Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 11

Combined Association and Linkage Analysis for General Pedigrees and Genetic Models

Ola Hössjer*

*University of Stockholm, Sweden, ola@math.su.se

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 including polygenic and shared environmental effects and allows for missing marker data. It is multipoint, taking marker genotypes from several loci into account simultaneously. The score vector has one association and one linkage component, which can be used to define separate tests for association and linkage. For complete marker data, we give closed form expressions for the efficiency of the linkage, association and combined tests. These are examplified 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 test is a robust alternative, which never performs much worse than the best of the linkage and association tests, and sometimes significantly better than both of them. It should be particularly useful when little is known about the genetic model.

KEYWORDS: Association, linkage, multipoint test, noncentrality parameter, score test

*The author wishes to thank two anonymous referees for valuable comments that considerably improved the paper.

1 Introduction

Association and linkage analysis are complementary methods for localizing genes that increase susceptibility to a certain inheritable disease. In linkage analysis, regions are sought for where marker allele transmissions from parents to children are correlated with phenotypes. This is due to presence of crossovers in meioses within families and occurs for all markers linked to the disease locus. In association analysis, one searches regions of non-independence between phenotypes and marker genotypes in a whole population. In absence of population stratification, this is caused a mutation many generations ago. The mutated haplotype has been transmitted and spread to a subset of individuals of the present generation. During this process, crossovers in meioses of previous generations break up association at larger distances from the disease locus, since the mutated haplotype is only intact over small distances around the mutation.

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.

Since association as well as linkage tests use marker and phenotype data from a number of families, one might argue that a combined linkage and association test optimally extracts information from data and hence should have greater power in detecting a disease susceptibility locus. By a joint association and linkage test we mean one that has power even when there is no association between marker and disease genotypes. This is not the case for methods based on transmitted and non-transmitted founder alleles, such as the TDT-tests (Falk and Rubinstein, 1987, Terwilliger and Ott, 1992, Spielman et al., 1993). Hence, we consider the TDT-test and its extensions as pure association tests, even though they can be used for testing linkage in presence of association.

In recent years, some authors have considered joint tests of linkage and association. Xiong and Jin (2000) considered a single point likelihood ratio test, which generalizes the classical lod score for pure linkage analysis. It can be applied to arbitrary pedigree structures and genetic models without polygenic or shared environmental effects. Association is modeled between marker and disease alleles of the founders and linkage is modeled in allele transmissions to nonfounders. Farnir et al. (2002) considered a similar likelihood ratio test for an animal genetic application with quantitative phenotypes and half sib families, allowing for several markers to be analyzed jointly. Zhao et al. (1998) defined a semiparametric estimating equations method, with one linkage and one association component in the score vector, Terwilliger and Göring (2000) developed a method based on pseudomarkers and Fulker et al. (1999), Sham et al. (2000) and Meuwissen et al. (2002) considered variance components tests for quantitative traits. A multipoint Bayesian method for combined linkage and association analysis was defined by Pérez-Enciso (2003). It is defined for arbitrary pedigree structure and quantitative traits, but can in principle be defined for general genetic models.

The focus of this paper is on score tests for linkage and/or association. 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, 2005) 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. Self et al. (1991), Schaid and Sommer (1994), Schaid (1996) and Lunetta et al. (2000) developed family based score tests for association. These tests are conditional on phenotypes and observed founder 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 defined in this paper is valid for arbitrary pedigree structures and missing marker information. It is multipoint, in the sense that it uses information from all markers at the same chromosome as the marker locus being tested for association and linkage. It is based on biologically based models with association between marker and disease causing alleles and penetrance parameters of the disease genes. As in Hössjer (2005), we allow for polygenic effects and a large class of phenotypes. 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. These two score components can also be used for defining separate association and linkage tests. Moreover, very explicit expressions can be derived for the efficiency (noncentrality parameter) of the linkage test, various versions of the association test (with or without conditioning on founder genotypes and in the latter case, with known or estimated marker allele frequencies) as well as for the combined association and linkage test.

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. Asymptotic efficiency is considered in Section 6 for the combined test as well as for the pure association and linkage tests. 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 sample of N outbred pedigrees of arbitrary (and possibly different) form. Based on phenotypes and marker data from all families, we wish to test presence of a disease locus τ along a genomic region Ω, which may consist of (parts of) one or several chromosomes. Marker data is collected at K loci $x_1, \ldots, x_K$ in Ω, assumed to be ordered within each chromosome. For each x_i we wish to test

$$\begin{array}{rl} H_0(x_i): & x_i \text{ is unlinked to } \tau \text{, marker data is not associated to} \\ & \text{disease genotypes at } \tau, \\ H_1(x_i): & x_i = \tau. \end{array} \tag{1}$$

Consider one of the N families and assume it consists of n individuals; f founders and $n - f$ nonfounders. Let $Y = (Y_1, \ldots, Y_n)$ be the vector of phenotypes Y_k of all pedigree members k, including the possibility $Y_k =' ?'$ when k has unknown phenotype. We denote the genotypes at the disease locus τ as $G = (G_1, \ldots, G_n)$, 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. Marker data is written $M = (M_1, \ldots, M_K)$, where $M_i = \{H_{ik};\, k \in \mathcal{T}\}$ is marker data at locus x_i, $\mathcal{T} \subset \{1, \ldots, n\}$ is the set of genotyped family members and $H_{ik} = (b_{i,2k-1} b_{i,2k})$ is the marker genotype of k at locus x_i, which consists of two alleles, one transmitted from the mother and one from the father. Write $H_i = (H_{i1}, \ldots, H_{in})$ and notice that $M_i = H_i$ when all family members are genotyped.

To begin with, we consider a fixed locus $x = x_i$ and construct a pointwise score test for (1), taking marker data from x and all other loci into account. The genetic model parameters are

$$\theta = (q_1, \ldots, q_K, \Delta, \psi, p),$$

where $q_i = (q_{i0}, \ldots, q_{i,d_i-1})$ is the vector of marker allele frequencies of the d_i-allelic marker at x_i, Δ quantifies association between G and $H = H_i$, ψ contains penetrance parameters that quantify association between Y and G, and $p = (p_0, p_1)$ consists of disease allele frequencies at the biallelic disease locus.

The two structural parameters are ψ and Δ, whereas the others are nuisance parameters. The effect of misspecifying p and marker allele frequencies $q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_K$ at loci $x_j \neq x$ is not so serious and hence we can use plug-in estimates of these parameters in the score tests without taking sample variability into account[1]. Hence we reduce the parameter vector to

$$\theta = (q, \Delta, \psi),$$

where $q = q_i = (q_0, \ldots, q_{d-1})$ contains marker allele frequencies for the $d_i = d$-allelic marker at x.

For one family, we use the retrospective likelihood (Prentice and Pyke, 1979), i.e. the conditional probability

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

of observed marker data given phenotypes. 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

[1] In fact, essentially, p only enters in the score test through a certain dominance variance fraction c, defined in Section 5 and $q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_K$ in the multipoint probability $P(b, v|M)$, defined in Section 4.

M), see for instance Kraft and Thomas (2001). For N families, we take the product of the familywise likelihoods (2).

Subscript x in (2) means '$x = \tau$'. As will be seen in Section 4, this is a valid assumption also under the null hypothesis, since all likelihood computations can be made assuming $x = \tau$. Indeed, $H_0(x)$ 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.

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 $b = (b_1, \ldots, b_{2f})$, with $b_k = b_{ik}$, be the vector of founder marker alleles at x (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

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

assuming

$$P_{x,\theta}(M|b, v, Y) = P(M|b, v), \tag{4}$$

that is, given (b, v), marker data at loci different from x are independent of Y and genetic model parameters. This is clear when $M = M_i$ only contains marker data at x. Otherwise, it holds under Haldane's model of no interference for crossovers and when there is no *residual* linkage disequilibrium (LD) between G and markers H_j, $j \neq i$, that is not already accounted for as LD between H and G. This crucial assumption will be further discussed in Section 9.

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 (so that $M_i = H$), 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) for which $P(M|b, v)$ have the same nonzero value, whereas $P(M|b, v) = 0$ for all other pairs. The 2^f pairs compatible with M are obtained by shifting phase of all founders independently (Kruglyak et al., 1996). Hence, for complete marker data,

$$L(x,\theta;M) = 2^f P(M|b,v)P_{x,\theta}(b,v|Y), \tag{5}$$

where (b,v) is any of the 2^f pairs compatible with M. Since $2^f P(M|b,v)$ is independent of θ, it will cancel out when computing score functions for complete marker data. For this reason, it is often more convenient to use

$$L^{\mathrm{c}}(\theta;b,v) = P_{x,\theta}(b,v|Y), \tag{6}$$

where superscript c is short for 'complete' and dependence on x on the LHS is implicit in b and v.

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 $L^{\mathrm{c}}(\theta;b,v)$ by conditioning on the vector $a = (a_1,\ldots,a_{2f})$ of *founder alleles* at the disease locus,

$$\begin{array}{rcl} L^{\mathrm{c}}(\theta;b,v) & = & \sum_a P_{q,\Delta}(b|a)P_{p,\psi}(a,v|Y) \\ & = & C\sum_a P_{p,q,\Delta}(a,b)P_{\psi}(Y|a,v), \end{array} \tag{7}$$

where $C = 2^{-m}/P_{p,\psi}(Y)$ is a proportionality constant that depends on ψ but not on b and v. Hence it will only affect the score function in Section 4 by means of an additive constant. We devote the next section to specifying the association term $P_{p,q,\Delta}(a,b)$ and penetrance term $P_{\psi}(Y|a,v)$ in more detail.

3 Modeling of Association and Penetrance

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

$$\begin{array}{rcl} P_{p,q,\Delta}(a,b) & = & \prod_{j=1}^{2f} P_{a_j,b_j}, \\ P_{a_j,b_j} & = & p_{a_j}q_{b_j}(1+\Delta s(a_j,b_j)), \end{array} \tag{8}$$

where $P_{a_jb_j}$ 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

$$\Delta = \frac{P_{00}P_{11} - P_{01}P_{10}}{(p_0p_1q_0q_1)^{1/2}} \tag{9}$$

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

$$s = \begin{pmatrix} \left(\frac{p_1 q_1}{p_0 q_0}\right)^{1/2} & -\left(\frac{p_1 q_0}{p_0 q_1}\right)^{1/2} \\ -\left(\frac{p_0 q_1}{p_1 q_0}\right)^{1/2} & \left(\frac{p_0 q_0}{p_1 q_1}\right)^{1/2} \end{pmatrix}. \tag{10}$$

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. Other measures of linkage disequilibrium are discussed by Devlin and Risch (1995). □

Since we assume that v is the inheritance vector for loci x *and* τ, it specifies how founder alleles a are spread to all nonfounders, so that G is a deterministic function of a and v. Moreover, the penetrance factor $P_\psi(Y|a,v)$ is a function of a and v only through G. Hence we denote it as $P_\psi(Y|G)$ in the sequel. We write the penetrance vector as $\psi = (\psi(0), \psi(1), \psi(2))$, where $\psi(j)$ is the penetrance factor for 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

$$f(Y,\mu) = \prod_{k=1}^{n} f_k(Y_k;\mu_k), \tag{11}$$

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 (11), define

$$f_k(Y_k;\mu_k) = \mu_k^{(Y_k=1)}(1-\mu_k)^{(Y_k=0)}. \tag{12}$$

□

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

$$f(Y;\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n|\Sigma|^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y-\mu)\Sigma^{-1}(Y-\mu)^T)\right), \tag{13}$$

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 $\mathrm{Corr}(Y|G) = \Sigma = (\Sigma_{kl})$ are independent of 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 (2005), 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 (13), 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 $P(Y|X) = 1_{\{X\in A\}}$, where $A = \{X;\ X_k > z \text{ if } Y_k = 1, X_k \leq z \text{ if } Y_k = 0 \text{ or } X_k \text{ arbitrary if } Y_k =' ?'\}$. □

4 Score functions and tests

Following McPeek (1999) and Hössjer (2003, 2005), we consider a one-dimensional trajectory $\{\psi_\varepsilon\}_\varepsilon$ of penetrance vectors

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

and hence rewrite the parameter vector as

$$\theta = (q, \Delta, \varepsilon). \tag{14}$$

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

$$\begin{array}{ll} H_0(x): & \tau = x, \Delta = \varepsilon = 0, \\ H_1(x): & \tau = x, \varepsilon \neq 0. \end{array}$$

Adding '$\tau = x$' under $H_0(x)$ may seem contradictory in view of (1). However, $\tau = x, \varepsilon = 0$ refers to a 'disease locus' τ with no effect on the phenotypes, which, for all likelihood computations, is mathematically equivalent to a disease locus with effect on phenotypes that is unlinked to τ. With this notation, we make it clear that the parameter spaces of $H_0(x)$ and $H_1(x)$ are disjoint but not isolated from each other and that all likelihood computations can be performed assuming $\tau = x$. Compared to (1), we have added the additional requirement $\varepsilon \neq 0$ under H_1. This means that under H_1, τ is indeed a true disease locus that has some effect on the phenotype.

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. However, it is shown in the appendix that the scores of Δ and ε vanish for outbred pedigrees. For this reason, we reparametrize to

$$\epsilon = (\epsilon_0, \epsilon_1, \epsilon_2), \tag{15}$$

where $\epsilon_0 = q = (q_1, \ldots, q_{d-1})$ contains nuisance parameters, and $\epsilon_1 = \Delta\varepsilon$ and $\epsilon_2 = \varepsilon^2$ are the structural parameters[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 at locus x becomes

$$S(x) = \left. \frac{\partial \log L(x, \epsilon; M))}{\partial \epsilon} \right|_{\epsilon=(q,0,0)} = (S_0(x), S_1(x), S_2(x)), \tag{16}$$

where $S_i(x) = S_i(x; M, Y)$ 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

$$S(x) = \sum_{b,v} P_q(b, v|M) S(b, v), \tag{17}$$

[2] To be exact, ϵ is not a one-to-one function of θ in (14), since (q, Δ, ε) and $(q, -\Delta, -\varepsilon)$ correspond to the same ϵ. However, the likelihood is locally invariant around $(q, 0, 0)$ with respect to this ambiguity. Alternatively, we might impose $\varepsilon \geq 0$ to make the reparametrization a one-to-one mapping.

where, with a slight abuse of notation, $S(b,v) = (S_0(b,v), S_1(b,v), S_2(b,v))$ is the score function for complete marker data, defined as

$$S(b,v) = \left.\frac{\partial \log L^c(\epsilon; b, v))}{\partial \epsilon}\right|_{\epsilon=(q,0,0)}. \qquad (18)$$

With $N > 1$ pedigrees, we simply add the score vectors (17) of each pedigree to obtain a total score vector $S(x)$. The validity of (17) depends crucially on (4), as shown in the appendix.

We consider both the case when q is known and unknown. In the latter case, we have to plug in a maximum-likelihood estimate of q into S_1 and S_2, and this will affect the Fisher information. Define $I_{ij} = E(S_i^T S_j)$, where expectation is under H_0, and $I = (I_{ij})_{i,j=1}^2$. Let $J = (J_{ij})_{i,j=1}^2$ be the 2×2 Fisher information matrix for (ϵ_1, ϵ_2) at the null value $(0,0)$, defined as

$$J = \begin{cases} I, & q \text{ known}, \\ I - (I_{01}, I_{02})^T I_{00}^{-1} (I_{01}, I_{02}), & q \text{ is estimated}. \end{cases} \qquad (19)$$

As shown in the appendix, the combined test statistic for linkage and association, when testing H_0 against H_1 is

$$T_{\text{combined}}(x) = \begin{cases} (S_1(x), S_2(x)) J^{-1} (S_1(x), S_2(x))^T, & \text{if } \tilde{S}_2(x) \geq 0, \\ S_1^2(x)/J_{11}, & \text{if } \tilde{S}_2(x) < 0, \end{cases} \qquad (20)$$

where $\tilde{S}_2(x) = (S_1(x), S_2(x)) J^{-1/2} g^T$, $g = \left((0,1)J^{1/2}\right)^\perp$ and $e^\perp = (-e_2, e_1)$ is a vector orthogonal to $e = (e_1, e_2)$. When J is diagonal, as for complete marker data, we have $\tilde{S}_2(x) = S_2(x)$. The null hypothesis 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, see the appendix and Self and Liang (1987).

The 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

$$\begin{aligned} T_1(x) &= S_1^2(x)/J_{11}, \\ T_2(x) &= S_2(x)/\sqrt{J_{22}}, \end{aligned} \qquad (21)$$

and H_0 is rejected when T_1 or T_2 exceed a given threshold. It is customary in linkage analysis not to account for the influence of estimating q and put

$J_{22} = I_{22}$ in (21) regardless of whether q is estimated or not. The reason is that misspecification of q is not as serious for linkage analysis. Indeed, we show below that $S_2(b, v) = S_2(v)$, and hence $S_2(x) = \sum_v P_q(v|M)S_2(v)$. 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.

For practical purposes, for incomplete marker data, computation of $T_{\text{combined}}(x)$, $T_1(x)$ and $T_2(x)$ requires an explicit expression for $P_q(b, v|M)$ in (17) for each family. In the most general case, when nearby loci x_i are close enough to be in LD, the LD-structure has to modeled and incorporated in order to compute $P_q(b, v|M)$. For sparser marker maps, when nearby x_i are in linkage equilibrium (LE), one may use the multipoint algorithm of Lander and Green (1987) for this purpose, as shown in the appendix.

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 (12), that

$$\omega_k = Y_k - K_p. \tag{22}$$

Further, $\omega_{kl} = \omega_k\omega_l$ when $k \neq l$, and this is general property in absence of polygenic and shared environmental effects (11). If k and l is a monozygotic twin pair, it follows that the relative risk ratio (Risch, 1990) equals

$$\lambda = 1 + \omega_{kl}\varepsilon^2 + o(\varepsilon^2). \tag{23}$$

If $E(\psi(|G_k|)) = 0$, the prevalence is independent of ε and the remainder term in (23) 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 $\sigma_g^2 = 1$,

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

where Σ_{kl}^{-1} is the $(k,l)^{\text{th}}$ entry of Σ^{-1}, see Hössjer (2005). 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 (2005),

$$\begin{aligned}\omega_k &= \sigma_g \int ((X-\mu_0)\Sigma^{-1})_k P(X|Y)dX\\ \omega_{kl} &= \sigma_g^2 \int \left(((X-\mu_0)\Sigma^{-1})_k((X-\mu_0)\Sigma^{-1})_l - \Sigma_{kl}^{-1}\right) P(X|Y)dX,\end{aligned} \tag{24}$$

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 $\sigma_g^2 = (1-K_p)^2K_p^2/\phi^2(z)$, where $\phi = \Phi'$. Then (24) reduces to (22), 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, 2005). It can be shown that (23) 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

$$\begin{array}{rcl} S_0(b,v) = S_0(b) & = & (n_1/q_1 - n_0/q_0, \ldots, n_d/q_d - n_0/q_0), \\ S_1(b,v) = S_1(H) & = & \sqrt{1-c} \cdot \sum_{k=1}^{n} \omega_k (g(b_{2k-1}) + g(b_{2k})) - C_1 \\ S_2(b,v) = S_2(v) & = & \sum_{1 \le k < l \le n} \omega_{kl} \left((1-c)\mathrm{IBD}_{kl}/2 + c 1_{\{\mathrm{IBD}_{kl}=2\}}\right) - C_2, \end{array} \tag{25}$$

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 (25) 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 (8). In the linkage case, T_2 coincides with the test statistic in Hössjer (2005), 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

$$S_1^{\mathrm{NF}}(H) = \sqrt{1-c} \sum_{k=f+1}^{n} \omega_k \left(g^0(b_{2k-1}) + g^0(b_{2k})\right), \tag{26}$$

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 (16) 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

$$T_1^{\mathrm{NF}}(x) = (S_1^{\mathrm{NF}}(x))^2 / J_{11}^{\mathrm{NF}}, \tag{27}$$

where $J^{\mathrm{NF}} = (J_{ij}^{\mathrm{NF}})_{i,j=1}^2$ is defined analogously to J in (19). 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 the 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

$$\eta^2_{\text{combined}} = 2E_\theta \left(\log \frac{L(\tau, \hat{\theta})}{L(\tau, \hat{\theta}_0)} \right) \tag{28}$$

as twice the expected increase of the log likelihood under $\hat{\theta} = (\hat{q}, \Delta, \varepsilon)$ compared to $\hat{\theta}_0 = (\hat{q}, 0, 0)$ and with $\hat{q}$ the ML-estimator of q. When q is known we replace $\hat{\theta}$ and $\hat{\theta}_0$ by $\theta = (q, \Delta, \varepsilon)$ and $\theta_0 = (q, 0, 0)$ respectively. The noncentrality parameter grows with N if θ is kept fixed. However, for θ close to θ_0, likelihood theory implies, under certain regularity assumptions (see (A.1) and (A.9) in the appendix) that the asymptotic approximation

$$\eta^2_{\text{combined}} = (\epsilon_1, \epsilon_2) J (\epsilon_1, \epsilon_2)^T. \tag{29}$$

may be used, provided that the RHS of (29) stays bounded when N grows. For complete marker data, it is shown in the appendix that

$$J = \begin{pmatrix} J_{11} & 0 \\ 0 & I_{22} \end{pmatrix}, \tag{30}$$

which implies

$$\eta^2_{\text{combined}} = \eta_1^2 + \eta_2^2, \tag{31}$$

where

$$\begin{array}{rcl} \eta_1^2 & = & (\Delta\varepsilon)^2 J_{11}, \\ \eta_2^2 & = & \varepsilon^4 I_{22}, \end{array} \tag{32}$$

are the noncentrality parameters of the pure association and linkage tests respectively. The asymptotic distributions of the combined association test and linkage tests at the disease locus are

$$\begin{aligned} T_{\text{combined}}(\tau) &= (X_1+\eta_1)^2 + (X_2+\eta_2)^2 1_{\{X_2+\eta_2\geq 0\}}, \\ T_1(\tau) &\in \chi^2(1,\eta_1^2), \\ T_2(\tau) &\in N(\eta_2,1), \end{aligned} \tag{33}$$

where $\chi^2(p,\eta^2)$ is a noncentral χ^2-distribution with p degrees of freedom and noncentrality parameter η^2 and X_1 and X_2 are independent standard normal random variables. Notice that $T_{\text{combined}}(\tau)$ does *not* have a $\chi^2(2,\eta^2_{\text{combined}})$-distribution, since the linkage parameter ϵ_2 is constrained to be nonnegative.

Formula (32) for η_1 holds for all three versions of the association test, with

$$J_1 = \begin{cases} J_{11}^{\text{km}} = I_{11}, & \text{no founder cond., known } q, \\ J_{11}^{\text{em}} = I_{11} - I_{01}^T I_{00}^{-1} I_{01}, & \text{no founder cond., estimated } q, \\ J_{11}^{\text{NF}} = I_{11}^{\text{NF}}, & \text{founder conditioning,} \end{cases} \tag{34}$$

and superscripts 'km' and 'em' are short for known and 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 $N_1^{\text{em}}(\alpha,\beta)$, $N_1^{\text{km}}(\alpha,\beta)$, $N_1^{\text{NF}}(\alpha,\beta)$, $N_2(\alpha,\beta)$ and $N_{\text{combined}}(\alpha,\beta)$ needed for level α tests T_1^{em}, T_1^{km}, T_1^{NF}, T_2 and T_{combined} to attain power β. Write $N(\alpha,\beta)$ for any of these five quantities and η^2 for any of the four noncentrality parameters $(\eta_1^{\text{em}})^2$, $(\eta_1^{\text{km}})^2$, $(\eta_1^{\text{NF}})^2$ and η_2^2. Then

$$\eta^2 = N\eta^2(1), \tag{35}$$

where $\eta^2(1)$ is corresponding noncentrality parameter when $N=1$.

For a pointwise test T at one single x, we first choose threshold by solving for t in $P(T\geq t|H_0(x)) = \alpha$, where the H_0-distribution of T is obtained by putting $\eta = 0$ in (33). Then $N(\alpha,\beta)$ is found by inserting (35) into (33) and solving for N in $P(T\geq t|H_1(x)) = \beta$.

When linkage and/or association is tested at number of loci x simultaneously, we incorporate multiple testing correction into the definition of $N(\alpha,\beta)$. We have done this for the pure association and linkage tests (but not for the combined test) in the appendix, where it is shown that

$$N(\alpha,\beta) \approx \frac{(\xi_{\tilde{\alpha}} + \xi_{1-\tilde{\beta}})^2}{\eta^2(1)}, \tag{36}$$

with $\xi_\alpha = \Phi^{-1}(1-\alpha)$ the $(1-\alpha)$-quantile of a standard normal $N(0,1)$ random variable. A formula similar to (36) appears in Risch and Merikangas (1996). The numbers $\tilde{\alpha}$ $(= \tilde{\alpha}_1^{\text{em}}, \tilde{\alpha}_1^{\text{km}}, \tilde{\alpha}_1^{\text{NF}}$ or $\tilde{\alpha}_2)$ and $\tilde{\beta}$ $(= \tilde{\beta}_1^{\text{em}}, \tilde{\beta}_1^{\text{km}}, \tilde{\beta}_1^{\text{NF}}$ 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 $\tilde{\beta} \le \beta$. 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' $\tilde{K}$ of one-sided independent tests. The Bonferroni approximation for small α is $\tilde{\alpha} \approx \alpha/\tilde{K}$, 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 $\tilde{K}_1 = 2$ and $\tilde{K}_2 = 1$ when one locus is tested, where $\tilde{K}_1$ is any of $\tilde{K}_1^{\text{em}}$, $\tilde{K}_1^{\text{km}}$ or $\tilde{K}_1^{\text{NF}}$. When several loci are tested, $\tilde{K}_1/2$ and $\tilde{K}_2$ correspond to the effective number of 'independent loci' for association and linkage tests. When test statistics at different loci are independent, both $\tilde{K}_1/2$ and $\tilde{K}_2$ equal the number K of *actual* loci x_i. In general though they may be smaller than K due to dependency of test statistics at nearby loci. Since linkage disequilibrium decays faster than correlation of allele sharing statistics, $\tilde{K}_1/2$ is in general larger than $\tilde{K}_2$ and hence the numerator of (36) is larger for the association tests than for the linkage test.

Formula (36) can be generalized samples 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 test to attain power β is

$$N(\alpha,\beta) \approx \frac{(\xi_{\tilde{\alpha}} + \xi_{1-\tilde{\beta}})^2}{\int \eta^2(\phi)d\nu(\phi)}, \tag{37}$$

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.

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 (9). 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

$$\begin{array}{rcl} J_{11}^{\mathrm{km}} & = & 0.5 \cdot (1-c)\,((\sum_{k=1}^{n} \omega_k)^2 + (\omega_1 - \omega_2)^2 + \sum_{k=3}^{n} \omega_k^2) \\ J_{11}^{\mathrm{em}} & = & 0.5 \cdot (1-c)\,((\omega_1 - \omega_2)^2 + \sum_{k=3}^{n} \omega_k^2) \\ J_{11}^{\mathrm{NF}} & = & 0.5 \cdot (1-c) \sum_{k=3}^{n} \omega_k^2, \\ I_{22} & = & 0.125 \cdot (1 + 0.5 \cdot c^2) \sum_{3 \le k < l \le n} \omega_{kl}^2 \end{array} \qquad (38)$$

and this in turn gives explicit expressions for $N(\alpha, \beta)$, using formulas of the previous section.

8 Numerical Results

We evaluated the required sample size $N(\alpha, \beta)$ in a genomewide scan for different pedigree types and genetic models. Since (36) 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^{em} and

T_1^{NF} for larger pedigrees. This is not surprising, since I_{22} grows quadratically with pedigree size, whereas J_{11}^{em} and J_{11}^{NF} grow linearly with pedigree size. Among the three association tests, T_1^{km} is most efficient, followed by T_1^{em} and T_1^{NF}. This is evident from the figures, but can also be seen by comparing the Fisher informations in (38). (However, T_1^{km} is also most sensitive to model misspecification, followed by T_1^{em} and T_1^{NF}.) When parents have unknown phenotypes ($\omega_1 = \omega_2 = 0$), T_1^{em} and T_1^{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^{km} and T_1^{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. When c increases, the linkage test gets comparatively more efficient than the association tests, as can be seen from (38). This effect is small for when $c \leq 0.1$ but then rapidly gets more pronounced as c increases, see Hössjer (2004a) for details.

We also investigated the effect of varying δ in the range 0.01-0.3 cM in Hössjer (2004a). The conclusion is that $N_1(\alpha, \beta)$ is quite insensitive to the choice of δ. Hence the N_1-curves of Figures 1-6 are quite representative for a range of values of δ.

To evaluate the performance of the combined test, we computed the relative efficiency

$$N(\alpha, \beta)/\min(N_1(\alpha, \beta), N_2(\alpha, \beta), N_{\mathrm{combined}}(\alpha, \beta)) \qquad (39)$$

as function of η_2/η_1 for the combined, association and linkage tests in Figure 7. In this case, all sample sizes are for pointwise tests, using (33). It is evident from this figure that the combined test is very robust, and never performs much worse than the best of the association and linkage tests. On the other hand, for a range of intermediate values of η_2/η_1, the combined test is more efficient than both the association and linkage tests, since the extra cost of adding degrees of freedoms to the combined test is less than the gain of using more information from marker data.

We conjecture that the conclusions of Figure 7 extend to multiple testing over one or several chromosomes. For instance, it is reasonable to assume that the amount of multiple testing is about the same for the association test and combined tests, setting $\tilde{\alpha}$ and $\tilde{\beta}$ to the same values as for both tests in Figures 1-6. Hence, we conclude T_{combined} is a robust alternative to using either of T_1 or T_2. It is preferable to use the combined test when little is known about the genetic model. Only if prior knowledge of the genetic model is available that makes either of T_1 and T_2 much more powerful than the other, is it reasonable to use the best of these two tests.

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 and multipoint analysis 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. We have also defined separate tests for linkage and association, by using either of the two components of the combined score test separately.

The derived score tests are semiparametric in the sense described in Hössjer (2003b, 2005). By this we mean that they depend on the dominance variance fraction c and a few other genetic model parameters (such as prevalence and polygenic heritability) included in the weights ω_k and ω_{kl}. The latter can often be estimated from population data, and $c = 0$ is a good approximation when the disease allele frequency p_1 is reasonably small. However, we stress that misspecification of these parameters does not affect the significance level of the tests, only the power.

Our framework facilitates efficiency comparisons between the combined, association and linkage tests, 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 with correction for multiple testing. When comparing linkage and association, no method is uniformly superior, but association tests are comparatively more efficient than linkage tests for weak penetrance and strong association models, smaller families and less extreme (quantitative) phenotypes. The linkage methods are comparatively more informative when there is strong penetrance, weak association, larger families and extreme phenotypes. These results could be used to choose between linkage or association when some prior knowledge of the genetic model is available. The combined tests is a robust alternative, with performance close to the best of the linkage and association tests, and sometimes better. Hence it is a useful altnernative, in particular when little is known about the genetic model.

A natural continuation is to extend the analytical efficiency and power comparisons between linkage and various association methods to d-allelic markers with $d > 2$. Likewise, the effect of incomplete marker data should be studied. Incomplete marker data arises because of untyped family members, unknown haplotype phase, sparse marker maps or non-polymorphic markers. To a large extent, efficiency and power comparisons for incomplete marker data will employ simulation, since explicit analytical expressions for noncentrality parameters are hard to obtain except in certain special cases.

In the efficiency comparisons, we utilized large sample normal approximations to determine pointwise significance level and power. To adjust the significance level (and power) for multiple testing, we used analytically tractable methods that are fast to compute. For the association tests, we introduced an 'effective distance between independent loci' as a varying parameter and for the linkage test we used extreme value theory for Gaussian processes. Several modifications of the multiple testing correction are possible. First, it is possible the refine the Gaussian approximation (for the linkage test) by adjusting for non-normality and finite marker spacing (Feingold et al. 1993, Ängquist and Hössjer, 2005). Secondly, simulation could be used as a more accurate but time-consuming alternative. For the linkage test, one may use importance sampling (Ängquist and Hössjer, 2004). Another option is to employ permutation testing. It has the advantage of being applicable to any of the linkage, association and combined tests, both for pointwise and genomewide significance calculations. On the other hand, it requires exchangeability between all families or between subsets of families.

In the multipoint analysis, the crucial assumption (4) on marker and phenotype data is natural to use i) for sparse marker maps with all marker loci in LE and ii) for dense marker maps with polymorphic (haplotype) markers, so that the marker closest to τ captures essentially all LD with G, although there may be other markers nearby. Hence, we believe our joint association and linkage test is well suited for dense marker maps, although haplotype markers (e.g. ones with a certain number B of SNPs) are then preferred a) to utilize all available association between disease genotypes and markers to increase power and b) to make the assumption (4) for deriving the score test correct.

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

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

instead of (28), so our our results are not directly comparable.

Several other extensions are possible. For instance, the assumption of Hardy-Weinberg equilibrium could be relaxed and 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, 2005). For recessive diseases, it is also of interest to consider inbred pedigrees. Then the first order linkage score $\partial \log L/\partial\varepsilon$ will no longer vanish, requiring a reparametrization different from (15). Examples of linkage score functions for inbred pedigrees can be found in McPeek (1999) and Hössjer (2003b, 2005).

Appendix

Derivation of incomplete marker data score functions. Let $\dot{P}_{x,\epsilon}(b,v|Y) = \partial P_{x,\varepsilon}(b,v|Y)/\partial\epsilon$. Then

$$\begin{aligned} S(x) &= \left.\frac{\sum_{b,v} P(M|b,v)\dot{P}_{x,\epsilon}(b,v|Y)}{\sum_{b,v} P(M|b,v)P_{x,\epsilon}(b,v|Y)}\right|_{\epsilon=(q,0,0)} \\ &= \frac{\sum_{b,v} P(M|b,v)P_q(b,v)S(b,v)}{\sum_{b,v} P(M|b,v)P_q(b,v)} \\ &= \sum_{b,v} P_q(b,v|M)S(b,v), \end{aligned}$$

where in the first equality we used (4) and in the second (6), (18) and $P_{x,\epsilon}(b,v|Y) = P_q(b,v)$ when $\epsilon = (q,0,0)$. □

Motivation of combined score test. We omit x and ϵ_0 in the notation, so that $\epsilon = (\epsilon_1,\epsilon_2)$, $L(x,\epsilon) = L(\epsilon)$ and $S = (S_1,S_2) = (S_1(x),S_2(x))$. We consider a local (contiguous) sequence of true parameter values, meaning that the true value of ϵ is such that $\epsilon J\epsilon^T$ stays bounded when N (and hence also J) grows. We further assume that the likelihood surface is smooth enough to admit a second order asymptotic Taylor expansion

$$2\log(L(\epsilon)/L(0,0)) = 2S\epsilon^T - \epsilon J\epsilon^T + o_p(\epsilon J\epsilon^T) \qquad \text{(A.1)}$$

as $N \to \infty$, which is valid for all ϵ (not just the true parameter vector) such that $\epsilon J\epsilon^T$ stays bounded when N grows. The remainder term $o_p(\epsilon J\epsilon^T)$ is stochastic and of smaller order than $\epsilon J \epsilon^T$. Essentially, (A.1) requires that the likelihood, before reparametrization, has partial derivatives (mixed or not) with respect to Δ and ε up to order five, see Rotnitzky et al. (2000). We will derive the combined score test as an asymptotic approximation of

$$2\log(L(\hat{\epsilon})/L(0,0)), \tag{A.2}$$

where $\hat{\epsilon}$ is the ML-estimator of ϵ, i.e. the maximizer of $L(\epsilon_1, \epsilon_2)$ over the region $\{(\epsilon_1, \epsilon_2),\, \epsilon_2 \geq 0\}$, or a bounded subset thereof[3]. Define the two unit vectors $e = (1,0)J^{1/2}/|(1,0)J^{1/2}| = (e_1, e_2)$, $f = e^{\perp} = (-e_2, e_1)$ and the orthogonal 2×2 matrix $Q = (e^T, f^T)$. A reparametrization $\tilde{S} = SJ^{-1/2}Q = (\tilde{S}_1, \tilde{S}_2)$ and $\tilde{\epsilon} = \epsilon J^{1/2} Q$ gives

$$2\log(L(\tilde{\epsilon})/L(0,0)) = 2\tilde{S}\tilde{\epsilon}^T - \tilde{\epsilon}\tilde{\epsilon}^T + o_p(|\tilde{\epsilon}|^2), \tag{A.3}$$

which is maximized over[4] $\{\tilde{\epsilon} = (\tilde{\epsilon}_1, \tilde{\epsilon}_2);\, \tilde{\epsilon}_2 \geq 0\}$. Asymptotically, as $N \to \infty$, we may ignore the remainder term and notice that the ML-estimator of $\tilde{\epsilon}$ is

$$\hat{\tilde{\epsilon}} = \begin{cases} (\tilde{S}_1, \tilde{S}_2), & \tilde{S}_2 \geq 0, \\ (\tilde{S}_1, 0), & \tilde{S}_2 < 0. \end{cases}$$

Insertion into (A.3) yields

$$2\log(L(\hat{\tilde{\epsilon}})/L(0,0)) = \begin{cases} \tilde{S}_1^2 + \tilde{S}_2^2, & \tilde{S}_2 \geq 0, \\ \tilde{S}_1^2, & \tilde{S}_2 < 0. \end{cases} \tag{A.4}$$

Since the likelihood ratio is invariant with respect to reparametrization, the LHS of (A.4) coincides with (A.2). From the definition of $\tilde{S}$, we notice that the RHS of (A.4) is equivalent to (20). □

Expanding multipoint probabilities. Recall $x = x_i$ and that b and v are founder allele and inheritance vectors at x_i. When all markers belong to the same chromosome, let $M^{(i-)} = (M_1, \ldots, M_{i-1})$ and $M^{(i+)} = (M_{i+1}, \ldots, M_K)$. Then, in the case of LE between markers, the multipoint probability $P_q(b, v|M)$, can be written

$$P_q(b, v|M) \propto P_q(b, v, M) = \alpha_i(v)\beta_i(v)P(b)P(M_i|b, v)$$

[3] This is because Δ is bounded and, depending on the application, ε is either bounded or not.

[4] Even if the parameter space of ϵ is bounded, J grows with N, and hence asymptotically, this is the parameter space of $\tilde{\epsilon}$.

where $\alpha_i(v) = P(M^{(i-)}, v)$ are forward probabilities, updated recursively with respect to i from left to right, and $\beta_i(v) = P(M^{(i+)}|v)$ backward probabilities, updated recursively with respect to i from right to left. See Lander and Green (1987) for details. The term $P(M_i|b, v)$ is computed as in Appendix A of Kruglyak et al. (1996). □

Derivation of complete marker data score functions. To begin with, we use parameters (q, Δ, ε), as in (14), to motivate why a reparametrization (15) is appropriate. We regard a in (7) as hidden data and define the full data likelihood as

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

where superscript f short for 'full data' and the proportionality constant $C = 2^{-m}/P_{p,\psi}(Y)$ will only affect the score function by means of an additive (centering) constant.

Then, in view of (7), the complete marker data likelihood (6) satisfies

$$L^{\mathrm{c}}(q, \Delta, \varepsilon) = E(L^{\mathrm{f}}(q, \Delta, \varepsilon)|b, v). \qquad \text{(A.5)}$$

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 the complete marker data likelihood L^{c}. It follows, by differentiating (A.5), that

$$S_{ijk}(b, v) = E(S^{\mathrm{f}}_{ijk}(a, b, v)|b, v). \qquad \text{(A.6)}$$

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 (8) and (A.6) that $S_{100}(b, v) = S^{\mathrm{f}}_{100}(a, b, v)$ is identical to $S_0(b)$ in (25). Since $L^{\mathrm{f}}(q, \Delta, 0) = 2^{-m}P_{q,\Delta}(a, b)$, it follows from (A.5) that $L(q, \Delta, 0) = 2^{-m}P_q(b)$ is independent of Δ. Hence $S^{\mathrm{f}}_{0j0}(b, v) = 0$ for all $j > 0$. Further,

$$\begin{array}{rcl} S^{\mathrm{f}}_{011}(a, b, v) & = & \sigma_g^{-1} \sum_{j=1}^{2f} \sum_{k=1}^{2n} s(a_j, b_j)\omega_k u(|G_k|) - C \\ S^{\mathrm{f}}_{001}(a, b, v) & = & \sigma_g^{-1} \sum_{k=1}^{n} \omega_k u(|G_k|) - C \\ S^{\mathrm{f}}_{002}(a, b, v) & = & 2\sigma_g^{-2} \sum_{1\le k<l\le n} \omega_{kl} u(|G_k|)u(|G_l|) \\ & + & \sigma_g^{-2} \sum_{k=1}^{n} \omega_{kk} u^2(|G_k|) - C. \end{array} \qquad \text{(A.7)}$$

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 (8)), 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 (25). 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_{001}^f nor S_{002}^f depend on b, and we can apply results from Hössjer (2005) 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

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

A reparametrization (15) and (18) shows that indeed S_i in (25) are valid score functions for complete marker data. □

Proof of (30) and (34) for complete marker data. To prove (30), it suffices to verify that $I_{02} = (0,\ldots,0)^T$ and $I_{12} = 0$. 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 (25) 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

$$I_{12} = \sqrt{1-c}\cdot\sum_{k=1}^{n}\sum_{1\le k'<l'\le n}\left(\mathrm{Cov}(g(b_{2k-1}), S_{k'l'}(v)) + \mathrm{Cov}(g(b_{2k}), S_{k'l'}(v))\right). \tag{A.8}$$

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.8) $I_{12} = 0$ follows. Finally, (34) 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. □

Pointwise asymptotic distribution of test statistics, complete marker data. Omitting ϵ_0 and S_0 in the notation, we put $\epsilon = (\epsilon_1,\epsilon_2)$ and $S = (S_1,S_2) = (S_1(\tau),S_2(\tau))$. Assume that the likelihood surface is smooth enough so that (A.1) holds and the Central Limit Theorem can be applied to deduce that asymptotically,

$$S \in N(\epsilon J, J) \tag{A.9}$$

where ϵ is the true parameter vector, assumed to be contiguous, i.e. $\epsilon J \epsilon^T$ stays bounded as N grows. See Rotnitzky et al. (2000) for a discussion on regularity conditions admitting (A.9). By the definition of $\tilde{S} = (\tilde{S}_1, \tilde{S}_2)$ above (A.3), (A.9) is equivalent to

$$\tilde{S} \in N(\tilde{\epsilon}, I_2) = N(\epsilon J^{1/2} Q, I_2), \tag{A.10}$$

where I_2 is a 2×2 identity matrix.

For complete marker data, we use (30) and (32) to conclude that (A.10) is equivalent to

$$\begin{aligned} \tilde{S}_1 &= S_1(\tau)/\sqrt{J_{11}} = X_1 + \eta_1, \\ \tilde{S}_2 &= S_2(\tau)/\sqrt{J_{22}} = X_2 + \eta_2, \end{aligned}$$

with X_1 and X_2 independent standard normal random variables. Then (33) follows from (21) and representation (A.4) of T_{combined}. □

Required sample size (36) for one type of pedigree for association and linkage tests. Let $T(x)$ be any of $T_1^{\text{km}}(x)$, $T_1^{\text{em}}(x)$, $T_1^{\text{NF}}(x)$ and $T_2(x)$ and $t = t(\alpha)$ 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

$$\begin{aligned} P\left(\max_{1 \le i \le K} T(x_i) \ge t | H_0\right) &= 1 - P\left(T < t | H_0\right)^{K'} = \alpha \\ P\left(\max_{i; x_i \in \tilde{\Omega}} T(x_i) \ge t | H_1\right) &= \beta \end{aligned} \tag{A.11}$$

respectively, where $H_0 = \cap_{i=1}^{K} H_0(x_i)$, K' is the effective number of independent pointwise tests, $H_1 = H_1(\tau)$ 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 K of marker loci due to dependence of test statistics. Letting $\tilde{K}$ denote the effective number of independent *one-sided* tests, we have $K' = \tilde{K}$ when $T = T_2$ and $K' = \tilde{K}/2$ for the three association tests T_1^{km}, T_1^{em} and T_1^{NF}. Starting with the first equation in (A.11), formula (33), with $\eta_1 = \eta_2 = 0$, implies that $\tilde{\alpha}$, the pointwise one-sided significance level satisfies

$$\tilde{\alpha} = \begin{cases} P(T(x) \ge t | H_0)/2 = 1 - \Phi(\sqrt{t}), & T = T_1, \\ P(T(x) \ge t | H_0) = 1 - \Phi(t), & T = T_2, \end{cases} \tag{A.12}$$

where T_1 is any of the three association tests. The first equation of (A.11) then implies

$$\tilde{\alpha} = \begin{cases} (1-(1-\alpha)^{2/\tilde{K}})/2, & T = T_1, \\ 1-(1-\alpha)^{1/\tilde{K}}, & T = T_2. \end{cases}$$

Solving for t we find

$$t = \begin{cases} \xi_{\tilde{\alpha}}^2, & T = T_1, \\ \xi_{\tilde{\alpha}}, & T = T_2. \end{cases} \tag{A.13}$$

For the power calculations, we assume a) $x_i = \tau$ for some $i = 1, \ldots, K$ and b) that the power β in the second equation of (A.11) can be written as a function of the noncentrality parameter η^2 at τ, where η is defined in (32) ($\eta = \eta_1$ for the association tests and $\eta = \eta_2$ for the linkage test). 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 $N = 1$. From this (32) follows, which inserted into (33) gives the pointwise power[5]

$$\tilde{\beta} = P(T(\tau) \geq t|H_1) = \begin{cases} 1 - \Phi(\sqrt{t} - \sqrt{N}\eta(1)), & T = T_1, \\ 1 - \Phi(t - \sqrt{N}\eta(1)), & T = T_2. \end{cases} \tag{A.14}$$

It satisfies $\tilde{\beta} \leq \beta$ because of (A.11) and (A.14). 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.11), or equivalently, solving for N in (A.14). Using the latter approach (36) follows. □

Required sample size 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 (36) we find

$$N(\alpha, \beta) = \frac{(\xi_{\tilde{\alpha}} + \xi_{1-\tilde{\beta}})^2}{\sum_{j=1}^{N(\alpha,\beta)} \eta^2(\phi_j)/N(\alpha, \beta)}. \tag{A.15}$$

When $N(\alpha, \beta)$ is large, the denominator of (A.15) is close to $\int \eta^2(\phi) d\nu(\phi)$ by the law of large numbers and this leads to (37). □

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

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 $\tilde{K} = C + 2\rho L t^2$, 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 (2005) for details. According to (A.11), t is the solution of

$$\Phi(t)^{C+2\rho L t^2} = 1 - \alpha. \tag{A.16}$$

and then $\tilde{\alpha}$ is computed from (A.13). For power, we use the asymptotic approximation

$$\beta = \tilde{\beta} + \phi(t-\eta)\left(\frac{2}{\eta d} - \frac{1}{\eta(2d-1)+t}\right), \tag{A.17}$$

of Feingold et al. (1993, formula (A.8)), where $\tilde{\beta}$ is defined in (A.14), $\phi = \Phi'$ is the standard normal density and d a constant that is close to 1 for most pedigrees (Hössjer, 2003c). 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}, T_2 when marker data is complete. 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

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

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

$$I_{11} = (1-c) \cdot \sum_{k,l=1}^{n} \omega_k \omega_l r_{kl}. \tag{A.18}$$

When $d = 2$, write $S_0 = (n_1 - E(n_1))/(q_0 q_1)$. Since for any $k \in \{1, \ldots, 2f\}$, $\text{Cov}(g(b_k), n_1) = (g(1) - g(0))\text{Cov}(b_k, n_1) = (g(1) - g(0))\text{Var}(b_k) = \sqrt{q_0 q_1/2}$, we get $I_{01} = \sqrt{1-c}/(q_0q_1) \cdot \sum_{k=1}^{n} \left(2\omega_k\sqrt{q_0q_1/2}\right) = \sqrt{2(1-c)/q_0q_1}\sum_{k=1}^{n}\omega_k$. Combining this with $I_{00} = 2f/(q_0q_1)$, it follows that

$$I_{01}^2/I_{00} = (1-c)(\sum_{k=1}^{n}\omega_k)^2/f. \tag{A.19}$$

Define S_{kl} as in (A.8), 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

$$I_{22} = \sum_{kl,k'l'} \omega_{kl}\omega_{k'l'}\Sigma_{kl,k'l'}. \tag{A.20}$$

In particular, for a nuclear family with two parents ($k = 1,2$) and $n-2$ children ($k = 3,\dots,n$), we have $f = 2$, $r_{kk} = 1$, $r_{12} = r_{21} = 0$ and $r_{kl} = 0.5$ for all other k,l with $k \ne l$. 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.18)-(A.20) simplify to

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

which in turn implies the first two and the fourth equations in (38). □

Fisher information for $T_1^{\mathbf{NF}}$ when marker data is complete. Let $1 \le k \le 2n$ be a given allele and $1 \le j_k = j_k(v) \le 2f$ the founder allele that is transmitted to k. Introduce $p_{kj} = P(j_k(v) = j)$ and, for any pair $1 \le k,l \le 2n$ 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

$$E(b_k^0 b_l^0) = \alpha_{kl}q_0q_1. \tag{A.21}$$

Combining (A.21) with the definition (26) of the nonfounder score, it follows, for complete marker information, that

$$\begin{array}{rcl} I_{11}^{\mathrm{NF}} & = & (1-c)\sum_{k,l=1}^{n}\omega_k\omega_l\left(E(g^0(b_{2k-1})g^0(b_{2l-1})) + E(g^0(b_{2k-1})g^0(b_{2l}))\right. \\ & & \left.+E(g^0(b_{2k})g^0(b_{2l-1})) + E(g^0(b_{2k})g^0(b_{2l}))\right) \\ & = & (1-c)(g(1)-g(0))^2\sum_{k,l=1}^{n}\omega_k\omega_l\left(E(b_{2k-1}^0b_{2l-1}^0) + E(b_{2k-1}^0b_{2l}^0)\right. \\ & & \left.+E(b_{2k}^0b_{2l-1}^0) + E(b_{2k}^0b_{2l}^0)\right) \\ & = & 0.5(1-c)\sum_{k,l=f+1}^{n}\omega_k\omega_l(\alpha_{2k-1,2l-1} + \alpha_{2k-1,2l} + \alpha_{2k,2l-1} + \alpha_{2k,2l})), \end{array} \tag{A.22}$$

where, in the last step, we used that $\alpha_{kl} = 0$ if either k or l is a founder allele, i.e. if either $1 \le k \le 2f$ or $1 \le l \le 2f$. For a nuclear family, it is easy to see that $\alpha_{kl} = 0.5 \cdot 1_{\{k=l\}}$ for a nonfounder pair $2f + 1 = 5 \le k, l \le 2n$ of alleles. Hence (A.22) 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 (38). □

References

Ängquist, L. and Hössjer, O. (2004). Using importance sampling to improve simulation in linkage analysis. *Statistical Applications of Genetics and Molecular Biology*, **3**(1), article 5.

Ängquist, L. and Hössjer, O. (2005). 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. *Hum. Hered.* **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. *Genet. Epidemiol.* **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. *Theor. Pop. Biol.* **23**, 34-64.

Pérez-Enciso, M. (2003). Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: A Bayesian unified framework. *Genetics* **163**, 1497-1510.

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

Farnir, F. et al. (2002). Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: Revising the location of a quantitative trait locus with major effect on milk production on Bovine chromosome 14. *Genetics* **161**, 275-287.

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. *J. Amer. Statist. Assoc.*, **98**, 1035-1051.

Hössjer, O. (2003c). Spectral decomposition of score functions in linkage analysis. Tentatively accepted by *Bernoulli.*

Hössjer, O. (2004a). Combined association and linkage analysis for general pedigrees and genetic models. Research Report 2004:11, Mathematical Statistics, Stockholm University.

Hössjer, O. (2004b). Information and effective number of meioses in linkage analysis. *J. Math. Biol.* **50**(2), 208-232.

Hössjer, O. (2005). Conditional likelihood score functions in linkage analysis. *Biostatistics* **6**, 313-332.

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.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347-1363.

Kurbasic, A. and Hössjer, O. (2005). Relative risks for general phenotypes and genetic models. Mathematical Statistics, Stockholm University, Research Report 2005:4.

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.S. and Green, P. (1987). Construction of multilocus genetic maps in humans. *Proc. Natl. Acad. Sci. USA*, **84**, 2363-2367.

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

Lunetta, K., Faraone, S.V., Biederman, J. and Laird, N.M. (2000). Family-based test of association and linkage that use unaffected sibs, covariates and interactions. *Am. J. Hum. Genet.*, **66**, 605-614.

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

Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I. and Goddard, M.E. (2002). Fine mapping of quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**, 373-379.

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.

Rotnitzky, A., Cox, D.R., Bottai, M. and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli* **6**, 243-284.

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.

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.

Self, S.G., Longton, G., Kopecky, K.J. and Liang, K-Y. (1991). On estimating HLA/disease association with application to a study of plastic anemia. *Biometrics* **47**, 53-62.

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.

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. *Hum. Biol.* **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.

Xiong, M. and Jin, L. (2000). Combined linkage and linkage disequilibrium mapping for genome screens. *Genet. Epidemiol.* **19**, 211-234.

Zhao, L.P., Aragaki, C., Hsu, L. and Quiaoit, F. (1998). Mapping complex traits by single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **63**, 225-240.

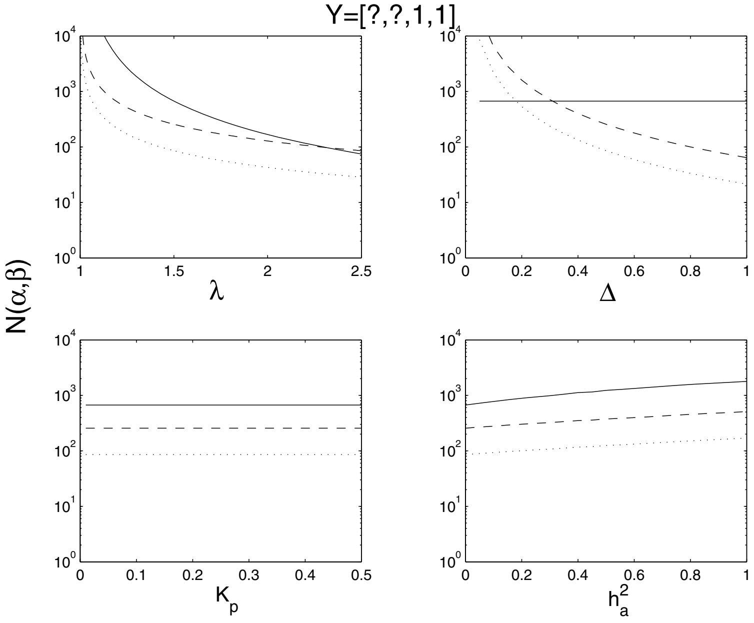

Figure 1: The number of affected sib pairs ($Y = (?, ?, 1, 1)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan. The curves correspond to T_1^{km} (dotted), T_1^{em} and T_1^{NF} (dashed) and T_2 (solid). Only one parameter is varied, 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 multiple testing correction, we assume $C = 22$ chromosomes of total length $L = 3575$ 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.16) and (A.17) with $\rho = 0.02$ cM^{-1} and $d = 1$.

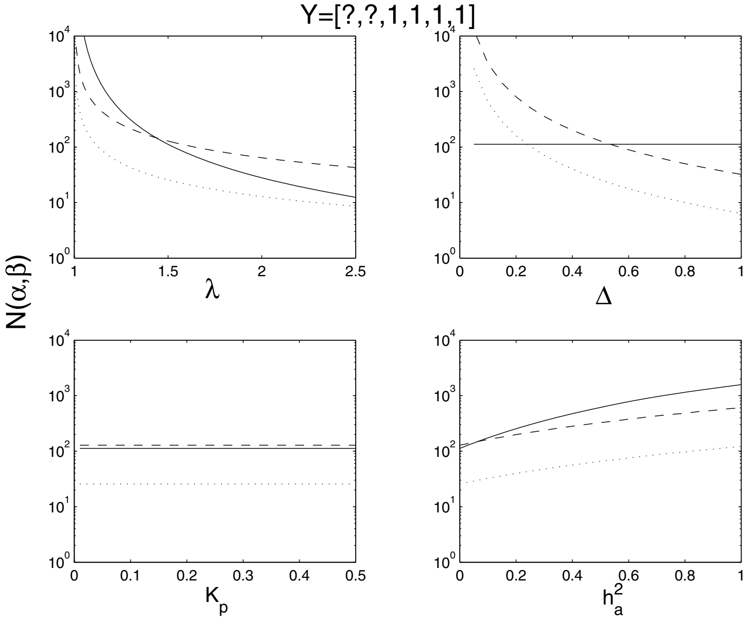

Figure 2: The number of affected sib quartets ($Y = (?, ?, 1, 1, 1, 1)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan. The curves correspond to T_1^{km} (dotted), T_1^{em} and T_1^{NF} (dashed) and T_2 (solid). See Figure 1 for further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

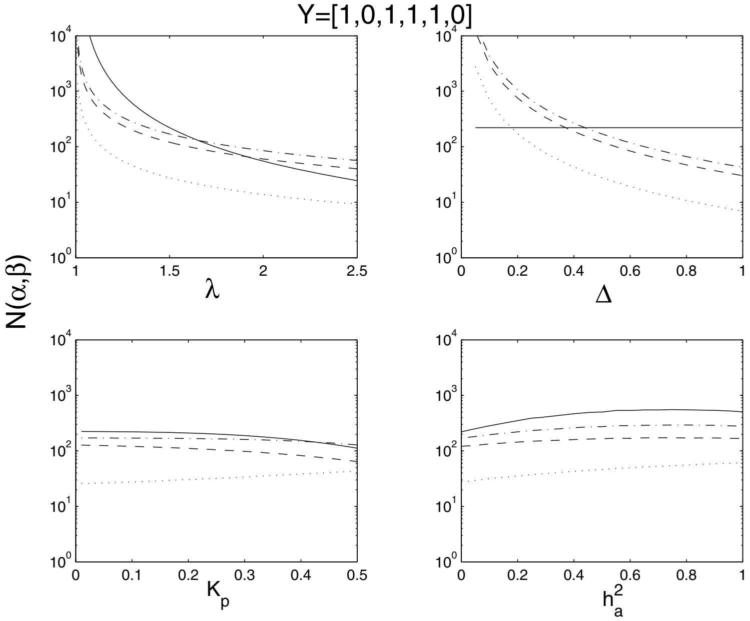

Figure 3: The number of nuclear families with binary phenotypes $Y = (1, 0, 1, 1, 1, 0)$ required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan. The curves correspond to T_1^{km} (dotted), T_1^{em} (dashed), T_1^{NF} (dash-dotted) and T_2 (solid).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

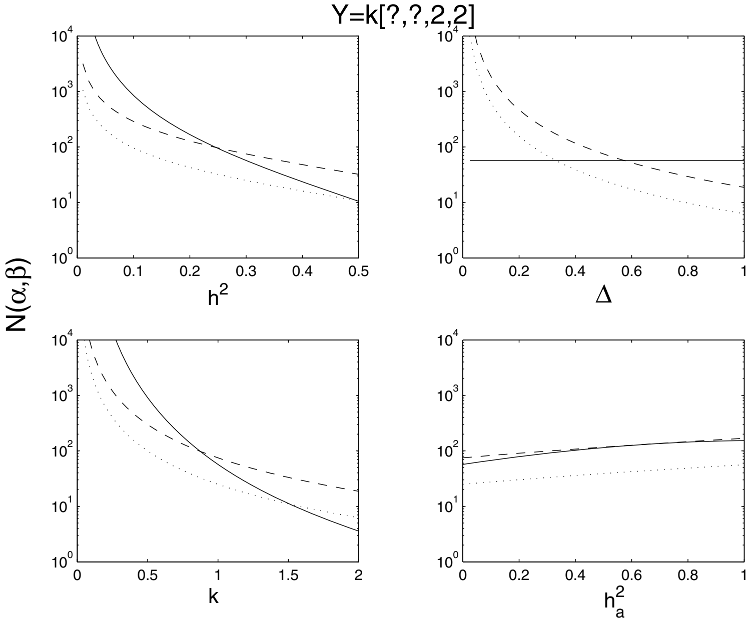

Figure 4: The number of concordant sib pairs (quantitative phenotypes, $Y = k(?, ?, 2, 2)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan. The curves correspond to T_1^{km} (dotted), T_1^{em} and T_1^{NF} (dashed) and T_2 (solid). Only 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. See Figure 1 for details on multiple testing correction.

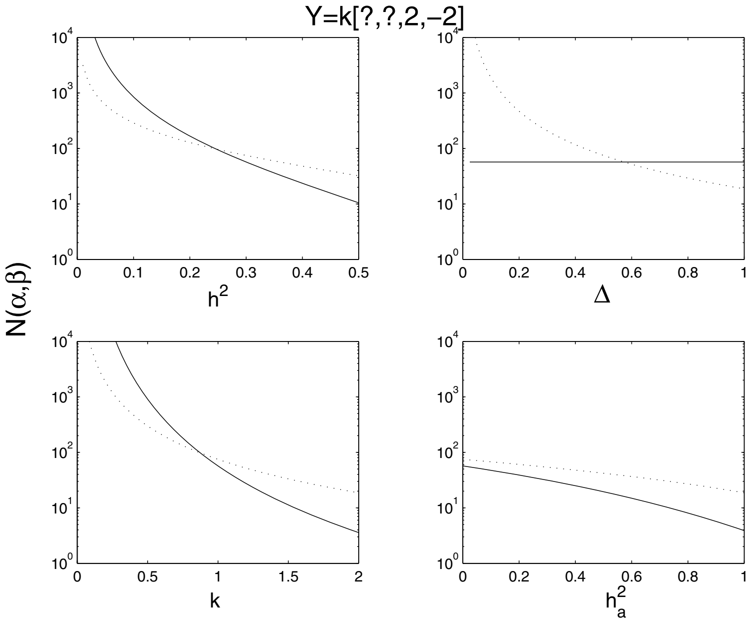

Figure 5: The number of discordant sib pairs (quantitative phenotypes, $Y = k(?, ?, 2, -2)$) required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan. The curves correspond to T_1^{km}, T_1^{em} and T_1^{NF} (dotted) and T_2 (solid). See Figures 1 and 4 for details on multiple testing correction and parameter values.

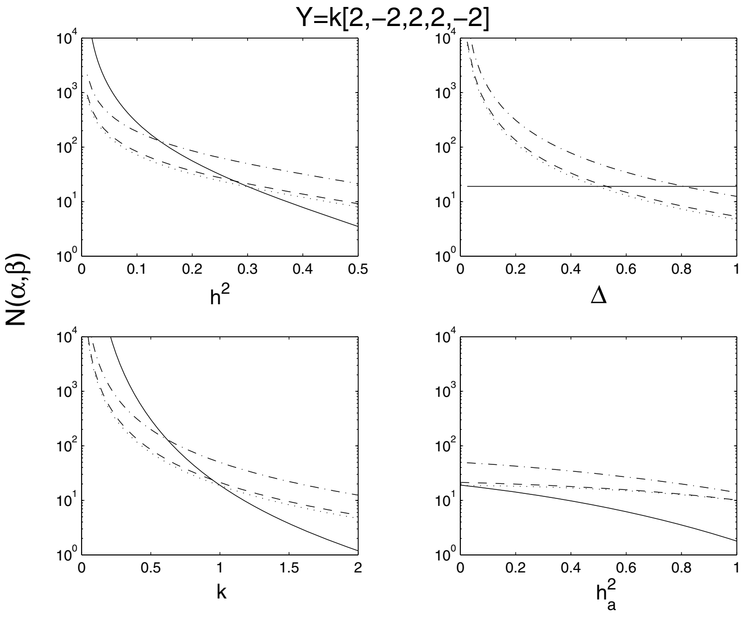

Figure 6: The number nuclear families with quantitative phenotypes and phenotype vector $Y = k(2, -2, 2, 2, -2)$ required to attain power $\beta = 0.8$ as function of various parameters when $\alpha = 0.05$ in a genomewide scan. The curves correspond to T_1^{km} (dotted), T_1^{em} (dashed) T_1^{NF} (dash-dotted) and T_2 (solid). See Figures 1 and 4 for details on multiple testing correction and parameter values.

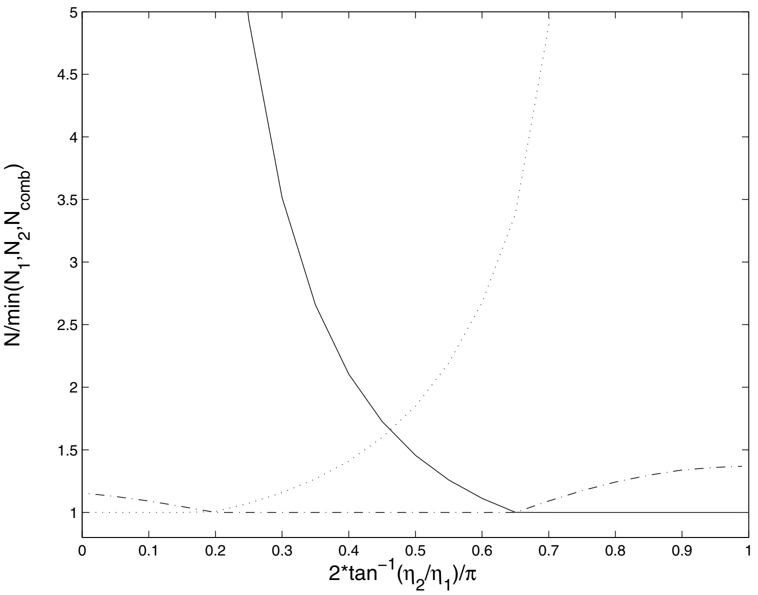

Figure 7: Normalized required sample size (39) for the association test (dotted), linkage test (solid) and combined test (dash-dotted). $N_{\mathrm{combined}}(\alpha, \beta)$ is computed by Monte Carlo, using 100 000 iterates.