distribution for ρ_i at $\beta = 0$, conditional on the family structure and phenotype Φ_i . For censored failure data the weight is

$$a(\phi_{ij}) = \phi_{ij} - \Lambda_0(t_{ij})\eta_i \text{ where } \eta_i = E_{\beta=0}[e^{\rho_i} | \Phi_i]. \quad (18)$$

The quantitative phenotype is useful to illustrate this extension because the expected phenotype is available in closed form. In this case the weight for the j^{th} individual in the i^{th} family is

$$a(\phi_{ij}) = \phi_{ij} - [v_i \bar{\phi}_i + (1 - v_i)\rho_0]. \quad (19)$$

Here $\bar{\phi}_i$ is the mean of the phenotypes for the $n_i + 2$ family members, and

$$v_i = \frac{\tau^2}{\tau^2 + \sigma^2 / (n_i + 2)}.$$

Thus, when the inter-family variance component τ^2 is large compared to the total variance $\sigma^2 = 1$, the null expectation in the weights a is approximately the family-specific mean phenotype. In contrast when τ^2 is small relative to σ^2 , the null expectation is approximately ρ_0 , and all families receive roughly equal weights. In general, when the inter-family variance component $\tau^2 = 0$, the expected phenotypes reduce to the expressions (13), (15), and (16) derived in the absence of residual phenotype correlation within families.

COMPARISON WITH TESTS BASED ON COX'S PARTIAL LIKELIHOOD

We used simulations to compare the performance of the family-based score tests for censored survival data with those of score tests based on Cox's partial likelihood function [Cox, 1972]. Here we assumed the proportional hazard model (14), and stratified the partial likelihood score tests on families to take into account potentially different baseline hazard functions across families. More specifically, the partial likelihood score for a family when both parents have known genotypes and phenotypes is

$$\sum_{j \in \{M, F\} \cup \{1, \dots, n_i\}} \varepsilon_{ij} \left\{ z(g_{ij}) - \frac{\sum_{k \in R_{ij}} z(g_{ik})}{|R_{ij}|} \right\}. \quad (20)$$

Here R_{ij} is the risk set in family i at time t_{ij} , and $|R_{ij}|$ is the size of this risk set. In the case that one or both parents are untyped, we exclude untyped parents in the summation as well as R_{ij} . If parental data are excluded as in the TDT, the scores becomes

$$\sum_{j=1}^{n_i} \varepsilon_{ij} \left\{ z(g_{ij}) - \frac{\sum_{k \in \tilde{R}_{ij}} z(g_{ik})}{|\tilde{R}_{ij}|} \right\}, \quad (21)$$

where $\tilde{R}_{ij}$ is the risk set in family i at time t_{ij} with parents excluded. It can be seen that the partial likelihood function accommodates the failure data of typed parents, but not of untyped parents. Also it does not use information on the genetic relationship between parents and offspring. The resulting score test may thus perform poorly when the family sizes are too small (since the risk sets will be small) or when the allele of interest is rare (since many families will not contribute to the score because all their members have the same genotype). We describe next how our statistics deal with these problems.

Notice that the family-based scores (10) and (11) differ from the partial likelihood scores (20) and (21) in the null expectations of the $z(g_{ij})$ and in the weights attached to each individual's contribution. For the partial likelihood scores, the null expectation for the carrier status $z(g_{ij})$ of an individual is the mean $z(g_{ik})$ averaged over all family members at risk of disease at his or her same age. For the family-based scores, in contrast, this "empirical" average is replaced by the null expectation of $z(g_{ij})$ based on an assumed genetic model. The weight for an individual's contribution to a partial likelihood score is his/her disease indicator ε_{ij} , so that diseased individuals receive a weight of $\varepsilon_{ij} = 1$, and censored individuals receive a weight of $\varepsilon_{ij} = 0$. In contrast, the weight for an individual in the family-based scores

is $a(\phi_{ij})$ of (15). The contributions of diseased individuals are weighted positively, while those of censored individuals are weighted negatively. Moreover individuals with early-onset disease receive more positive weights than do those with late-onset disease, and older censored individuals receive more negative weights than do those who are censored early. Therefore, one might expect the family-based statistics to be more powerful than those based on Cox's partial likelihood when the allele of interest is associated with early disease onset.

The simulations support this conjecture. Each simulated dataset consists of censored times to disease occurrence and genotypes from 100 nuclear families, each with three offspring. These families were selected to have at least two affected family members, at least one of which was an offspring. The parental genotypes were generated under random mating and Hardy-Weinberg frequencies, with deleterious allele frequency of $\gamma = 0.1$. Given the parental genotypes, offspring genotypes were generated according to Mendelian probabilities. The (uncensored) times X to disease were generated according to the proportional hazards model (14) with an additive genetic model ($z(g) = g$) and random effects, as described in the preceding section. We chose the baseline hazard at $\rho_i = 0$ to be the Weibull hazard $\lambda_0(t) = (3.28 \times 10^{-8})(t - 30)_+^3$, where $x_+ = \max(x, 0)$. This corresponds to a 95% survival probability at age 80 years. We considered three values of the association parameter β : $\beta = 0$ (no association), $\beta = 0.405 = \ln 1.5$ (50% increase in baseline hazard for heterozygotes, i.e., individuals with $g = 1$), and $\beta = 0.693 = \ln 2$ (100% increase in baseline hazard for heterozygotes). The corresponding survival probabilities at age 80 years are 95%, 89%, and 81%. The censoring times C were generated according to a Weibull hazard $\mu(t) = (2.4 \times 10^{-5})(t - 40)_+^2$. For $\beta = 0.405$ and $\rho = 0$, about 90, 85, and 80% of failures are censored for individuals carrying zero, one and two deleterious alleles, respectively. The observed phenotype consists of the censored time to disease occurrence $T = \min(X, C)$ and a censoring indicator $\varepsilon = I[T = X]$. The results for type 1 error and power are based on 5,000 and 3,000 simulated datasets, respectively. We also used the true additive model $z(g) = g$ in the analysis.

Table I compares the type I error and power of the Total, Nonfounder and Founder statistics with those of Cox partial likelihood score tests. When there are no missing parental genotypes, the Nonfounder statistic performs comparably with the partial likelihood score test, while both Total and Founder statistics have much higher power for detecting association. When 80% of parental genotypes are missing, the partial likelihood score test suffers from substantial power loss. In contrast, the Total, Nonfounder, and Founder statistics all maintain good power.

Of course, we might expect power advantage for the genetically-based score statistics relative to power for the partial likelihood score statistics, since the former require assumptions about the baseline hazard and about the distribution of parental genotypes that are avoided by the partial likelihood. In particular, misspecification of the baseline hazard leads to suboptimal weights for the genotype contributions of family members, and thus to reduced power. We evaluated the extent of such power loss when the baseline hazard is misspecified in certain ways. Table II shows the impact of misspecification of baseline hazard on Type I error and power in the absence of missing parental genotypes and within-family phenotype correlation. Here the shape parameter of the baseline Weibull hazard is assumed to be the true value $r = 3$ or is misspecified to be $r = 0$ or $r = 3.3$. These parameter choices imply that the disease-free probability at age 80 years, which has true value of 0.95, is misspecified to be either close to 1 or 0.86. As expected, misspecification of the baseline hazard does not affect the partial likelihood score test since it does not use this information. Also as expected,

TABLE I. Type I Error and Power of Score Tests for No Association ($H_0: \beta = 0$) at Significance Level 0.05, with 0 or 80% Parental Genotypes Missing, and in the Absence ($\tau = 0$) or Presence ($\tau = 0.35$) of Within-Family Phenotype Correlation Due to an $N(0, \tau^2)$ Random Effect*

τ	% missing parental genotypes	β	P (reject H_0)				
			Cox (Φ_p included) ^a	Cox (Φ_p excluded)	Total	Nonfounder	Founder
0	0	0	0.050	0.053	0.052	0.048	0.050
		0.405	0.352	0.213	0.843	0.348	0.780
		0.693	0.874	0.476	1.000	0.783	0.998
	80	0	0.054	0.052	0.049	0.050	0.049
		0.405	0.242	0.203	0.823	0.426	0.718
		0.693	0.504	0.388	0.998	0.833	0.988
0.35	0	0	0.053	0.048	0.048	0.049	0.049
		0.405	0.359	0.200	0.885	0.413	0.813
		0.693	0.790	0.437	1.000	0.830	0.996
	80	0	0.051	0.052	0.053	0.052	0.052
		0.405	0.233	0.197	0.845	0.481	0.786
		0.693	0.536	0.459	0.998	0.918	0.994

*Based on 5,000 replications for Type I error and 3,000 replications for power.

^a Φ_p denotes parental phenotypes.

misspecification of the baseline hazard does not affect the Type I error of the genetically-based statistics, but it does decrease their power. However the power loss is relatively small.

Table III shows the impact of misspecification of within-family phenotype correlation on Type I error and power in the absence of missing parental genotypes, but in the presence of within-family phenotype correlation. Here the within-family phenotype correlation is induced by an $N(0, \tau^2)$ random effect, and we have misspecified the true value $\tau = 0.35$ to be $\tau = 0$. With $\tau = 0.35$, the random effect allows the baseline hazard for a family to vary from half to twice the value of the “overall population” baseline hazard λ_0 . It can be seen from Table III that such misspecification does not affect type I error, and for this level of correlation, the misspecification has little impact on power.

In conclusion, at least for the parameter ranges we have considered, misspecification of

TABLE II. Impact of Misspecification of Baseline Hazard Function on Type I Error and Power of Score Tests for No Association ($H_0: \beta = 0$) at Significance Level 0.05 in the Absence of Missing Parental Genotypes and Within-Family Phenotype correlation*

β	r	P (reject H_0)				
		Cox (Φ_p included) ^a	Cox (Φ_p excluded)	Total	Nonfounder	Founder
0	3 (true)	0.053	0.048	0.048	0.049	0.049
	0	0.057	0.055	0.051	0.057	0.054
	3.3	0.046	0.045	0.050	0.057	0.056
0.405	3 (true)	0.359	0.200	0.885	0.413	0.813
	0	0.366	0.208	0.861	0.422	0.776
	3.3	0.381	0.178	0.634	0.339	0.499
0.693	3 (true)	0.790	0.437	1.000	0.830	0.996
	0	0.804	0.445	0.998	0.809	0.992
	3.3	0.790	0.412	0.975	0.715	0.892

*Based on 5,000 replications for Type I error and 3,000 replications for power. Here the shape parameter r of baseline Weibull hazard is assumed to be the true value 3 or is misspecified to be 0 or 3.3.^a Φ_p denotes parental phenotypes.

TABLE III. Impact of Misspecification of Within-Family Correlations on Type I Error and Power of Score Tests for No Association ($H_0: \beta = 0$) at Significance Level 0.05 in the Absence of Missing Parental Genotypes*

β	τ	P (reject H_0)				
		Cox (Φ_p included) ^a	Cox (Φ_p excluded)	Total	Nonfounder	Founder
0	0.35 (true)	0.050	0.053	0.052	0.048	0.050
	0	0.053	0.047	0.053	0.048	0.056
0.405	0.35 (true)	0.352	0.213	0.843	0.348	0.780
	0	0.382	0.210	0.857	0.394	0.763
0.693	0.35 (true)	0.874	0.476	1.000	0.783	0.998
	0	0.821	0.459	0.999	0.828	0.995

*Based on 5,000 replications for Type I error and 3,000 replications for power. Here the standard deviation τ of the family-specific random effect is assumed to be the true value $\tau = 0.35$ or is misspecified to be $\tau = 0$.

^a Φ_p denotes parental phenotypes.

the baseline hazard and ignoring residual phenotype correlation have little adverse effect on the power of the genetically-based score statistics.

APPLICATION TO FAMILIAL PROSTATE CANCER

We illustrate the test statistics by applying them to genotypes from 126 nuclear families ascertained because of multiple cases of prostate cancer. The genotypes give the number of T (vs. A) alleles at a diallelic polymorphism of the steroid 5α -reductase gene (SRD5A2). The A/T polymorphism reflects a missense mutation called “A49T” that replaces alanine (A) at codon 49 with threonine (T). The phenotypes give the censored ages of prostate cancer occurrence and censoring indicators. These data are taken from an unpublished study (Hsieh et al.). Tu et al. [2000] analyzed the data using the qualitative phenotype (affected/unaffected), and found a statistically significant association between prostate cancer risk and homozygosity for the TT allele. Here, we re-analyzed these data using censored age at prostate cancer occurrence as phenotype.

Because prostate cancer is a late-onset disease (median age at diagnosis is 71 years), genotypes were available for only six fathers and eight mothers. The T allele was found in 11/126 families. This subset of families includes 24 typed affected sons (with genotype frequencies of 6 TT, 10 AT, and 8 AA), and 7 typed unaffected sons (with frequencies 0 TT, 3 AT, and 4 AA). None of the six typed fathers carried the T allele. Female offspring with known genotypes were included in the analysis to contribute information about parental genotypes, when these were incomplete.

We obtain a maximum likelihood estimate of 0.0356 for the T allele frequency, with 95% confidence interval (0.0206, 0.0563), under the null hypothesis and the assumptions of Hardy-Weinberg parental genotype frequencies and random mating of parents with respect to the genotypes of the A/T polymorphism. Table IV summarizes the values of the Total, Founder, and Nonfounder statistics and their one-sided P values. The score statistics based on Cox’s partial likelihood including parental phenotypes are also presented for comparison. The results based on Cox’s partial likelihood excluding parental phenotypes are the same as those for including parental phenotypes, and are therefore not presented in Table IV. Here the baseline hazard function was assumed known and equal to the maximum likelihood estimate of the population hazard function using prostate cancer incidence rates from the SEER program of the US National Cancer Institute. It has the form of a Weibull hazard, with certain constants

TABLE IV. Score Statistics for Association Between Prostate Cancer and the T Allele of the A49T Polymorphism of the SRD5A2 Gene in 126 Nuclear Families*

τ^2	Model	Score statistic (1-sided <i>P</i> -value)			
		Cox ^a	Total	Nonfounder	Founder
0	Dominant	1.100 (0.136)	-0.821 (0.794)	-0.106 (0.542)	-1.048 (0.853)
	Additive	0.936 (0.175)	1.208 (0.114)	0.375 (0.354)	0.703 (0.241)
	Recessive	0.617 (0.269)	3.170 (0.001)	2.023 (0.022)	2.500 (0.006)
0.04	Dominant	1.100 (0.136)	-0.817 (0.793)	-0.110 (0.544)	-1.038 (0.850)
	Additive	0.936 (0.175)	1.202 (0.115)	0.369 (0.356)	0.718 (0.236)
	Recessive	0.617 (0.269)	3.180 (0.001)	2.040 (0.021)	2.501 (0.006)

*The estimated allele frequency of the T allele is 0.0356, and models include both the presence and absence of within-family correlation (variance τ^2 of family-specific random effects set to 0 or 0.04).

^aParental phenotypes included.

reflecting differences in prostate cancer risk with respect to differences in calendar periods and ethnicity. The Total, Founder, and Nonfounder statistics all indicate a statistically significant association between early-onset prostate cancer and the homozygous TT genotype. In contrast, we found no significant association between prostate cancer risk and the A/T polymorphism using the score tests based on Cox's partial likelihood. The simulations suggest that this discrepancy is due the lower power of Cox's partial likelihood for data such as these, for which the majority of parental genotypes are missing and the deleterious allele is rare.

DISCUSSION

We have extended two score statistics for family genotype data to include censored times to failure and quantitative traits, and to accommodate residual within-family phenotype correlation. We also have extended the statistics to situations where the null distribution of parental genotypes is unknown and must be estimated. For the analysis of censored survival data in families, we have used simulations to compare the statistics to those based on Cox's partial likelihood function stratified on families, and to examine power loss for the genetically-based tests associated with misspecification of the baseline hazard function or the extent of residual phenotype correlation within families. The simulations suggest that the genetically-based tests are more powerful than those based on the stratified partial likelihood, and that this power advantage persists in the presence of model misspecification.

Application of the test statistics to genotype data from nuclear families with multiple cases of prostate cancer supports these conclusions and illustrates several features of the genetically-based tests. Missing parental genotypes are replaced by a probability distribution that is conditional on the genotypes observed in the family. The parameters in this distribution are estimated by maximizing the null likelihood of the data. If several nested models are fit to the parental genotype distribution and enough parents have been typed to discriminate among the models, then likelihood-ratio statistics could be used to test assumptions such as random parental mating and Hardy-Weinberg parental genotype proportions. The Founder statistics measures the deviation between observed (or inferred) parental genotype frequencies and their null expectations. The Nonfounder statistic extends the TDT and the score statistics of Schaid and Sommer [1994] and Schaid [1996] to include all offspring (regardless of phenotype), missing parental genotypes, and residual phenotype correlation within families.

The genetically-based tests have some limitations. The Founder statistic could be biased by inappropriate assumptions on the parental genotype distribution (e.g., random mating) or

by failure to adjust for differences in ethnic distribution among the test families. Comparison of the Founder and Nonfounder statistics provides a check on such bias; the presence of a statistically significant Founder statistic in the absence of a similar Nonfounder statistic would be grounds for suspicion. In principle, however, misspecification of the parental distribution also could induce bias in the Nonfounder statistic, although it is difficult to envisage situations that could lead to serious bias. The performance of these statistics in the presence of parental genotype misspecification will be the subject of subsequent work. The consequences of misspecifying the null phenotype probabilities, in contrast, are only suboptimal weights for summing contributions from individuals and families, and so such misspecification will not bias the test statistics. While power loss is an issue in principle, we have found it to be negligible in both simulations and data analysis.

There are alternatives to the Nonfounder statistic that do not require assumptions about the parental genotype distribution in the presence of untyped parents. Instead, these alternative statistics are conditioned more heavily on the observed genotypes among the offspring. The statistic proposed by Curtis [1997] and Knapp [1999] includes only families for which the parental genotypes can be reconstructed from those of their offspring, and its null distribution is conditioned on the types of families included. This approach loses power by ignoring information from families in which the parental genotypes are only partially reconstructible. The statistic proposed by Spielman and Ewens [1998], Schaid and Rowland [1998], and Horvath and Laird [1998] ignores parental genotypes. Instead, it compares genotypes of affected and unaffected offspring, conditioning on the observed genotypes and binary phenotypes of the offspring, but not on the pairing of genotype with phenotype. Such conditioning avoids the need for assumptions about parental genotypes, and the resulting statistic can be computed using standard software for conditional logistic regression. However, it loses power by excluding data from families in which all offspring are either phenotype-concordant (e.g., affected sib pairs) or genotype-concordant. Rabinowitz and Laird [2000], Laird et al. [2000], and Mokliatchouk et al. [2001] proposed conditioning the family-specific likelihood (22) on the minimal sufficient statistic $Y(\mathbf{g}, \mathbf{h})$ for γ , and using as test statistic the β -score evaluated at $\beta = 0$, which is independent of γ . When the two parental genotypes $\mathbf{g}$ are known, $Y = \mathbf{g}$ and this conditional likelihood is just (24). This approach avoids the need to estimate γ . However, the heavy conditioning also loses power by excluding families with genotype-concordant offspring and ignoring certain parental genotype information available from the offspring. For example, if one offspring is homozygous AA and another is homozygous BB , then both of the untyped parents must be heterozygous AB . Thus, the null probability of the observed ordered offspring outcome (AA, BB) is $1/16$. However, the conditioning described by Rabinowitz and Laird [2000] assigns null probability $1/2$ to each of the outcomes (AA, BB) and (BB, AA) and ignores all other possible outcomes. Because the Nonfounder statistic conditions only on the parental genotype information provided by the observed family genotypes, it includes data that are excluded by all of these more heavily conditioned tests. There is need for further study of the relative power of these tests under a broad spectrum of true models [see Knapp et al., 1995].

The likelihood function (22) lends itself to several possible generalizations. Extension to loci with $m > 2$ alleles is conceptually straightforward, although notationally more cumbersome. The scores in the score statistics are now vectors of dimension $m - 1$, and their variances are $(m - 1) \times (m - 1)$ matrices. However, the resulting test of the global null hypothesis of no association of any allele with disease may have poor power due to the increased degrees of freedom. Estimation of the association parameter β , with options for including covariates in the underlying penetrance model, is an extension of considerable interest. But as illustrated by

the work of Clayton [1999], such an extension forfeits the relative simplicity of the calculations under the null hypothesis $\beta = 0$, which make computing the score statistics tractable. Finally, there is need to extend the statistics to accommodate more varied types of family structures, including three-generation families and families containing half-sibs.

REFERENCES

- Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *Am J Hum Genet* 65:1170–7.
- Cox DR. 1972. Regression models and life tables (with discussion). *J R Stat Soc Series B* 34:187–220.
- Curtis D. 1997. Use of siblings as controls in case-control association studies. *Ann Hum Genet* 61:319–33.
- Horvath S, Laird NM. 1998. A discordant-sibship test for disequilibrium and linkage: No need for parental data. *Am J Hum Genet* 63:1886–97.
- Knapp M. 1999. The transmission/disequilibrium test and parental-genotype reconstruction: The reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet* 64:861–70.
- Knapp M, Wassmer G, Baur MP. 1995. The relative efficiency of the Hardy-Weinberg equilibrium-likelihood and the conditional on parental-genotype-likelihood methods for candidate-gene association studies. *Am J Hum Genet* 57:1476–85.
- Laird NM, Horvath S, Xu X. 2000. Implementing a unified approach to family-based tests of association. *Genet Epidemiol (Suppl 1)* 19:S36–S42.
- Mokliatchouk O, Blacker D, Rabinowitz D. 2001. Association tests for traits with variable age of onset. *Hum Hered* 51:46–53.
- Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 50:211–23.
- Schaid DJ. 1996. General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 13:423–49.
- Schaid DJ, Rowland C. 1998. Use of parents, sibs and unrelated controls for detection of association between genetic markers and disease. *Am J Hum Genet* 63:1492–1506.
- Schaid DJ, Sommer SS. 1994. Comparison of statistics for candidate-gene association between genetic markers and disease. *Am J Hum Genet* 55:402–9.
- Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *Am J Hum Genet* 62:450–8.
- 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* 52:506–16.
- Tu I, Balise RR, Whittemore AS. 2000. Detection of disease genes by use of family data. II. Application to nuclear families. *Am J Hum Genet* 66:1341–50.
- Whittemore AS, Tu I. 2000. Detection of disease genes by use of family data. I. Likelihood-based theory. *Am J Hum Genet* 66:1328–40.

APPENDIX A

Likelihood, Score, and Variance Contributions for a Family

We summarize briefly a family's contributions to the three test statistics. Because of equation (4), we need only consider the contribution from a family with known parental genotypes $\mathbf{g}_i = (g_{iM}, g_{iF})$. Let $\mathbf{h}_i = (h_{i1}, \dots, h_{in_i})$ denote the vector of offspring genotypes, and let $\Phi_i = (\Phi_{iF}, \Phi_{iNF})$ denote the vector of observed phenotypes for family members. (The phenotypes of some members may be missing.)

Likelihood and Score Contributions

The family's likelihood $L_i(\theta; \mathbf{g}_i, \mathbf{h}_i)$ is proportional to the probability of the family's genotype data $(\mathbf{g}_i, \mathbf{h}_i)$, given Φ_i :

$$L_i(\theta; \mathbf{g}_i \mathbf{h}_i) = \frac{P(\mathbf{g}_i; \gamma) P(\Phi_{iF} | \mathbf{g}_i, \beta) P(\mathbf{h}_i | \mathbf{g}_i) P(\Phi_{iNF} | \mathbf{h}_i, \beta)}{\sum_{\mathbf{g}'} P(\mathbf{g}'; \gamma) P(\Phi_{iF} | \mathbf{g}', \beta) \sum_{\mathbf{h}''} P(\mathbf{h}'' | \mathbf{g}') P(\Phi_{iNF} | \mathbf{h}'', \beta)} \quad (22)$$

We factor this likelihood as the product of a contribution

$$L_i^F(\theta; \mathbf{g}_i) = \frac{P(\mathbf{g}_i; \gamma) P(\Phi_{iF} | \mathbf{g}_i, \beta) \sum_{\mathbf{h}'} P(\mathbf{h}' | \mathbf{g}_i) P(\Phi_{iNF} | \mathbf{h}', \beta)}{\sum_{\mathbf{g}'} P(\mathbf{g}'; \gamma) P(\Phi_{iF} | \mathbf{g}', \beta) \sum_{\mathbf{h}''} P(\mathbf{h}'' | \mathbf{g}') P(\Phi_{iNF} | \mathbf{h}'', \beta)} \quad (23)$$

from the parental genotypes, times a contribution

$$L_i^{NF}(\beta; \mathbf{h}_i | \mathbf{g}_i) = \frac{P(\mathbf{h}_i | \mathbf{g}_i) P(\Phi_{iNF} | \mathbf{h}_i, \beta)}{\sum_{\mathbf{h}'} P(\mathbf{h}' | \mathbf{g}_i) P(\Phi_{iNF} | \mathbf{h}', \beta)} \quad (24)$$

from the offspring genotypes.

The corresponding family-specific score vectors are

$$\begin{aligned} \mathbf{u}_i(\theta; \mathbf{g}_i) &= \frac{d}{d\theta} \log L_i(\theta; \mathbf{g}_i; \mathbf{h}_i), \\ \mathbf{u}_i^F(\theta; \mathbf{g}_i) &= \frac{d}{d\theta} \log L_i^F(\theta; \mathbf{g}_i) \end{aligned} \quad (25)$$

and

$$\mathbf{u}_i^{NF}(\theta; \mathbf{g}_i) = \frac{d}{d\theta} \log L_i^{NF}(\beta; \mathbf{h}_i | \mathbf{g}_i). \quad (26)$$

Variance Contributions

The variance matrix of $\mathbf{u}_i(\mathbf{x}_i, \theta)$ is

$$\mathbf{J}_i = \mathbf{J}_i(\theta) = E[\mathbf{u}_i(\mathbf{x}_i, \theta) \mathbf{u}_i(\mathbf{x}_i, \theta)^T] = \begin{pmatrix} E[(u_{i1})^2] & E[u_{i1}u_{i2}] \\ E[u_{i1}u_{i2}] & E[(u_{i2})^2] \end{pmatrix}.$$

Using the relations $u_{i1} = u_{i1}^F + u_{i1}^{NF}$ and $u_{i2} = u_{i2}^F$, we can decompose this matrix as

$$\begin{aligned} \mathbf{J}_i &= \begin{pmatrix} E[(u_{i1}^F)^2] & E[u_{i1}^F u_{i2}^F] \\ E[u_{i1}^F u_{i2}^F] & E[(u_{i2}^F)^2] \end{pmatrix} + \begin{pmatrix} 2E[u_{i1}^F u_{i1}^{NF}] & E[u_{i1}^{NF} u_{i2}^F] \\ E[u_{i1}^{NF} u_{i2}^F] & 0 \end{pmatrix} + \begin{pmatrix} E[(u_{i1}^{NF})^2] & 0 \\ 0 & 0 \end{pmatrix} \\ &= \mathbf{J}_i^F + \mathbf{B}_i + \mathbf{J}_i^{NF}. \end{aligned} \quad (27)$$

The family's contribution to the variance matrix used for the Total and Founder statistics are, respectively, $\mathbf{J}_i$ and $\mathbf{J}_i^F$. Here the expectations are obtained by averaging over all possible realizations of the parental genotypes, weighted by their population frequencies. The contribution to the variance matrix used for the Nonfounder statistic is

$$\mathbf{J}_i^* = \begin{bmatrix} E[(u_{i1}^{NF})^2] & E[u_{i1}^{NF} u_{i2}^F] \\ E[u_{i1}^{NF} u_{i2}^F] & E[(u_{i2}^{NF})^2] \end{bmatrix}, \quad (28)$$

where the expectations are conditional on the observed or inferred parental genotypes.

For families with known parental genotypes, u_{i1}^{NF} is uncorrelated with both u_{i1}^F and u_{i2}^F , and so $\mathbf{B}_i = \mathbf{0}$. Thus, from (27) $\mathbf{J}_i = \mathbf{J}_i^F + \mathbf{J}_i^{NF}$, and from (28) $\mathbf{J}_i^* = \mathbf{J}_i^{NF}$ is diagonal and independent of γ . In contrast, when one or both parental genotypes are uncertain, $\mathbf{J}_i$ does not decompose into a sum of parental and offspring contributions, because the covariances $E[u_{i1}^{NF} u_{i1}^F]$ and $E[u_{i1}^{NF} u_{i2}^F]$ are nonzero.

We conclude by describing how to compute the entries of $\mathbf{J}_i$ of (27) and $\mathbf{J}_i^*$ of (28) for families with uncertain parental genotype. We illustrate the computation for just one of the entries, as they all follow the same pattern. Thus, for example, for J_i of (27),

$$\begin{aligned} E\{[u_{i1}^{NF}(\mathbf{x}_i)]^2\} &= \sum_{\mathbf{x}'} P(\mathbf{x}'; \gamma) \{u_{i1}^{NF}(\mathbf{x}')\}^2 \\ &= \sum_{g'_M} \sum_{g'_P} \sum_{h'_1=0}^2 \cdots \sum_{h'_n=0}^2 P(g'_M) P(g'_P) \prod_{j=1}^n P(h'_j | g'_M, g'_P) \{u_{i1}^{NF}(\mathbf{x}')\}^2. \end{aligned} \quad (29)$$

Here $u_{i1}^{NF}(\mathbf{x}')$ is given by equation (8), and

$$\mathbf{x}' = \begin{cases} (g'_M, g'_P, h'_1, \dots, h'_n) & \text{if } g_M \text{ and } g_P \text{ are both observed} \\ (*, g'_P, h'_1, \dots, h'_n) & \text{if, say, } g_M \text{ is missing} \\ (*, *, h'_1, \dots, h'_n) & \text{if } g_M \text{ and } g_P \text{ are both missing.} \end{cases}$$

These formulae give expectations with respect to the null distribution of parental and offspring genotypes. However, for the Nonfounder statistic we calculated expectations conditional on what can be inferred about the parental genotypes. Thus in the summation (29), we took

$$\mathbf{x}' = \begin{cases} (g_M, g_P, h'_1, \dots, h'_n) & \text{if } g_M \text{ and } g_P \text{ are both observed} \\ (*, g_P, h'_1, \dots, h'_n) & \text{if, say, } g_M \text{ is missing} \\ (*, *, h'_1, \dots, h'_n) & \text{if } g_M \text{ and } g_P \text{ are both missing.} \end{cases}$$

Also the prior probabilities $P(g'_M)$ and $P(g'_P)$ are replaced by the corresponding posterior probabilities for the two parental genotypes, given the genotypes, in the entire family. Note that all these variance formulae are based on the expected information matrix (i.e., the expected value of minus the second derivative of the loglikelihood with respect to θ , under the null hypothesis), while the variance formulae provided by Clayton [1999] are based on the observed information matrix.

APPENDIX B

Family-Specific Scores for Three Kinds of Phenotype

To derive expression (10) and (11), we first assume that, given their genotypes at the test locus, the phenotypes of family members are mutually independent:

$$P(\Phi_i | \mathbf{g}_i, \mathbf{h}_i) = \prod_{j \in \{M, F\}} P(\phi_{ij} | g_{ij}, \beta) \prod_{j=1, \dots, n_i} P(\phi_{ij} | h_{ij}, \beta).$$

By substituting this expression in equations (23) and (25), differentiating the result with respect to β , evaluating at $\beta = 0$, and then using relation (12), we obtain expression (10). Similar calculations using (24) and (26) give expression (11).

When a random effect ρ_i is used to accommodate residual within-family phenotype correlations, we assume that, given the value of the random effect and the genotypes of family members for the polymorphism of interest, the phenotypes in the family are mutually independent:

$$P(\Phi_i | \mathbf{g}_i, \mathbf{h}_i) = \int \left(\prod_{j \in \{M, F\}} P(\phi_{ij} | g_{ij}, \rho, \beta) \prod_{j=1, \dots, n_i} P(\phi_{ij} | h_{ij}, \rho, \beta) \right) \psi(\rho) d\rho,$$

where $\psi(\cdot)$ is the density of ρ_i . By substituting this expression in

$$L_i^F(\theta; \mathbf{g}_i) = \frac{P(\mathbf{g}_i; \gamma) \sum_{\mathbf{h}'} P(\mathbf{h}' | \mathbf{g}_i) P(\Phi_i | \mathbf{g}_i, \mathbf{h}', \beta)}{\sum_{\mathbf{g}'} P(\mathbf{g}'; \gamma) \sum_{\mathbf{h}''} P(\mathbf{h}'' | \mathbf{g}') P(\Phi_i | \mathbf{g}', \mathbf{h}'', \beta)}$$

and

$$L_i^{NF}(\beta; \mathbf{h}_i | \mathbf{g}_i) = \frac{P(\mathbf{h}_i | \mathbf{g}_i) P(\Phi_i | \mathbf{g}_i, \mathbf{h}_i, \beta)}{\sum_{\mathbf{h}'} P(\mathbf{h}' | \mathbf{g}_i) P(\Phi_i | \mathbf{g}_i, \mathbf{h}', \beta)},$$

differentiating the results with respect to β and evaluating at $\beta = 0$, we again obtain (10) and (11). In this derivation we have used the fact that

$$\frac{d}{d\beta} \log P(\Phi_i | \mathbf{g}_i, \mathbf{h}_i; \beta)_{\beta=0} = \sum_{j \in \{M, F\}} a(\phi_{ij}) z(g_{ij}) + \sum_{j=1}^{n_i} a(\phi_{ij}) z(h_{ij})$$

where

$$a(\phi) z(g) = \int \frac{d}{d\beta} \log P(\phi | g, \rho, \beta)_{\beta=0} P(\rho | \Phi) d\rho$$

and

$$P(\rho | \Phi) = \frac{P(\Phi | \rho) \psi(\rho)}{\int P(\Phi | \rho') \psi(\rho') d\rho'}.$$

Note that the weights $a(\phi)$ for the three phenotypes, given by (17)–(19), differ from those occurring in the absence of residual correlation.