Mathematical Statistics
Stockholm University

Combined Association and Linkage Analysis for General Pedigrees and Genetic Models

Ola Hössjer

Research Report 2004:11

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2004:11**,
http://www.math.su.se/matstat

Combined Association and Linkage Analysis for General Pedigrees and Genetic Models

Ola Hössjer

Abstract

A combined score test for association and linkage analysis is introduced, based on a biologically plausible model with association between markers and causal genes and penetrance between phenotypes and the causal gene. The test is based on a retrospective likelihood of marker data given phenotypes, treating the alleles of the causal gene as hidden data. It is defined for arbitrary outbred pedigrees, a wide class of genetic models and allows for missing marker data. For complete marker data, we give closed form expressions for the efficiency of the linkage and association parts of the test. These are exemplified for binary and quantitative phenotypes with or without polygenic effects. The conclusion is that association tests are comparatively more efficient than linkage tests for strong association, weak penetrance models, small families and non-extreme phenotypes, whereas the linkage test is more efficient for weak association, strong penetrance models, large families and extreme phenotypes. The combined tests is a robust alternative, which might be particularly useful when little is known about the genetic model.

KEY WORDS: Association, linkage, noncentrality parameter, retrospective likelihood, score test.

1 Introduction

Association and linkage analysis are complementary methods for localizing genes that increase susceptibility to a certain disease. Linkage analysis is often used in a first step for coarse mapping and then association methods are employed to fine map the regions pinpointed by linkage. Risch and Merikangas (1996) showed, on the other hand, for family trios/sib pairs, that association methods can be more powerful for complete genome scans, although current genotyping technology still makes such an approach expensive. Other authors have proposed joint tests for linkage and association to be used for genome scans. Terwilliger and Göring (2000) developed a method based on pseudomarkers, and Fulker et al. (1999) and Sham et al. (2000) considered variance components methods for quantitative traits. However, less attention has been paid on a general framework for association and linkage, using biologically based models with association between marker and disease causing alleles and penetrance parameters of the disease genes. As we show in this paper, it is possible to construct a joint score test of linkage and association for such models.

Whittemore (1996) developed score tests for linkage analysis by differentiating the retrospective log likelihood of marker data given phenotypes with respect to model parameters. McPeek (1999) and Hössjer (2003, 2004a) used the same approach, with biologically based models involving disease allele frequencies and penetrance parameters of the causal gene. These score tests put parametric and nonparametric linkage analysis on a common ground and accommodate a large class of genetic models, including generalized linear models with hidden Gaussian regimes. A variety of different kinds of phenotypes can be handled and polygenic effects are allowed for.

Association score tests are traditionally derived by modeling penetrance directly between marker genotypes and phenotypes. Schaid and Sommer (1994) and Schaid (1996) developed family based score tests for association that generalize the classical TDT test of Terwilliger and Ott (1992) and Spielman et al. (1993). These tests are conditional on phenotypes and observed parental genotypes and avoid spurious association due to population admixture as well as the need to estimate marker allele frequencies. The price to pay is loss of information and possibly reduced power. Indeed, Clayton (1999), Whittemore and Tu (2000), Tu et al. (2002) and Shih and Whittemore (2002) have recently developed more general score tests based on the conditional distribution of marker data given phenotypes, and shown that increased power is possible in many cases.

The joint score test for linkage and association that we develop in this paper

is valid for arbitrary pedigrees and missing marker information. As in Hössjer (2004a), we allow for polygenic effects and a large class of phenotypes. We consider biologically based models, with disease and marker allele frequencies, parameters for penetrance of the causal gene and association parameters between the causal and marker gene. Such a model seems complicated, but it turns out that the resulting score vector has a very explicit form. It is two-dimensional, consisting of one association and one linkage score. Moreover, very explicit expressions can be derived for the efficiency (noncentrality parameter) of the both the linkage component and various versions of the association component (with or without conditioning on founder genotypes and in the latter case, with known or estimated marker allele frequencies).

The paper is organized as follows. In Section 2 we define the retrospective likelihood of marker data given phenotypes. Since only marker alleles are observed, the alleles of the disease causing gene are treated as hidden variables. In Section 3 we describe the association and penetrance parts of the likelihood for a wide class of genetic models. Score functions and test statistics are derived in Sections 4 and 5. The asymptotic form of the noncentrality parameter is computed in Section 6 for the combined test as well as for the linkage and association components. This in turn gives the sample size N required for each test to attain a certain power. In Sections 7 and 8 we specialize to biallelic markers and nuclear families and give closed form expressions as well as numerical examples of N. Possible extensions are discussed in Section 9 and more technical results are gathered in the appendix.

2 Likelihood for combined association and linkage

Consider a pedigree with n individuals, f founders and $n-f$ nonfounders. A vector $Y = (Y_1, \ldots, Y_n)$ is observed, which contains phenotypes Y_k of all pedigree members (including the possibility $Y_k = '?'$, corresponding to an unknown phenotype of k). The genotypes at an unknown disease locus τ are $G = (G_1, \ldots, G_n)$, where $G_k = (a_{2k-1}a_{2k})$ consists of the the two alleles a_{2k-1} and a_{2k} that are transmitted to k through maternal and paternal meioses. We wish to test if τ is located in close vicinity of another locus x of known position, at which a specific marker is located. It has genotypes $H = (H_1, \ldots, H_n)$, where $H_k = (b_{2k-1}b_{2k})$ is the marker genotype of k which consists of two alleles. The marker data M contains parts of H and possibly also alleles from other markers surrounding x. Formally, we write the

pointwise hypothesis test at x as

$$\begin{aligned} H_0 &: \ x \text{ unlinked to } \tau, \ H \text{ is not associated to } G, \\ H_1 &: \ x = \tau. \end{aligned} \tag{1}$$

In order to test (1) we have phenotype data Y and marker data M from N families.

The genetic model parameters are $\theta = (p, \psi, q, \Delta)$, where $p = (p_0, p_1)$ consists of disease allele frequencies at the biallelic disease locus, ψ contains penetrance parameters, $q = (q_0, \ldots, q_{d-1})$ marker allele frequencies of the d possible marker alleles of H and Δ quantifies association between G and H. For all other markers than H, allele frequencies are assumed to be known, and hence not included as parameters. For one family, we use the retrospective likelihood (Prentice and Pyke, 1979), i.e. the conditional probability

$$L(\theta; M) = P_\theta(M|Y) \tag{2}$$

observed marker data given phenotypes. For N families, we take the product of the familywise likelihoods (2). An advantage of retrospective likelihood is that the ascertainment rule, i.e. the rule for sampling pedigrees, need not be modeled explicitly, as long as it depends on Y only (and not on M), see for instance Kraft and Thomas (2001).

As we will be seen in Section 4, all likelihood computations can be made assuming $x = \tau$. The reason is that H_0 is equivalent to choosing a subset of the parameter space at which the position of τ does not affect the likelihood. We emphasize that $x = \tau$ does not mean that H and G are identical. Instead it means that x and τ are so close that 1) H and G are likely to be in linkage disequilibrium and 2) crossovers between x and τ for all meioses in the pedigrees can be ignored. Therefore, let $v = (v_1, \ldots, v_m)$ be the *common* inheritance vector of a particular pedigree at loci x and τ, where $m = 2(n - f)$ is the number of meioses. It is a binary vector, where v_j equals 0 or 1 depending on whether a grand-paternal or grand-maternal allele was transmitted during formation of the j^{th} germ cell (Donnely, 1983). Let also $b = (b_1, \ldots, b_{2f})$ be the vector of marker alleles at x for the founders (assuming founders are numbered as $1, \ldots, f$). The purpose of marker data is to retain as much information about (b, v) as possible. Therefore, for each pedigree, we expand the familywise likelihood (2) as

$$L(\theta; M) = \sum_{b,v} P(M|b, v) P_\theta(b, v|Y), \tag{3}$$

assuming $P_\theta(M|b, v, Y) = P(M|b, v)$, that is, the conditional distribution of observed marker data given phenotypes and (b, v) is independent both of phenotypes and model parameters. This is clear when M only contains marker

data at x (a subset of H). Otherwise, it holds when the remaining markers are in linkage equilibrium with G. In absence of spurious association due to population admixture, this requirement will be fulfilled if these markers are linked to but sufficiently far apart from x.

In general, (b, v) is more informative than H, especially when H is a marker with low degree of polymorphism. When this is the case, and even if all individuals are genotyped at locus x (so that M contains all of H), other markers surrounding x are still needed to give additional information about v. In fact, for complete marker data, there are of 2^f pairs (b, v) with $P(M|b, v) = 1$, whereas $P(M|b, v) = 0$ for all other pairs. The 2^f pairs compatible with M all have the same probability and are obtained by shifting phase of all founders independently (Kruglyak et al., 1996). Hence, for complete marker data,

$$L(\theta; M) = 2^f P(b, v|Y), \tag{4}$$

where (b, v) is any of the 2^f pairs compatible with M. In view of (4), we will write $L(\theta; M) = L(\theta; b, v)$ for complete marker data in the sequel.

Göring and Terwilliger (2000) and Terwilliger and Göring (2000) noted that linkage and association analysis can be put into a unified framework by conditioning on disease locus genotypes G. We will follow a similar route and expand the complete marker data likelihood $L(\theta; b, v)$ by conditioning on the vector $a = (a_1, \ldots, a_{2f})$ of *founder alleles* at the disease locus,

$$\begin{aligned} L(\theta; b, v) &= 2^f \textstyle\sum_a P_{q,\Delta}(b|a) P_{p,\psi}(a, v|Y) \\ &\propto \textstyle\sum_a P_{p,q,\Delta}(a, b) P_\psi(Y|a, v) \end{aligned} \tag{5}$$

where the proportionality constant $2^f P(v)/P_{p,\psi}(Y) = 2^{f-m}/P_{p,\psi}(Y)$ depends on p and ψ but not on b and v. Notice that the association and penetrance parameters Δ and ψ appear in different factors in (5). We devote the next section to specifying these factors in more detail.

3 Modeling of Association and Penetrance

Assuming random mating and Hardy-Weinberg equilibrium, we model association between a and b as

$$\begin{aligned} P_{p,q,\Delta}(a, b) &= \textstyle\prod_{j=1}^{2f} P_{a_j,b_j}, \\ P_{a_j,b_j} &= p_{a_j} q_{b_j} (1 + \Delta s(a_j, b_j)), \end{aligned} \tag{6}$$

where $P_{a_j b_j}$ is the joint probability of (a_j, b_j) and $s = (s(i, j))_{ij}$ a $2 \times d$ matrix. In order to keep marginal allele frequencies fixed when Δ varies, we impose $\sum_j s(i, j) q_j = \sum_i s(i, j) p_i = 0$ for all i and j.

Example 1 (Biallelic markers.) For biallelic markers ($d = 2$), if

$$\Delta = \frac{P_{00}P_{11} - P_{01}P_{10}}{(p_0 p_1 q_0 q_1)^{1/2}} \tag{7}$$

is the correlation coefficient of a 2×2 table with cell probabilities P_{ij}, then

$$s = \begin{pmatrix} \left(\frac{p_1 q_1}{p_0 q_0}\right)^{1/2} & -\left(\frac{p_1 q_0}{p_0 q_1}\right)^{1/2} \\ -\left(\frac{p_0 q_1}{p_1 q_0}\right)^{1/2} & \left(\frac{p_0 q_0}{p_1 q_1}\right)^{1/2} \end{pmatrix}. \tag{8}$$

Other measures are also possible, such as $P_{00}P_{11} - P_{01}P_{10}$, which is the expected difference, for cells $(0,0)$ and $(1,1)$, between actual cell probabilities and those expected under independence. In this case $s(i,j) = 1$ if $i = j$ and -1 otherwise. See Devlin and Risch (1995) for more discussion on various measures of linkage disequilibrium. □

Since we assume that v is the inheritance vector for loci x and τ, the penetrance factor $P_\psi(Y|a,v)$ does not involve any crossovers between x and τ. With Y fixed, it is a function of G and can be written as $P_\psi(Y|G)$, where $\psi = (\psi(0), \psi(1), \psi(2))$ are the penetrance parameters and $\psi(j)$ corresponds to an individual with j copies of the disease causing allele, say 1. Other penetrance parameters, such as regression coefficients, polygenic and environmental variance components are considered fixed (or estimated from population data) and hence suppressed in the notation. Let $|G_k| = a_{2k-1} + a_{2k}$ be the number of disease causing alleles of the k^{th} pedigree member, put $\mu = (\psi(|G_1|), \ldots, \psi(|G_n|))$ and

$$P_\psi(Y|G) = f(Y;\mu).$$

For instance, in absence of polygenic and shared environmental effects

$$f(Y,\mu) = \prod_{k=1}^{n} f_k(Y_k;\mu_k), \tag{9}$$

although this restriction is not needed in general. Here $f_k(Y_k;\mu_k) = P(Y_k|G_k)$ is the penetrance factor for individual k. Dependence of f_k on k allows for individual covariates.

Example 2 (Binary phenotypes.) Let $Y_k = 1$ for an affected individual and $Y_k = 0$ for an unaffected one. In absence of polygenic and shared environmental effects (9), define

$$f_k(Y_k;\mu_k) = \mu_k^{(Y_k=1)}(1-\mu_k)^{(Y_k=0)}. \tag{10}$$

□

Example 3 (Quantitative phenotypes.) For quantitative traits, it is common to use a multivariate distribution $Y|G \in N(\mu, \sigma^2\Sigma)$, that is

$$f(Y;\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n|\Sigma|^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y-\mu)\Sigma^{-1}(Y-\mu)^T)\right), \qquad (11)$$

where ^{T} denotes vector transposition. This mixed model incorporates effects of the major gene G only in the mean vector, whereas $\sigma^2 = \mathrm{Var}(Y_k|G_k)$ and the correlation matrix $\mathrm{Corr}(Y|G) = \Sigma = (\Sigma_{kl})$ do not depend on G. For instance, if Σ incorporates additive polygenic effects, we have $\Sigma_{kl} = (1-h_a^2)1_{\{k=l\}} + h_a^2 r_{kl}$, where r_{kl} is the coefficient of relationship of k and l, i.e. the proportion of alleles shared identical-by-descent by k and l and h_a^2 the additive polygenic heritability. See Ott (1979) for more details. □

Example 4 (Gaussian liabilities.) In Hössjer (2004a), a large class of models was defined with a Gaussian liability $X|G \in N(\mu;\sigma^2\Sigma)$ as in Example 3. Observed phenotypes Y then depend on X through, for instance, a liability threshold, generalized linear or Cox proportional hazards model. If $\tilde{f}(\cdot;\mu)$ is a multivariate normal density as in (11), the penetrance function for this class of models can be written

$$f(Y;\mu) = \int P(Y|X)\tilde{f}(X;\mu)dX.$$

Consider, for instance, a liability threshold model. It is a generalization, for binary phenotypes, of Example 2, to incorporate polygenic and shared environmental effects. Let z be a given threshold and put $Y_k = 1_{\{X_k>z\}}$, so that a liability $\geq z$ implies disease. Then $P(Y|X) = 1_{\{X\in\Omega\}}$, where $\Omega = \{X;\ X_k > z \text{ if } Y_k = 1, X_k \leq z \text{ if } Y_k = 0 \text{ or } X_k \text{ arbitrary if } Y_k = '?'\}$. □

4 Score functions and tests

Following McPeek (1999) and Hössjer (2003, 2004a), we consider a one-dimensional trajectory $\{\psi_\varepsilon\}_\varepsilon$ of genetic models

$$\psi_\varepsilon = (m^*, m^*, m^*) + \varepsilon(u(0), u(1), u(2)).$$

Here $\varepsilon = 0$ corresponds to no genetic effect, $P(Y|G, \varepsilon = 0) = P(Y)$ of the major gene G and small ε gives a weak penetrance model, where G has weak

impact on Y. We reformulate the hypothesis testing problem as

$$\begin{aligned} H_0: &\quad \Delta = \varepsilon = 0, \\ H_1: &\quad \tau = x, \varepsilon \neq 0, \end{aligned}$$

which is mathematically equivalent to (1) except for the additional requirement $\varepsilon \neq 0$ under H_1 that makes H_0 and H_1 disjoint. Whereas $\varepsilon \neq 0$ is needed in H_1 for testing both association and linkage, $\Delta \neq 0$ is only needed in H_1 for testing association. Hence, we do not impose $\Delta \neq 0$ in a joint test for linkage and association. Further, we regard q as a nuisance parameter that needs to be estimated whereas p is kept fixed. Then the model parameters can be reduced to (q, Δ, ε). However, it is shown in the appendix that the scores of Δ and ε vanish for outbred pedigrees. For this reason, we reparametrize to $\epsilon = (\epsilon_0, \epsilon_1, \epsilon_2)$, where $\epsilon_0 = (q_1, \ldots, q_{d-1})$, $\epsilon_1 = \Delta\varepsilon$ and $\epsilon_2 = \varepsilon^2$. Notice that we only included $d-1$ components of q because of the constraint $\sum_{j=0}^{d-1} q_j = 1$. The score vector is

$$S(M) = \left.\frac{\partial \log L(\epsilon; M))}{\partial \epsilon}\right|_{\epsilon=(q,0,0)} = (S_0(M), S_1(M), S_2(M)), \tag{12}$$

where $S_i(M)$ is the partial derivative of $\log L$ with respect to ϵ_i. Using standard results for likelihood scores with missing data (Dempster et al., 1977), it follows, by differentiating (3) with respect to ϵ, that

$$S(M) = \sum_{b,v} P_q(b, v|M) S(b, v), \tag{13}$$

where $S(b, v) = (S_0(b, v), S_1(b, v), S_2(b, v))$ is the score function for complete marker data, defined as in (12), but with (b, v) instead of M. With $N > 1$ pedigrees, we simply add the score vectors (13) of each pedigree to obtain a total score vector $S(M)$.

We are primarily interested in score components S_1 and S_2, although S_0 is needed to account for the fact that we have to plug in a maximum-likelihood estimate of q into S_1 and S_2. Define $I_{ij} = E(S_i^T S_j)$, where expectation is under H_0, and let $J = (J_{ij})_{i,j=1}^2$ be the 2×2 Fisher information matrix for (ϵ_1, ϵ_2). In case q is known we put $J = I = (I_{ij})_{i,j=1}^2$, and when q has to be estimated, $J = I - (I_{01}, I_{02})^T I_{00}^{-1} (I_{01}, I_{02})$. The combined test statistic for linkage and association, when testing H_0 against H_1 is

$$T_{\text{combined}} = \begin{cases} (S_1, S_2) J^{-1} (S_1, S_2)^T, & \text{if } S_2 \geq 0, \\ (1, 0) J^{-1} (S_1, S_2)^T / (1, 0) J^{-1} (1, 0)^T, & \text{if } S_2 < 0, \end{cases} \tag{14}$$

and H_0 is rejected when T_{combined} exceeds a given threshold. Notice that $\epsilon_2 \geq 0$, whereas no sign constraint is put on ϵ_1. For this reason, T_{combined} is defined

differently depending on whether S_2 is negative or positive. Asymptotically, for large samples ($N \to \infty$), the null distribution of T_{combined} is a $0.5 : 0.5$ mixture of $\chi^2(1)$ and $\chi^2(2)$ distributions . This is a typical limit distribution when the null parameter is at the boundary of the parameter space (Self and Liang, 1987). The corresponding tests for pure association (H_1: $\tau = x$, $\Delta \neq 0$ and $\varepsilon \neq 0$) and pure linkage (H_1: $\tau = x$ and $\varepsilon \neq 0$) have tests statistics

$$\begin{aligned} T_1 &= S_1^2/(I_{11} - I_{01}^T I_{00}^{-1} I_{01}), \\ T_2 &= S_2/\sqrt{I_{22} - I_{02}^T I_{00}^{-1} I_{02}}, \end{aligned} \tag{15}$$

and H_0 is rejected when T_1 or T_2 exceed a given threshold. In case q is known we drop the terms $I_{01}^T I_{00}^{-1} I_{01}$ and $I_{02}^T I_{00}^{-1} I_{02}$ in the denominators. The same is true in the linkage case for complete marker data, since then $I_{02} = 0$. In any case, it is customary in linkage analysis not to account for the influence of estimating q and simply write $T_2 = S_2/\sqrt{I_{22}}$. The reason is that misspecification of q is then not as serious. Indeed, we show below that $S_2(b, v) = S_2(v)$, and hence $S_2(M) = \sum_v P_q(v|M) S_2(v)$. Therefore, q only enters in the conditional inheritance distribution $P_q(v|M)$ and not in centering of the score function. Asymptotically for large samples T_1 has a $\chi^2(1)$ and T_2 an $N(0,1)$-distribution under H_0.

5 Scores for complete marker data

For complete marker data, the scores S_i have a very explicit form. Let $\mu_0 = (m^*, \ldots, m^*)$, $\sigma_g^2 = \text{Var}(u_{|G_k|})$ and define

$$\begin{aligned} \omega_k &= \omega_k(Y) = \sigma_g \, \partial f(Y;\mu)/\partial \mu_k|_{\mu=\mu_0} / f(Y;\mu_0) \\ \omega_{kl} &= \omega_{kl}(Y) = \sigma_g^2 \, \partial^2 f(Y;\mu)/\partial \mu_k \partial \mu_l|_{\mu=\mu_0} / f(Y;\mu_0) \end{aligned}$$

as family-specific weights assigned to individuals and pairs of individuals.

Example 5 (Binary phenotypes, contd.) Let $K_p = P(Y_k = 1) = m^*$ be the prevalence of the disease when $\varepsilon = 0$. If $\sigma_g^2 = K_p^2(1 - K_p)^2$, it follows, by differentiating (10), that

$$\omega_k = Y_k - K_p. \tag{16}$$

Further, $\omega_{kl} = \omega_k \omega_l$ when $k \neq l$, and this is general property in absence of polygenic and shared environmental effects (9). If k and l is a monozygotic twin pair, it follows that the relative risk ratio (Risch, 1990) equals

$$\lambda = 1 + \omega_{kl} \varepsilon^2 + o(\varepsilon^2). \tag{17}$$

If $E(\psi(|G_k|)) = 0$, the prevalence is independent of ε and the remainder term in (17) vanishes. □

Example 6 (Quantitative phenotypes, contd.) Let $r = (Y - \mu_0)/\sigma = (r_1, \ldots, r_n)$ be the standardized vector of residuals. Then, if $\sigma_g^2 = 1$,

$$\begin{aligned} \omega_k &= (r\Sigma^{-1})_k, \\ \omega_{kl} &= (r\Sigma^{-1})_k (r\Sigma^{-1})_l - \Sigma^{-1}_{kl}, \end{aligned}$$

where Σ^{-1}_{kl} is the $(k,l)^{\text{th}}$ entry of Σ^{-1}, see Hössjer (2004a). Moreover, if $h^2 = \text{Var}(E(Y|G_k))/\text{Var}(Y_k)$ is the heritability at the main locus, then

$$\varepsilon^2 = \frac{h^2}{1 - h^2}.$$

□

Example 7 (Gaussian liabilities, contd.) If $\sigma^2 = 1$, then, as shown in Hössjer (2004a),

$$\begin{aligned} \omega_k &= \sigma_g \textstyle\int ((X - \mu_0)\Sigma^{-1})_k P(X|Y) dX \\ \omega_{kl} &= \sigma_g^2 \textstyle\int \left(((X - \mu_0)\Sigma^{-1})_k ((X - \mu_0)\Sigma^{-1})_l - \Sigma^{-1}_{kl} \right) P(X|Y) dX, \end{aligned} \tag{18}$$

where $P(X|Y) \propto P(Y|X)\tilde{f}(X; \mu_0)$ is the posterior distribution of X when $\varepsilon = 0$. Consider in particular the liability threshold model with $m^* = 0$. Then $\mu_0 = (0, \ldots, 0)$, $K_p = 1 - \Phi(z)$ is the prevalence $P(Y_k = 1)$ when $\varepsilon = 0$ and Φ is the distribution function of a standard normal random variable. Put $\sigma_g^2 = (1 - K_p)^2 K_p^2 / \phi^2(z)$, where $\phi = \Phi'$. Then (18) reduces to (16), in the special case of no polygenic or shared environmental effects, i.e. when Σ is an identity matrix. In general, the relative risk ratio of a monozygotic twin pair k, l can be written as $\lambda = \lambda^{\text{other}} \cdot \lambda^{\text{main}}$. It has two factors, of which the first one, λ^{other}, is due to polygenic and shared environmental effects, and the other, λ^{main}, is caused by the major gene G (Kurbasic and Hössjer, 2004). It can be shown that (17) generalizes to

$$\lambda^{\text{main}} = 1 + \omega_{kl}\varepsilon^2 + o(\varepsilon^2).$$

□

Decompose the genetic variance $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ into additive and dominant variance components $\sigma_a^2 = 2p_0p_1(p_0(\psi(1) - \psi(0)) + p_1(\psi(2) - \psi(1)))^2$ and $\sigma_d^2 = (p_0p_1)^2(\psi(2) - 2\psi(1) + \psi(0))^2$, and let $c = \sigma_d^2/\sigma_g^2$ be the fraction of variance due to dominance effects. Let also n_i be the number of marker alleles of type i among the founders, $i = 0, \ldots, d-1$. Then, it is shown in the appendix that for one outbred pedigree and complete marker data, the components of the score vector have the form

$$\begin{array}{rcl} S_0(b,v) = S_0(b) & = & (n_1/q_1 - n_0/q_0, \ldots, n_d/q_d - n_0/q_0), \\ S_1(b,v) = S_1(H) & = & \sqrt{1-c} \cdot \sum_{k=1}^{n} \omega_k (g(b_{2k-1}) + g(b_{2k})) - C_1 \\ S_2(b,v) = S_2(v) & = & \sum_{1 \le k < l \le n} \omega_{kl} \left((1-c)\mathrm{IBD}_{kl}/2 + c1_{\{\mathrm{IBD}_{kl}=2\}} \right) - C_2, \end{array} \tag{19}$$

where $g(b_1) = s(1, b_1)\sqrt{p_1/(2p_0)}$, IBD_{kl} is the number of alleles shared identical-by-descent by k and l and C_i is a centering constant that assures $E(S_i(b,v)) = 0$. For a set of N pedigrees, the familywise score components (19) are simply added to obtain the total score vector S.

For pure association testing, T_1 can be viewed as a generalization of the test statistics in Clayton (1999), Whittemore and Tu (2000) and Shih and Whittemore (2002), since polygenic and shared environmental effects are allowed for. For biallelic marker alleles ($d = 2$), g only attains two values, so without loss of generality we may rescale and put $g(b_k) = b_k$. Therefore, in this case, the influence of the marker allele is additive. Non-additive effects can be attained by dropping the assumption of Hardy-Weinberg equilibrium in (6). In the linkage case, T_2 coincides with the test statistic in Hössjer (2004a), see also McPeek (1999) and Hössjer (2003b).

The association score may be split into founder and nonfounder terms, $S_1 = S_1^{\mathrm{F}} + S_1^{\mathrm{NF}}$. The nonfounder score is defined conditionally on observed phenotypes *and* founder genotypes. Let $g^0(b_k) = g(b_k) - E(g(b_k)|b)$. Then

$$S_1^{\mathrm{NF}}(H) = \sqrt{1-c} \sum_{k=f+1}^{n} \omega_k \left(g^0(b_{2k-1}) + g^0(b_{2k}) \right), \tag{20}$$

where the sum ranges over all nonfounders $\{f+1, \ldots, n\}$, see Clayton (1999), Whittemore and Tu (2000) and Shih and Whittemore (2002). The analogous score for incomplete marker data is defined as in (12) with S_1^{NF} instead of S on the RHS. Moreover, 'nonfounder versions' of the combined association and linkage test T_{combined}, as well as the pure association test T_1 can be defined by replacing S_1 with S_1^{NF} everywhere. For instance, the modified version of T_1 is

$$T_1^{\mathrm{NF}} = (S_1^{\mathrm{NF}})^2/(I_{11}^{\mathrm{NF}} - (I_{01}^{\mathrm{NF}})^T I_{00}^{-1} I_{01}^{\mathrm{NF}})$$

when q has to be estimated, with $I_{i1}^{\mathrm{NF}} = E(S_i S_1^{\mathrm{NF}})$. When q is known, we drop the term $(I_{01}^{\mathrm{NF}})^T I_{00}^{-1} I_{01}^{\mathrm{NF}}$. This is a generalization of the classical TDT test to arbitrary pedigrees and phenotypes. An advantage of T_1^{NF} compared to T_1 is less sensitivity to model misspecification in terms of marker allele frequencies and spurious association due to population admixture. The price to pay is decreased efficiency. For complete marker data S_1^{NF} is orthogonal to S_0 and marker allele frequencies need not be estimated. An even more robust approach is to condition on a minimal sufficient statistic under the null hypothesis, see Rabinowitz and Laird (2000) for details. The distribution of resulting test statistic is independent of marker allele frequencies and adjusts for association due to population admixture for any kind of marker data.

6 Noncentrality Parameters and Efficiency

We define the noncentrality parameter

$$\eta^2_{\mathrm{combined}} = 2E_\theta \left(\log \frac{L(\theta)}{L(\theta_0)} \right) \tag{21}$$

as twice the expected increase of the log likelihood under θ compared to θ_0, the null hypothesis parameters. It grows with N if θ is kept fixed. However, for θ close to θ_0, standard likelihood theory implies that the asymptotic approximation

$$\eta^2_{\mathrm{combined}} = (\epsilon_1, \epsilon_2) J (\epsilon_1, \epsilon_2)^T. \tag{22}$$

may be used, provided that the RHS of (22) stays bounded when N grows. Moreover, the distribution of the untruncated version $(S_1, S_2) J^{-1} (S_1, S_2)^T$ of T_{combined} has approximately a noncentral χ^2 distribution with two degrees of freedom and noncentrality parameter $\eta^2_{\mathrm{combined}}$. For complete marker data, it is shown in the appendix that the noncentrality parameter (22) simplifies to

$$\eta^2_{\mathrm{combined}} = (\Delta\varepsilon)^2 J_{11} + \varepsilon^4 I_{22} =: \eta_1^2 + \eta_2^2. \tag{23}$$

The two terms on the RHS of (23) are the noncentrality parameters of the pure association and linkage tests respectively, since asymptotically T_1 has a noncentral χ^2-distribution with one degree of freedom and noncentrality parameter η_1^2 and T_2 a $N(\eta_2, 1)$-distribution. Formula (23) holds for all three versions of the association part of the combined test, with J_{11} taking values

$$\begin{array}{rcl} J_{11}^{\mathrm{km}} & = & I_{11}, \\ J_{11}^{\mathrm{em}} & = & I_{11} - I_{01}^T I_{00}^{-1} I_{01}, \\ J_{11}^{\mathrm{NF}} & = & I_{11}^{\mathrm{NF}}. \end{array} \tag{24}$$

Superscripts 'km' and 'em' denote association tests without conditioning on founder genotypes, using known or estimated marker allele frequencies respectively.

Consider a sample of one single *pedigree type*, i.e. a sample where all pedigrees have the same structure and identical phenotypes. Our goal is to assess the sample sizes $N_1^{\mathrm{em}}(\alpha,\beta)$, $N_1^{\mathrm{km}}(\alpha,\beta)$, $N_1^{\mathrm{NF}}(\alpha,\beta)$ and $N_2(\alpha,\beta)$ needed for level α tests T_1^{em}, T_1^{km}, T_1^{NF} and T_2 to attain power β. Write $N(\alpha,\beta)$ for any of these three quantities and $\eta^2(N)$ for any of the four noncentrality parameters $(\eta_1^{\mathrm{em}})^2$, $(\eta_1^{\mathrm{km}})^2$, $(\eta_1^{\mathrm{NF}})^2$ and η_2^2 based on sample size N. We also assume that a number of loci x are tested, so that multiple testing correction has to be incorporated into the definition of $N(\alpha,\beta)$. It is shown in the appendix that

$$N(\alpha,\beta) \approx \frac{(\lambda_{\tilde{\alpha}} + \lambda_{1-\tilde{\beta}})^2}{\eta^2(1)}, \tag{25}$$

where $\lambda_\alpha = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of a standard normal $N(0,1)$ random variable. A formula similar to (25) appears in Risch and Merikangas (1996). The numbers $\tilde{\alpha}$ ($= \tilde{\alpha}_1^{\mathrm{em}}$, $\tilde{\alpha}_1^{\mathrm{km}}$, $\tilde{\alpha}_1^{\mathrm{NF}}$ or $\tilde{\alpha}_2$) and $\tilde{\beta}$ ($= \tilde{\beta}_1^{\mathrm{em}}$, $\tilde{\beta}_1^{\mathrm{km}}$, $\tilde{\beta}_1^{\mathrm{NF}}$ or $\tilde{\beta}_2$) can be interpreted as the pointwise one-sided significance level and power after correction for multiple testing. The adjusted pointwise power satisfies $\tilde{\beta} \leq \beta$. The larger the region around τ is where a significant test result is considered as a true positive, the smaller is $\tilde{\beta}$, and $\tilde{\beta} = \beta$ if a significant test at τ is required to be characterized as a true positive. The adjusted pointwise significance levels can be interpreted by viewing the multiple testing problem under H_0 as performing an 'effective number' K of one-sided independent tests. The Bonferroni approximation for small α is $\tilde{\alpha} \approx \alpha/K$, although exact expressions can be found in the appendix. Since one two-sided test roughly corresponds to two one-sided tests for small α, we put $K_1 = 2$ and $K_2 = 1$ when one locus is tested, where K_1 is any of K_1^{em}, K_1^{km} or K_1^{NF}. In general, $K_1/2$ and K_2 are the effective number of one-sided association and linkage tests respectively. If test statistics of nearby *actual* loci are dependent under H_0, then number of actual loci is larger than the effective number of 'independent' loci. Since linkage disequilibrium decays faster than correlation of allele sharing statistics, $K_1/2$ is in general larger than K_2 and hence the numerator of (25) is larger for the two association tests than for the linkage test.

Using T_1^{em} as benchmark to compare the other test statistics with, the asymptotic relative efficiency (ARE)

$$\mathrm{ARE}(T, T_1^{\mathrm{em}}) = \lim_{\Delta,\varepsilon\to 0} \frac{N_1^{\mathrm{em}}(\alpha,\beta)}{N(\alpha,\beta)} = \mathrm{MT} \cdot \frac{\eta^2(1)}{(\eta_1^{\mathrm{em}})^2(1)}, \tag{26}$$

is the asymptotic ratio of $N(\alpha,\beta)$ and $N_1^{\rm em}(\alpha,\beta)$, see Noether (1955). In the last equality we used (25), and introduced the multiple testing factor $\mathrm{MT} = (\lambda_{\tilde{\alpha}_1^{\rm em}} + \lambda_{1-\tilde{\beta}_1^{\rm em}})^2/(\lambda_{\tilde{\alpha}} + \lambda_{1-\tilde{\beta}})^2$. Typically, MT is greater than one for linkage ($T = T_2$), but close to 1 for the two association tests ($T = T_1^{\rm km}$ or $T_1^{\rm NF}$).

Formula (25) can be generalized to the case when the sample is drawn from a *population* of pedigree types, as in Hössjer (2003a). Then, if $\eta^2(\phi)$ is the noncentrality parameter for a pedigree of type ϕ, the sample size required for a level α to attain power β is

$$N(\alpha,\beta) \approx \frac{(\lambda_{\tilde{\alpha}} + \lambda_{1-\tilde{\beta}})^2}{\int \eta^2(\phi) d\nu(\phi)}, \tag{27}$$

where ν is the distribution of pedigree types in the population. In other words, if $\tilde{\alpha}$ and $\tilde{\beta}$ are independent of pedigree type ϕ, the required sample size is the harmonic mean of the required sample sizes of each pedigree type.

Likewise, the ARE can be defined for a collection of N different pedigree types. It is shown in the appendix that

$$\mathrm{ARE}(T, T_1^{\rm em}) \approx \sum_{j=1}^{N} w_j \mathrm{ARE}_j(T, T_1^{\rm em}), \tag{28}$$

where $\{w_j\}$ are weights summing to one and ARE_j the asymptotic relative efficiency (26) of the $j^{\rm th}$ pedigree type.

7 Biallelic Markers and Nuclear Families

We now specialize to biallelic markers and complete marker data, and further assume that Δ is the correlation coefficient (7). In the appendix, we give general expressions for the Fisher information. In particular, for one nuclear family ($N = 1$) with two parents ($k = 1, 2$) and $n - 2$ siblings ($k = 3, \ldots, n$) we get

$$\begin{array}{rcl} J_{11}^{\rm km} & = & 0.5 \cdot (1-c)\left((\sum_{k=1}^n \omega_k)^2 + (\omega_1 - \omega_2)^2 + \sum_{k=3}^n \omega_k^2\right) \\ J_{11}^{\rm em} & = & 0.5 \cdot (1-c)\left((\omega_1 - \omega_2)^2 + \sum_{k=3}^n \omega_k^2\right) \\ J_{11}^{\rm NF} & = & 0.5 \cdot (1-c)\sum_{k=3}^n \omega_k^2, \\ I_{22} & = & 0.125 \cdot (1 + 0.5 \cdot c^2)\sum_{3 \le k < l \le n} \omega_{kl}^2 \end{array} \tag{29}$$

Combining (29) with (23) and (25), we get explicit expressions for the sample size required for T_1 and T_2 to attain power β.

Because of (25) and (29), and assuming MT = 1 for the association tests, we can write

$$\mathrm{ARE}(T, T_1^{\mathrm{em}}) = \begin{cases} \mathrm{MT} \cdot \mathrm{Dom} \cdot \mathrm{Phen} \cdot \Delta^2/\varepsilon^2, & T = T_2, \\ \mathrm{Phen}, & T = T_1^{\mathrm{km}} \text{ or } T_1^{\mathrm{NF}}. \end{cases} \tag{30}$$

where $\mathrm{Dom} = (1 + 0.5c^2)/(1 - c)$ is a dominance factor and Phen a factor depending on phenotypes and the genetic model through the weights $\{\omega_k\}$ and $\{\omega_{kl}\}$. In more detail, Phen is J_{11}^{km}, J_{11}^{NF} or I_{22} divided by J_{11}^{em} when $c = 0$.

8 Numerical Results

We evaluated the required sample size $N(\alpha, \beta)$ in a genomewide scan for different pedigree types and genetic models. Since (25) is based on asymptotic approximations, it is accurate mainly when ε is small (linkage) or when ε and Δ are small (association). For the linkage test, we used extreme value theory of stochastic processes to adjust for multiple testing, see the figure captions and appendix for details. No such theory is available for association analysis. Instead, we used the distance $\delta = 0.1$ cM between two adjacent 'effectively independent' association tests, since it is reasonable to assume that δ is is the range 0.05-0.2 cM (Reich et al., 2001). We used $\tilde{\beta}_1 = \beta = 0.8$ in all plots, requiring, for the association tests, a significant peak at the disease locus itself to be declared as a true positive. For the linkage test, a less stringent criterion was used in defining a true positive. With a smaller value of $\tilde{\beta}_1$, $N(\alpha, \beta)$ would decrease somewhat for all association tests.

In Figures 1-3, $N(\alpha, \beta)$ is calculated for three different pedigrees types with binary phenotypes. The model parameters of Examples 5 and 7 are varied one at a time. When $h_a^2 > 0$, the integral expressions for ω_k and ω_{kl} are evaluated by means of a rapid importance sampling algorithm (using 10000 samples) for multivariate normal distributions truncated on a rectangular region (Gottlow and Sadeghi, 1999).

In general the associations tests are more efficient than the linkage test for weak penetrance models (small λ or small ε) and strong association (large Δ) whereas the linkage test is more efficient for strong penetrance models and weak association. This can easily be explained since $N(\alpha, \beta)$ is inversely proportional to $(\Delta\varepsilon)^2$ for the association tests and inversely proportional to ε^4 for the linkage test. The prevalence has small effect on efficiency (with λ fixed), whereas increased polygenic variance often decreases efficiency, at least for concordant phenotypes. The linkage test is more efficient relative

to $T_1^{\rm em}$ and $T_1^{\rm NF}$ for larger pedigrees. This is not surprising, since I_{22} grows quadratically with pedigree size, whereas $J_{11}^{\rm em}$ and $J_{11}^{\rm NF}$ grow linearly with pedigree size. Among the three association tests, $T_1^{\rm km}$ is most efficient, followed by $T_1^{\rm em}$ and $T_1^{\rm NF}$. This is evident from the figures, but can also be seen by comparing the Fisher informations in (29). (However, $T_1^{\rm km}$ is also most sensitive to model misspecification, followed by $T_1^{\rm em}$ and $T_1^{\rm NF}$.) When parents have unknown phenotypes ($\omega_1 = \omega_2 = 0$), $T_1^{\rm em}$ and $T_1^{\rm NF}$ are equally efficient, as in Figures 1 and 2.

Figures 4-6 display $N(\alpha, \beta)$ for three different pedigree types using the Gaussian model of Example 6. Similar remarks can be made regarding the effect of penetrance, association and pedigree size. In addition, more extreme phenotypes (larger k if the figures) make the linkage test more efficient compared to the association tests. This follows since I_{22} is proportional to k^4 when there are no polygenic effects whereas J_{11} is proportional to k^2 for all three association tests. The effect of polygenic variance depends on the pedigree type. For concordant (discordant) phenotypes, the efficiency decreases (increases) when h_a^2 increases, and more so for the linkage test than for the association tests. Notice that $T_1^{\rm km}$ and $T_1^{\rm em}$ have identical efficiency when $\sum_k \omega_k = 0$, as in Figure 5. In this case the marker allele frequencies (ϵ_0) are orthogonal to ϵ_1, so no asymptotic efficiency loss is induced by estimating them.

The fraction of dominance variance at the disease locus, c, is zero in Figures 1-6. To see the effect of increasing c, we plot, in Figure 7, the dominance term (Dom) of the ARE between the linkage and association tests. Dom is close to one for small values of c (≤ 0.1), but then increases rapidly.

We investigated the effect of varying δ in Figure 8, where the multiple testing term of the ARE is plotted as function of δ. MT is quite insensitive to variations in δ. When $\alpha = 0.05$ and $\beta = 0.8$, the increased multiple testing for association compared to linkage leads to a decrease in efficiency by a factor 1.5 (2) for chromosomewide (genomewide) tests. When $\alpha = 0.01$, the decrease in efficiency is somewhat smaller.

Summarizing, the relative efficiency of linkage and association tests depends on both the pedigree type and the genetic model. There is no method that is uniformly better than the other. The required sample size $N_{\rm combined}(\alpha, \beta)$ of the combined test can obtained by putting $\eta^2_{\rm combined}$ in the denominator of (25). The numerator will have a more complicated form. Due to the increased number of degrees of freedoms, we conjecture it to be slightly larger than for the association part, which in turn is larger than for the linkage part. Therefore, we conjecture that the efficiency of the combined test is close T_1 when the association test is most powerful and a bit less efficient than T_2

when the linkage test is most powerful. This would make T_{combined} a robust alternative to using either T_1 or T_2. However, if prior knowledge of the genetic model is available, and it is known which of T_1 and T_2 is most powerful, it is probably a better strategy not to use the combined test, but the most efficient of T_1 and T_2.

9 Discussion

In this paper we derived a combined score test for linkage and association. It can be used for arbitrary combinations of (outbred) pedigree structures and allows for missing marker information in a general way. The test uses biologically based genetic models, with 1) marker allele frequencies, 2) association between markers and the causal gene and 3) penetrance between the causal gene and phenotypes as parameters. The genotypes at the causal gene are treated as hidden variables in the likelihood computations. Our framework facilitates efficiency comparisons between the linkage and association parts of the score test, as well as between various versions of the association test, with or without estimating marker allele frequencies, and with and without conditioning on founder marker genotypes. We derived general efficiency formulas for complete marker data and biallelic markers. A conclusion of these is that the association tests are comparatively more efficient than the linkage tests for weak penetrance and strong association models, smaller families and less extreme phenotypes (quantitative phenotypes). No methods is uniformly superior, and a combined linkage and association test might be a useful robust alternative when little is known about the genetic model.

Our approach has some similarities with that of Chapman et al. (2003), who also treat disease gene alleles as hidden variables. Their focus is mainly on population based case-control studies and prospective likelihoods $P(Y|M)$, while we focus on family-based association studies and retrospective likelihoods $P(M|Y)$. In Chapman et al. (2003), the relation between marker and disease causing alleles are modeled in terms of linear regression, while we use the joint distribution of marker and disease alleles among founders. We believe some ideas of Chapman et al. (2003) could be incorporated into our framework. Firstly, a pure association test can be derived by keeping $\Delta \neq 0$ fixed and only differentiating the log likelihood with respect to penetrance parameters. Secondly, when haplotype marker alleles are used and d, the length of q, is large, Δ can be chosen as a vector. The choice of dimensionality of such a vector is a trade-off between size of noncentrality parameter and number of degrees of freedoms. For instance, if the haplotype consists of B biallelic markers, $d = 2^B$, and the dimensionality of Δ should be somewhere

between B (locus scoring) and $d-1$ (haplotype scoring). Chapman et al. (2003) and Clayton et al. (2004) conclude that in many cases locus scoring is more powerful. It would be interesting to see if such a conclusion is valid also in our framework of family-based association studies.

For quantitative traits, Fulker et al. (1999) and Sham et al. (2000) introduced a joint likelihood ratio test for linkage and family-based association. They model $P(Y|M)$ through a multivariate normal distribution $N(\mu, \Sigma)$, and thereby avoid summing over disease genotypes. The association and linkage parameters are contained in μ and Σ respectively. The corresponding score vector obtained by differentiating $\log P(Y|M)$ with respect all model parameters, is closely related to our score vector (S_1, S_2) (assuming q is known). In fact, for additive models, with identical within and between family association parameters, calculations (not shown here) reveal that S_1 and S_2 are the mean and covariance part of the Fulker et al. score vector. Hence, in view of the asymptotic equivalence between likelihood ratio and score tests, their LR test for combined association and linkage is asymptotically equivalent to T_{combined}.

Sham et al. (2000) compute noncentrality parameters for their joint test of linkage and association, which they subsequently split into linkage and association terms. As in our paper, they conclude that the linkage noncentrality parameter is proportional to h^4 for small heritabilities h^2, whereas the association parameter is proportional to $\Delta^2 h^2$ for biallelic markers and small Δ and h^2. However, they use a prospective rather than retrospective likelihood and define the noncentrality parameter as

$$\eta^2_{\text{combined}} = 2\left(E_\theta \log L(\theta) - E_{\theta_0} \log L(\theta_0)\right)$$

instead of (21), so our our results are not directly comparable.

Several other extensions are possible. The effect of incomplete marker data on efficiency and power should be studied more carefully. This arises either because of unknown marker haplotype phase or untyped pedigree members. To a large extent, such an investigation will employ simulations, since explicit analytical expressions for noncentrality parameters are hard to obtain for incomplete marker data except in certain special cases. Another extension is to relax the assumption of Hardy-Weinberg equilibrium. Finally, rare-disease models could be analyzed, where the penetrance parameters are kept fixed but the frequency of the disease causing allele tends to zero. In the pure linkage case, such score tests has been derived by McPeek (1999) and Hössjer (2003b, 2004a).

Appendix

Derivation of complete marker data score functions. To begin with, we use parameters (q, Δ, ε) to motivate why a reparametrization is appropriate. Since p is fixed we drop it in the notation. We regard a in (5) as hidden data and define the full data likelihood as

$$L^{\mathrm{f}}(q, \Delta, \varepsilon; a, b, v) \propto P_{q,\Delta}(a, b) P_{\psi_\varepsilon}(Y|a, v),$$

with superscript f for 'full data'. The proportionality constant is the same as in (5). It depends on ε but is independent of (a, b, v). The complete marker data likelihood $L(q, \Delta, \varepsilon) = L(q, \Delta, \varepsilon; b, v)$ satisfies

$$L(q, \Delta, \varepsilon) = E(L^{\mathrm{f}}(q, \Delta, \varepsilon)|b, v). \tag{A.1}$$

Let $S^{\mathrm{f}}_{ijk}(a, b, v) = L^{\mathrm{f}}(q, 0, 0)^{-1} \cdot \partial^{i+j+k} L^{\mathrm{f}}(q, \Delta, \varepsilon)/(\partial^i q \partial^j \Delta \partial^k \varepsilon)\big|_{(q,\Delta,\varepsilon)=(q,0,0)}$ be the full data score function of order (i, j, k) and define $S_{ijk}(b, v)$ analogously for complete marker data. It follows, by differentiating (A.1), that

$$S_{ijk}(b, v) = E(S^{\mathrm{f}}_{ijk}(a, b, v)|b, v). \tag{A.2}$$

Next we compute the leading S_{ijk} terms needed in a Taylor series expansion of $\log L(q, \Delta, \varepsilon)$ around $(q, 0, 0)$. We let C denote a centering constant that assures $E(S^{\mathrm{f}}_{ijk}) = 0$ or $E(S_{ijk}) = 0$, whose value may differ from line to line. It follows from (6) and (A.2) that $S_{100}(b, v) = S^{\mathrm{f}}_{100}(a, b, v)$ is identical to $S_0(b)$ in (19). Since $L^{\mathrm{f}}(q, \Delta, 0) = 2^{f-m} P_{q,\Delta}(a, b)$, it follows from (A.1) that $L(q, \Delta, 0) = 2^{f-m} P_q(b)$ is independent of Δ. Hence $S^{\mathrm{f}}_{0j0}(b, v) = 0$ for all $j > 0$. Further,

$$\begin{array}{rcl} S^{\mathrm{f}}_{011}(a, b, v) & = & \sigma_g^{-1} \sum_{j=1}^{2f} \sum_{k=1}^{2n} s(a_j, b_j) \omega_k u(|G_k|) - C \\ S^{\mathrm{f}}_{001}(a, b, v) & = & \sigma_g^{-1} \sum_{k=1}^{n} \omega_k u(|G_k|) - C \\ S^{\mathrm{f}}_{002}(a, b, v) & = & 2\sigma_g^{-2} \sum_{1 \le k < l \le n} \omega_{kl} u(|G_k|) u(|G_l|) \\ & + & \sigma_g^{-2} \sum_{k=1}^{n} \omega_{kk} u^2(|G_k|) - C. \end{array} \tag{A.3}$$

Assuming an outbred pedigree, there are exactly two of the founder alleles a_j that are IBD to the alleles of $G_k = (a_{2k-1} a_{2k})$. Since $E(s(a_j, b_j)|b_j) = 0$ (see the discussion below (6)), it follows form independence of $\{(a_j, b_j)\}_{j=1}^{2f}$ that

$$\begin{aligned} S_{011}(b, v) &= \sigma_g^{-1} \textstyle\sum_{j=1}^{2f} \sum_{k=1}^{2n} E(s(a_j, b_j) \omega_k u(|G_k|)|b, v) - C \\ &= \sigma_g^{-1} \textstyle\sum_{k=1}^{2n} \omega_k \left(E(s(a_{2k-1}, b_{2k-1}) u(|G_k|)|b, v) + E(s(a_{2k}, b_{2k}) u(|G_k|)|b, v) \right) - C \\ &= \sigma_g^{-1} \textstyle\sum_{k=1}^{2n} \omega_k \left(E(s(a_{2k-1}, b_{2k-1}) u(|G_k|)|b_{2k-1}) + E(s(a_{2k}, b_{2k}) u(|G_k|)|b_{2k}) \right) - C \\ &= \sqrt{1 - c} \cdot \textstyle\sum_{k=1}^{n} \omega_k \left(g(b_{2k-1}) + g(b_{2k}) \right) - C, \end{aligned}$$

and the last line is identical to $S_1(H)$ in (19). In the last step, we used Lemma 1 in Hössjer (2003b) to conclude $u(|(a_1a_2)|) = \sigma_a(a_1+a_2)/\sqrt{2p_0p_1} + \sigma_d(a_1-p_0)(a_2-p_1)/(p_0p_1) + C$, where C is a constant, independent of a_1 and a_2. Neither S^f_{001} nor S^f_{002} depend on b, and we can apply results from Hössjer (2004a) to deduce, for an outbred pedigree, that $S_{001}(b,v) = 0$ and $S_{002}(b,v) = 2S_2(v)$. Summarizing, we have shown that

$$\log \frac{L(q',\Delta,\varepsilon)}{L(q,0,0)} = (q'-q)S_0^T + \Delta\varepsilon S_1 + \varepsilon^2 S_2 + o(|q'-q| + |\Delta\varepsilon| + \varepsilon^2).$$

A reparametrization $\epsilon_0 = q$, $\epsilon_1 = \Delta\varepsilon$ and $\epsilon_2 = \varepsilon^2$ shows that indeed S_i in (19) are valid score functions for complete marker data. □

Proof of (23) and (24) for complete marker data. To prove (23), it suffices to verify that $I_{02} = (0,\ldots,0)^T$ and $I_{12} = 0$, since then $J_{12} = J_{21} = 0$, $J_{22} = I_{22}$. Assuming independence of b and v (no segregation distortion) it follows immediately that $S_0 = S_0(b)$ and $S_2 = S_2(v)$ are independent, and hence $I_{02} = (0,\ldots,0)^T$. We prove $I_{12} = 0$ for the version (19) of S_1 without conditioning on founders' marker genotypes. (The proof for S_1^{NF} is analogous.) Let $S_{kl} = (1-c)\mathrm{IBD}_{kl}/2 + c1_{\mathrm{IBD}_{kl}=2}$ and notice that

$$I_{12} = \sqrt{1-c}\cdot\sum_{k=1}^{n}\sum_{1\le k'<l'\le n}\left(\mathrm{Cov}(g(b_{2k-1}),S_{k'l'}(v)) + \mathrm{Cov}(g(b_{2k}),S_{k'l'}(v))\right). \tag{A.4}$$

Since $\{b_j\}$ are independent and identically distributed, the conditional distribution $b_k|v = b_{j_k(v)}$ is independent of v, where $1 \le j_k(v) \le 2f$ is the founder allele number that has been transmitted to allele k, $k = 1,\ldots,2n$. Hence b_k and v are independent. Applying this in (A.4) $I_{12} = 0$ follows. Finally, (24) follows from the definition of J and I, and, for the nonfounder statistic, the fact that $I_{01}^{\mathrm{NF}} = 0$ for complete marker data. □

Required sample size (25) for one type of pedigree. Consider a set of marker loci $\Omega = \{x_1,\ldots,x_{\tilde{K}}\}$ on one or several chromosomes and let $T(x)$ be the value of test statistic T at locus x. As before, T could be any of T_1^{km}, T_1^{em}, T_1^{NF} and T_2. Let $t = t(\alpha)$ be the threshold for rejecting H_0. We wish to choose t and $N(\alpha,\beta)$ as solutions of the first and second equations in

$$\begin{array}{rcl} P\left(\max_{x\in\Omega} T(x) \ge t|H_0\right) & = & 1 - P\left(T < t|H_0\right)^{K'} = \alpha \\ P\left(\max_{x\in\tilde{\Omega}} T(x) \ge t|H_1\right) & = & \beta \end{array} \tag{A.5}$$

respectively, where K' is the effective number of independent pointwise tests and $\tilde{\Omega} \subset \Omega$ a region surrounding τ at which we declare rejections as true

positives. The larger $\tilde{\Omega}$ is, the more liberal we are in defining a 'true set of candidate loci'. Typically, K' is smaller than the actual number $\tilde{K}$ of marker loci due to dependence of test statistics. Letting K denote the effective number of independent *one-sided* tests, we have $K' = K$ when $T = T_2$ and $K' = K/2$ for the three association tests T_1^{km}, T_1^{em} and T_1^{NF}. Starting with the first equation in (A.5), the asymptotic approximation entails that for large samples N

$$T(x) \overset{H_0}{\in} \begin{cases} \chi^2(1), & T = T_1, \\ N(0,1), & T = T_2, \end{cases} \tag{A.6}$$

where T_1 is any of the three association tests. Then $\tilde{\alpha}$, the pointwise one-sided significance level satisfies

$$\tilde{\alpha} = \begin{cases} P(T(x) \geq t|H_0)/2 = 1 - \Phi(\sqrt{t}), & T = T_1, \\ P(T(x) \geq t|H_0) = 1 - \Phi(t), & T = T_2. \end{cases} \tag{A.7}$$

The first equation of (A.5) then implies $\tilde{\alpha} = (1 - (1-\alpha)^{2/K})/2$ for the association tests and $\tilde{\alpha} = 1 - (1-\alpha)^{1/K}$ for the linkage test. Solving for t we find

$$t = \begin{cases} \lambda_{\tilde{\alpha}}^2, & T = T_1, \\ \lambda_{\tilde{\alpha}}, & T = T_2. \end{cases} \tag{A.8}$$

For the power calculations, we assume a) $\tau \in \Omega$, and that $T(x)$ has maximal noncentrality parameter η at $x = \tau$ under H_1, given by (23) and b) that the power β in the second equation of (A.5) can be written as a function of η. Assuming N pedigrees with the same structure and with identical phenotypes, the Fisher information matrix satisfies $J(N) = NJ(1)$, where $J(1)$ is the Fisher information matrix when $N = 1$. Hence it follows from (23) that $\eta = \sqrt{N}\eta(1)$ if ϵ does not depend on N. Asymptotically we have

$$T(\tau) \in \begin{cases} \chi^2(1, N\eta^2(1)), & T = T_1, \\ N(\sqrt{N}\eta(1), 1), & T = T_2, \end{cases} \tag{A.9}$$

under H_1[1]. Define then the pointwise power[2]

$$\tilde{\beta} = P(T(\tau) \geq t|H_1) = \begin{cases} 1 - \Phi(\sqrt{t} - \sqrt{N}\eta(1)), & T = T_1, \\ 1 - \Phi(t - \sqrt{N}\eta(1)), & T = T_2, \end{cases} \tag{A.10}$$

[1]Technically, the asymptotic approximation (A.9) requires a sequence of contiguous alternatives. That is $\epsilon_1 = \Delta\varepsilon$ and $\epsilon_2 = \varepsilon^2$ tend to zero with N at a rate $N^{-1/2}$ in (23), so that $\eta_1(N)$ and $\eta_2(N)$ stay bounded, see Noether (1955) for details.

[2]The last identity is an approximation for $T = T_1$, assuming that the term $\Phi(-\sqrt{t} - \sqrt{N}\eta(1))$ can be ignored.

It satisfies $\tilde{\beta} \leq \beta$ because of (A.5) and (A.10). In fact, $\tilde{\beta}$ gets smaller the larger the region $\tilde{\Omega}$ is, with $\tilde{\beta} = \beta$ if $\tilde{\Omega} = \{\tau\}$. The required sample size $N(\alpha, \beta)$ is found by solving for N in the second part of (A.5), or equivalently, solving for N in (A.10). Using the latter approach (25) follows. □

Robustness against nonnormality. The required sample size formula (25) is fairly robust against deviations from normality. In fact, write $T = \tilde{T}^2$ if $T = T_1$ and $T = \tilde{T}$ if $T = T_2$. Assume there exists a monotone transformation G such that

$$G(\tilde{T}) \in N(\sqrt{N}\eta(1), 1)$$

both under H_0 ($\eta(1) = 0$) and H_1 ($\eta(1) \neq 0$). Then (25) follows by similar calculations as in the normal case ($G(t) = t$). □

Required sample size and ARE for combinations of different pedigree types. Let ϕ_j be the type of the j^{th} pedigree and $\eta^2(\phi_j)$ be corresponding noncentrality parameter of test statistic T at the disease locus. Then $\eta^2 = \sum_{j=1}^{N} \eta^2(\phi_j)$ for a sample of size N, since Fisher information is added over pedigrees. Then, by similar arguments as those leading to (25) we find

$$N(\alpha, \beta) = \frac{(\lambda_{\tilde{\alpha}} + \lambda_{1-\tilde{\beta}})^2}{\sum_{j=1}^{N(\alpha,\beta)} \eta^2(\phi_j)/N(\alpha, \beta)}. \tag{A.11}$$

Assuming pedigrees are drawn from a population, the denominator of (A.11) converges to the limit $\int \eta^2(\phi) d\nu(\phi)$ by the law of large numbers and this proves (27). Hence, for large samples, we may approximate the denominator of (A.11) by $\sum_{j=1}^{N} \eta^2(\phi_j)/N$, where N is the *given sample size*, and plug into (26). If also the multiple testing factor of T and T_1^{em} is assumed to be the same for all pedigree types ϕ_j, (28) follows after some calculations, with weights $w_j = (\eta_1^{\text{em}})^2(\phi_j) / \sum_{k=1}^{N} (\eta_1^{\text{em}})^2(\phi_k)$. □

Pointwise significance and power for linkage. For linkage analysis, when complete marker data and a dense marker map along C chromosomes of total length $L > 0$ cM is available, we use asymptotic extreme value theory for Gaussian processes from Feingold et al. (1993) and Lander and Kruglyak (1995) and put $K = C + 2\rho L t^2$, where ρ is the crossover rate. The value of ρ depends on the pedigree and the score function S_2. For most pedigrees its value is in the range 0.01-0.04, see Ängquist and Hössjer (2004) for details. According to (A.5), t is the solution of

$$\Phi(t)^{C+2\rho L t^2} = 1 - \alpha. \tag{A.12}$$

and then $\tilde{\alpha}$ is computed from (A.8). For power, we use the asymptotic approximation

$$\beta = \tilde{\beta} + \phi(t-\eta)\left(\frac{2}{\eta d} - \frac{1}{\eta(2d-1)+t}\right), \tag{A.13}$$

of Feingold et al. (1993, formula (A.8)), where $\tilde{\beta}$ is defined in (A.10), $\phi = \Phi'$ is the standard normal density and d a constant that is close to 1 for most pedigrees (Hössjer, 2004). This formula corresponds to choosing $\tilde{\Omega}$ as a region surrounding τ in which the largest peak of $T(x)$ is located with high probability under H_1

Fisher information for T_1^{km}, T_1^{em} and T_2. For biallelic markers, with Δ and s as in Example 1, it follows that $g(0) = -\sqrt{q_1/(2q_0)}$ and $g(1) = \sqrt{q_0/(2q_1)}$. Let z_{kl} denote the probability that alleles k and l are shared identical by descent. Then

$$\begin{aligned}
&\mathrm{Cov}\left(g(b_{2k-1}) + g(b_{2k}), g(b_{2l-1}) + g(b_{2l})\right) \\
&= (g(1)-g(0))^2 \mathrm{Var}(b_1)\left(z_{2k-1,2l-1} + z_{2k-1,2l} + z_{2k,2l-1}, z_{2k,2l}\right) \\
&= 2(g(1)-g(0))^2 \mathrm{Var}(b_1) r_{kl} \\
&= r_{kl},
\end{aligned}$$

where $r_{kl} = E(\mathrm{IBD}_{kl})/2$ is the coefficient of relationship, i.e. the proportion of alleles shared IBD, by k and l. It follows that

$$I_{11} = (1-c) \cdot \sum_{k,l=1}^{n} \omega_k \omega_l r_{kl}. \tag{A.14}$$

When $d = 2$, write $S_0 = (n_1 - E(n_1))/(q_0 q_1)$. Since for any $k \in \{1, \ldots, 2f\}$, $\mathrm{Cov}(g(b_k), n_1) = (g(1)-g(0))\mathrm{Cov}(b_k, n_1) = (g(1)-g(0))\mathrm{Var}(b_k) = \sqrt{q_0 q_1/2}$, we get $I_{01} = \sqrt{1-c}/(q_0 q_1) \cdot \sum_{k=1}^{n} \left(2\omega_k \sqrt{q_0 q_1/2}\right) = \sqrt{2(1-c)/q_0 q_1} \sum_{k=1}^{n} \omega_k$. Combining this with $I_{00} = 2f/(q_0 q_1)$, it follows that

$$I_{01}^2/I_{00} = (1-c)(\sum_{k=1}^{n} \omega_k)^2/f. \tag{A.15}$$

Define S_{kl} as in (A.4), and $\Sigma_{kl,k'l'} = \mathrm{Cov}(S_{kl}(v), S_{k'l'}(v))$ when $1 \le k < l \le n$ and $1 \le k' < l' \le n$. Then

$$I_{22} = \sum_{kl,k'l'} \omega_{kl} \omega_{k'l'} \Sigma_{kl,k'l'}. \tag{A.16}$$

In particular, for a nuclear family with two parents ($k = 1, 2$) and $n-2$ children ($k = 3, \ldots, n$), we have $f = 2$, $r_{kk} = 1$, $r_{12} = r_{21} = 0$ and $r_{kl} = 0.5$

for all other k, l with $k \neq l$. Further, $\Sigma_{kl,k'l'} = 0.125 + c^2 \cdot 0.0625$ when $(k,l) = (k',l')$ and both k and l are siblings and zero for all other k, l, k', l', see Hössjer (2004b). Hence (A.14)-(A.16) simplify to

$$\begin{aligned} I_{11} &= 0.5 \cdot (1-c) \left((\textstyle\sum_{k=1}^n \omega_k)^2 - 2\omega_1\omega_2 + \sum_{k=1}^n \omega_k^2 \right) \\ I_{01}^2/I_{00} &= 0.5 \cdot (1-c)(\textstyle\sum_{k=1}^n \omega_k)^2, \\ I_{22} &= 0.125 \cdot (1 + 0.5 \cdot c^2) \textstyle\sum_{3 \leq k < l \leq n} \omega_{kl}^2, \end{aligned}$$

which in turn implies the first two and the fourth equations in (29). □

Fisher information for $T_1^{\mathbf{NF}}$. Let $1 \leq k \leq 2n$ be a given allele and $1 \leq j_k = j_k(v) \leq 2f$ the founder allele that is transmitted to k. Introduce $p_{kj} = P(j_k(v) = j)$ and, for any pair $1 \leq k, l \leq 2n$ of alleles, $\alpha_{kl} = z_{kl} - \sum_{j=1}^{2f} p_{kj} p_{lj}$, where z_{kl} is the probability that k and l are shared IBD. If $b_k^0 = b_k - E(b_k|b)$, it follows after some calculations that

$$E(b_k^0 b_l^0) = \alpha_{kl} q_0 q_1 \tag{A.17}$$

Combining (A.17) with the definition (20) of the nonfounder score, it follows, for complete marker information, that

$$\begin{aligned} I_{11}^{\mathrm{NF}} &= (1-c) \textstyle\sum_{k,l=1}^n \omega_k \omega_l \left(E(g^0(b_{2k-1}) g^0(b_{2l-1})) + E(g^0(b_{2k-1}) g^0(b_{2l})) \right. \\ &\quad \left. + E(g^0(b_{2k}) g^0(b_{2l-1})) + E(g^0(b_{2k}) g^0(b_{2l})) \right) \\ &= (1-c)(g(1) - g(0))^2 \textstyle\sum_{k,l=1}^n \omega_k \omega_l \left(E(b_{2k-1}^0 b_{2l-1}^0) + E(b_{2k-1}^0 b_{2l}^0) \right. \\ &\quad \left. + E(b_{2k}^0 b_{2l-1}^0) + E(b_{2k}^0 b_{2l}^0) \right) \\ &= 0.5(1-c) \textstyle\sum_{k,l=f+1}^n \omega_k \omega_l (\alpha_{2k-1,2l-1} + \alpha_{2k-1,2l} + \alpha_{2k,2l-1} + \alpha_{2k,2l})), \end{aligned} \tag{A.18}$$

where, in the last step, we used that $\alpha_{kl} = 0$ if either k or l is a founder allele, i.e. if either $1 \leq k \leq 2f$ or $1 \leq l \leq 2f$. For a nuclear family, it is easy to see that $\alpha_{kl} = 0.5 \cdot 1_{\{k=l\}}$ for a nonfounder pair $2f + 1 = 5 \leq k, l \leq 2n$ of alleles. Hence (A.18) becomes

$$I_{11}^{\mathrm{NF}} = 0.5(1-c) \sum_{k=3}^n \omega_k^2,$$

which, for complete marker data, is identical to the third equation of (29). □

References

Ängquist, L. and Hössjer, O. (2004). Improving the calculation of statistical significance in genome-wide scans. To appear in *Biostatistics*.

Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Human Heredity* **56**, 18-31.

Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *Am. J. Hum. Genet.* **65**, 1170-1177.

Clayton, D., Chapman, J. and Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology* **27**(4), 415-428.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1-38.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311-322.

Donnely, P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology* **23**, 34-64.

Feingold, E., Brown, P. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. of Hum. Genet.* **53**, 234-251.

Fulker, D.W., Cherny, S.S., Sham, P.C. and Hewitt, J.K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259-267.

Gottlow, M. and Sadeghi, S. (1999). Forward inclusion in multiple linear regression. A conditional approach. Master Thesis 1999:E11, Mathematical Statistics, Lund University. (In Swedish.)

Göring, H.H.H. and Terwilliger, D. (2000). Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am. J. Hum. Genet.* **66**, 1310-1327.

Hössjer, O. (2003a). Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Ann. Statist* **31**, 1075-1109.

Hössjer, O. (2003b). Determining Inheritance Distributions via Stochastic Penetrances. *Journal of the American Statistical Association*, **98**, 1035-1051.

Hössjer, O. (2004a). Conditional likelihood score functions in linkage analysis. To appear in *Biostatistics*.

Hössjer, O. (2004b). Information and effective number of meioses in linkage analysis. To appear in *J. Math. Biol.*

Hössjer, O. (2004c). Spectral decomposition of score functions in linkage analysis. Mathematical Statistics, Stockholm University, Report 2003:21.

Kraft, P. and Thomas, D. (2001). Bias and efficiency in family-based gene characterization studies: Conditional, prospective, retrospective and joint likelihoods. *Am. J. Hum. Gen.* **66**, 1119-1131.

Kurbasic, A. and Hössjer, O. (2004). Relative risks for general phenotypes and genetic models. Working paper.

Lake, S.L., Blecker, D. and Laird, N.M. (2000). Family-based tests of association in the presence of linkage. *Am. J. Hum. Genet.* **67**, 1515-1525.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241-247.

McPeek, S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidimiol.* **16**, 225-249.

Noether, G.E. (1955). On a theorem of Pitman. *Ann. Math. Statist.* **26**, 64-68.

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigree analysis. *Am. J. Hum. Gen.* **31**, 161-175.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

Rabinowitz, D. and 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**, 211-223.

Reich, D.E. et al. (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199-204.

Risch, N. (1990). Linkage strategies for genetically complex traits I. Multilocus models. *Am. J. Hum. Genet.*, **46**, 222-228.

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

Schaid, D.J. (1996). General score tests for association of genetic markers with disease using cases and their parents. *Genet. Epidmiol.* **13**, 423-449.

Schaid, D.J. and Sommer, S.S. (1994). Comparison of statistics for candidate-gene association between genetic markers and disease. *Am. J. Hum. Genet.* **55**, 402-409.

Sham, P.C., Cherny, S.S., Purcell, S. and Hewitt, J.K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616-1630.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**, 605-610.

Shih, M-C. and Whittemore, A.S. (2002). Tests for genetic association using family data. *Genet. Epidmiol.* **22**, 128-145.

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

Terwilliger, J.D. and Göring, H.H.H. (2000). Gene mapping of the 20th and 21th centuries: Statistical methods, data analysis and experimental design. *Human Biology* **1**, 63-132.

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

Tu, I-P., Balise, R.R. and Whittemore, A.S. (2000). Detection of disease genes by use of family data. II. Application to nuclear families. *Am. J. Hum. Genet.* **66**, 1341-1350.

Whittemore, A. (1996). Genome scanning for linkage: An overview. *Biometrics* **59**, 704-716.

Whittemore, A.S. and Tu, I-P. (2000). Detection of disease genes by use of family data. I. Likelihood-based theory. *Am. J. Hum. Genet.* **66**, 3128-1340.

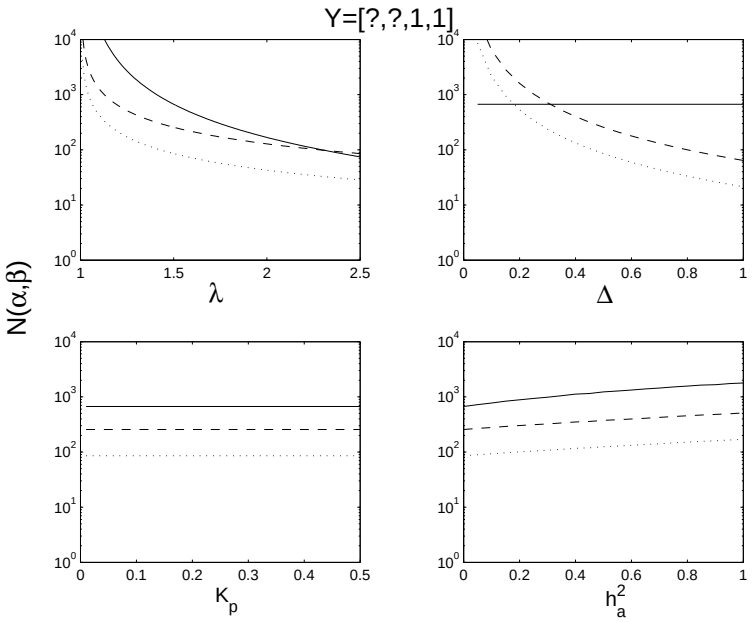

Figure 1: The number of affected sib pairs ($Y = (?, ?, 1, 1)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan including all 22 autosomes ($C = 22$, $L = 3575$ cM). The curves correspond to T_1^{km} (dotted), T_1^{em} and T_1^{NF} (dashed) and T_2 (solid). Only one parameter is varied, and the remaining ones are kept fixed at $K_p = 0.1$, $\Delta = 0.5$, $h_a^2 = 0$, $\delta = 0.1$ cM, $c = 0$ and $\lambda = 1.5$, where $\lambda = 1+(1-K_p)^2\varepsilon^2$ is the relative risk of an affected MZ twin pair in absence of polygenic effects. For details on multiple testing correction, see Figure 8.

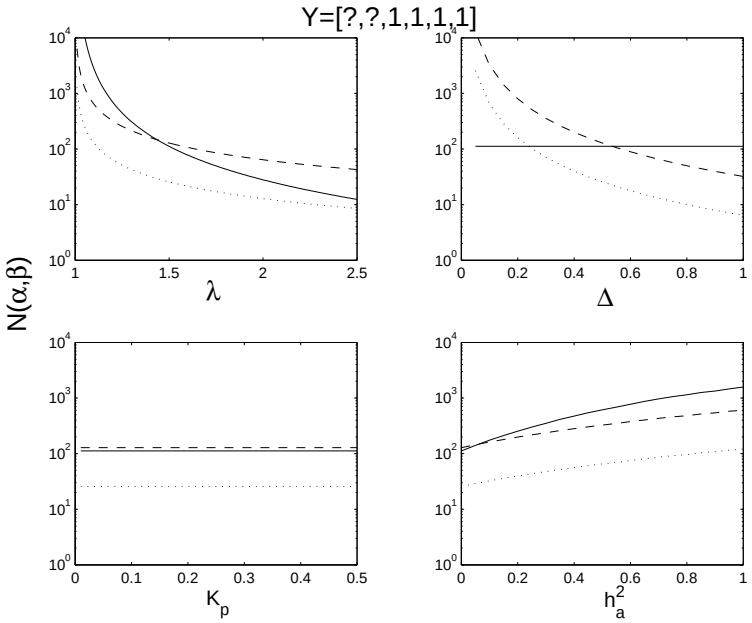

Figure 2: The number of affected sib quartets ($Y = (?, ?, 1, 1, 1, 1)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan ($C = 22$, $L = 3575$ cM). The curves correspond to T_1^{km} (dotted), T_1^{em} and T_1^{NF} (dashed) and T_2 (solid). See Figure 1 for details on parameter values.

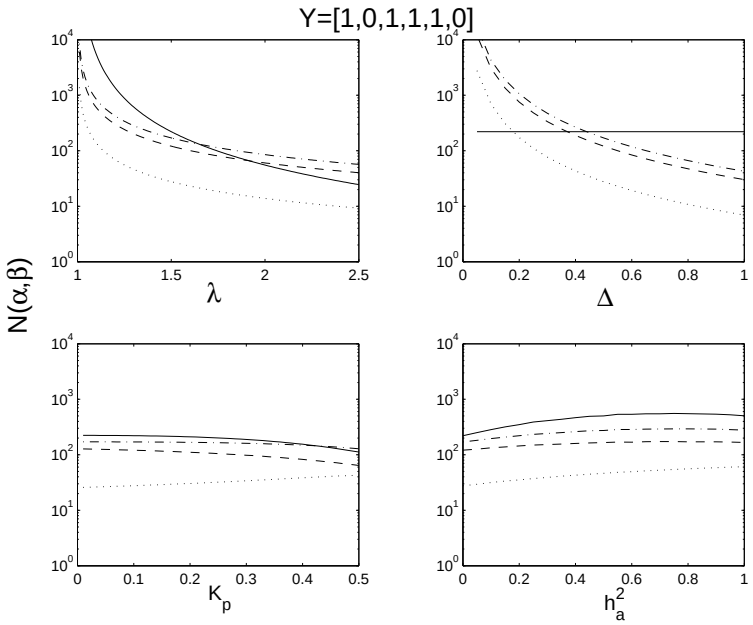

Figure 3: The number of nuclear families with binary phenotypes $Y = (1, 0, 1, 1, 1, 0)$ required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan ($C = 22$, $L = 3575$ cM). The curves correspond to T_1^{km} (dotted), T_1^{em} (dashed), T_1^{NF} (dash-dotted) and T_2 (solid). See Figure 1 for details on parameter values.

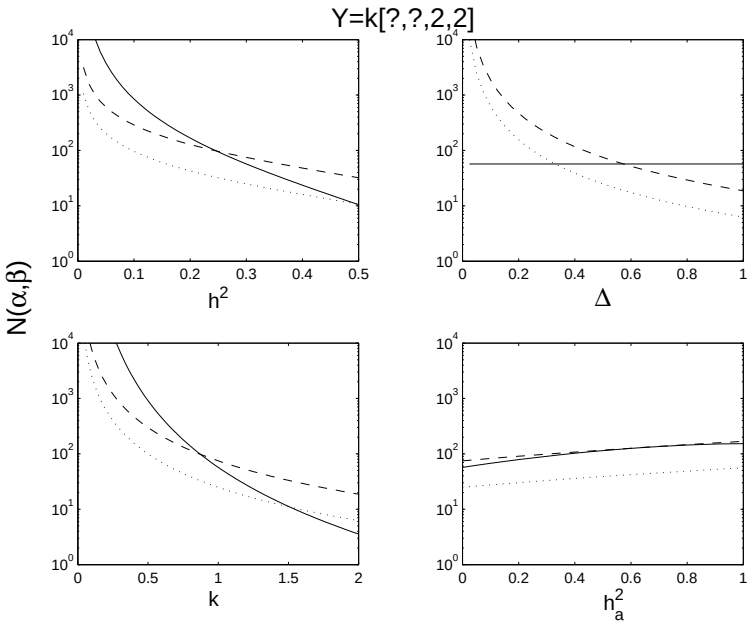

Figure 4: The number of concordant sib pairs (quantitative phenotypes, $Y = k(?, ?, 2, 2)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan ($C = 22$, $L = 3575$ cM). The curves correspond to T_1^{km} (dotted), T_1^{em} and T_1^{NF} (dashed) and T_2 (solid). Only one parameter value is varied, and the others equal $m^* = 0$, $\sigma = 1$, $k = 1$, $\Delta = 0.5$, $h^2 = 0.3$, $h_a^2 = 0$, $c = 0$ and $\delta = 0.1$ cM. For details on multiple testing correction, see Figure 8.

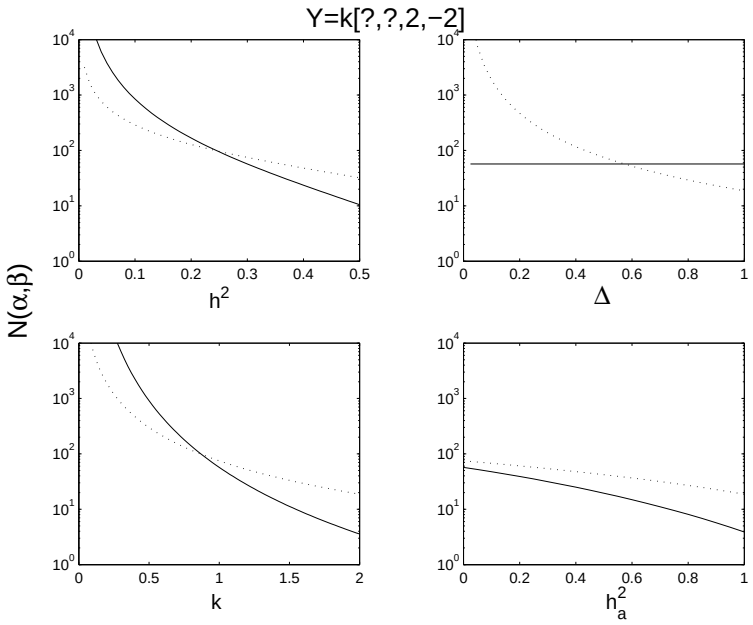

Figure 5: The number of discordant sib pairs (quantitative phenotypes, $Y = k(?, ?, 2, -2)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan ($C = 22$, $L = 3575$ cM). The curves correspond to T_1^{km}, T_1^{em} and T_1^{NF} (dotted) and T_2 (solid). For details on parameter values, see Figure 4.

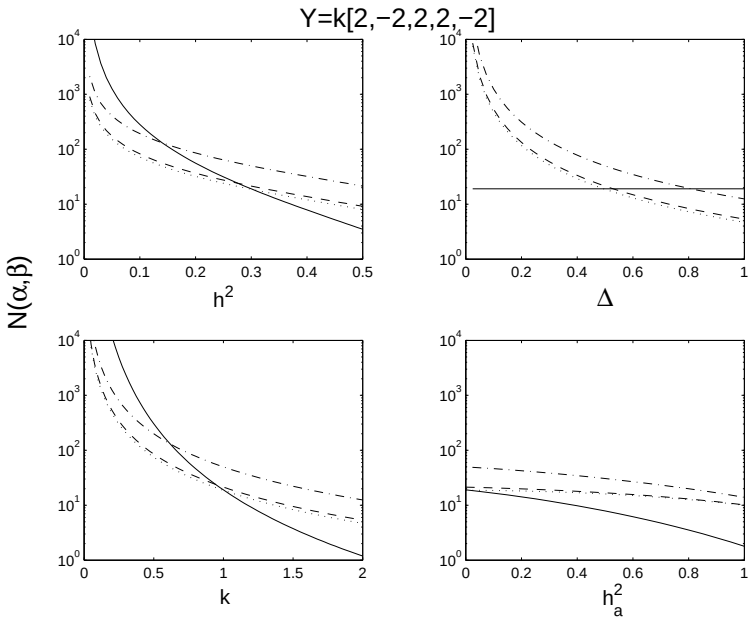

Figure 6: The number nuclear families with quantitative phenotypes and phenotype vector $Y = k(2, -2, 2, 2, -2)$ required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan ($C = 22$, $L = 3575$ cM). The curves correspond to T_1^{km} (dotted), T_1^{em} (dashed) T_1^{NF} (dash-dotted) and T_2 (solid). For details on parameter values, see Figure 4.

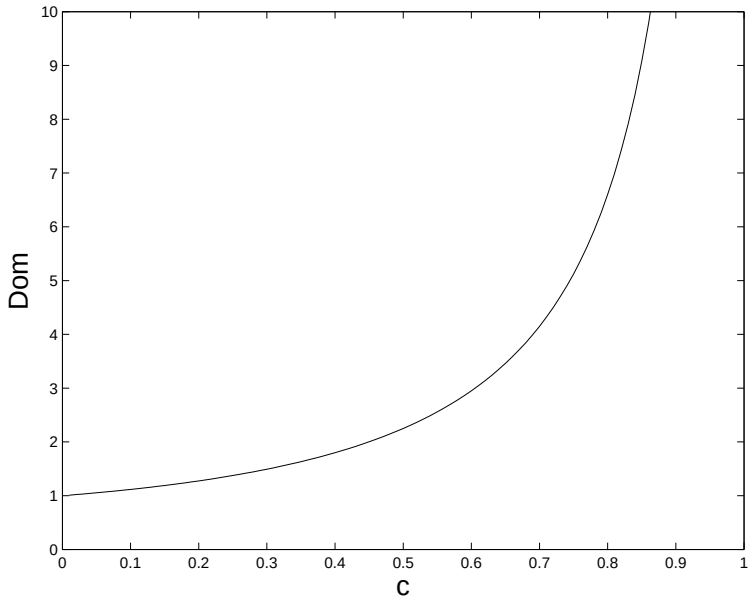

Figure 7: Dominance term Dom of the asymptotic relative efficiency (30) for linkage versus association tests as function of the fraction c of dominance variance at the causal locus.

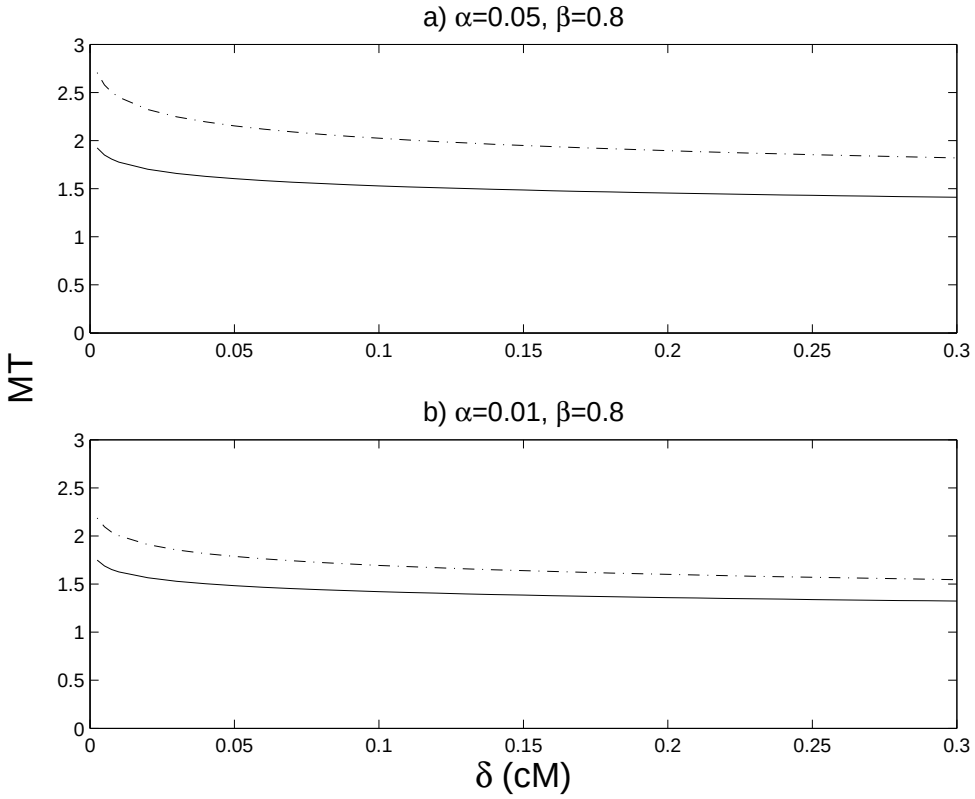

Figure 8: Multiple testing term MT of the asymptotic relative efficiency (30) for linkage versus association tests as function of the spacing δ between effectively independent association tests. Dash-dotted lines correspond to a chromosomewide scan ($C = 1$, $L = 150$ cM) and solid lines to a genomwide scan ($C = 22$, $L = 3575$ cM). For the association tests we use $\tilde{\beta}_1 = \beta$ and $K_1 = 2(L/\delta + C)$ and for the linkage tests the dense marker approximations (A.12) and (A.13) with $\rho = 0.02$ cM^{-1} and $d = 1$.