doi: 10.1111/j.1469-1809.2006.00266.x

Relative Risks and Effective Number of Meioses: A Unified Approach for General Genetic Models and Phenotypes

A. Kurbasic* and O. Hössjer†

Summary

Many common diseases are known to have genetic components, but since they are non-Mendelian, i.e. a large number of genetic factors affect the phenotype, these components are difficult to localize. These traits are often called complex and analysis of siblings is a valuable tool for mapping them. It has been shown that the power of the affected relative pairs method to detect linkage of a disease susceptibility locus depends on the locus contribution to increased risk of relatives compared with population prevalence (Risch, 1990a,b). In this paper we generalize calculation of relative risk to arbitrary phenotypes and genetic models, but also show that the relative risk can be split into the relative risk at the main locus and the relative risk due to interaction between the main locus and loci at other chromosomes. We demonstrate how the main locus contribution to the relative risk is related to probabilities of allele sharing identical by descent at the main locus, as well as power to detect linkage. To this end we use the effective number of meioses, introduced by Hössjer (2005a) as a convenient tool. Relative risks and effective number of meioses are computed for several genetic models with binary or quantitative phenotypes, with or without polygenic effects.

Keywords: Complex diseases, relative risk, linkage analysis, effective number of meioses.

Introduction

An important class of traits for study in humans are those caused by multiple loci which, in general, have a chronic and highly prevalent nature. Some examples are cancer, epilepsy and psychiatric disorders. They are often called quantitative traits since the disease phenotype is typically measured on a continuous scale or it is defined by thresholds to a continuous variable. Classical linkage analysis has low power in analyzing complex diseases, since there is no one-to-one correspondence between the genotype and disease phenotype which is typical for Mendelian diseases. To be able to detect linkage to a disease susceptibility locus large samples of pedigrees are needed, since each disease gene typically has a small effect by itself. For this reason relative pair families, most notably sib pairs, have proved a valuable tool, since they are relatively easy to collect in large quantities.

Before undertaking a linkage study one must have some understanding of how many relatives are needed to obtain evidence for linkage. The fundamental assumption is that the linkage information in pairs can be studied through the relationship between their identity-by-descent sharing (IBD) at the putative locus (loci) and their phenotypes. In a series of papers Risch (1990a,b) studied relative risk ratios for binary traits. He showed that these risk ratios are the essential parameters for determining probabilities of IBD-sharing, given phenotypes.

In this paper we study relative risk for arbitrary phenotypes and genetic models. We assume that the trait/phenotype is a result of one so called main locus, but also of other sources of covariation, i.e. polygenes unlinked to the main locus and shared environment. The relative risk is split into two terms, the risk at the main locus and the risk due to other sources of variation. We

* Address for correspondence: Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden. Email: azra@maths.lth.se, Phone: +46 46 222 9272, Fax: +46 46 222 4623.
† Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden.

generalize Risch's results by showing that relative risks at the main locus determine IBD-sharing probabilities given phenotypes at the main locus.

It turns out that the relationship between IBD-sharing probabilities and power can be formalized using the effective number of meioses for testing, m^{test}, a concept introduced by Hössjer (2005a) for general pedigrees. This can be interpreted as the amount of information inherent in the pedigrees for testing linkage, given phenotypes and the genetic model. In this paper we apply m^{test} to affected relative pairs. For a single pair it depends on IBD-sharing probabilities through a very explicit formula, and for a whole dataset it is simply added over all the families.

We believe relative risks and effective number of meioses are two complementary quantities. The former has a natural epidemiological interpretation, whereas the latter is more relevant for quantifying the information inherent in the data set to detect linkage. We compute relative risks and effective number of meioses for several types of genetic models, including binary and quantitative traits, with or without polygenes. In this way we are able to determine which phenotypes and genetic models a relative pair is informative for. In particular, for quantitative phenotypes we find that extreme pairs of phenotypes are most informative for linkage, thereby supporting the previous results of Carey & Williamson (1991), Cardon & Fulker (1994) and Risch & Zhang (1995, 1996).

Conditional Variances and Covariances

Consider a trait which is influenced by a polymorphism within a gene having two possible alleles. If the trait is related to a certain disease we think of the two alleles as the normal (0) and disease (1) alleles, respectively, having probabilities q and p, $q + p = 1$. Let Y_1 and Y_2 be the observed phenotypes for two individuals, and $f(Y_k)$ and $f(Y_1, Y_2)$ the marginal and joint probability (density) functions, respectively, of Y_1 and Y_2. Each Y_k may be scalar or vector-valued, and may include covariates. The three possible genotypes (00), (01) and (11) give rise to three penetrance values $\psi_0^{(k)} = f(Y_k|(00))$, $\psi_1^{(k)} = f(Y_k|(01))$ and $\psi_2^{(k)} = f(Y_k|(11))$ for each individual k, $k = 1, 2$. Under Hardy-Weinberg equilibrium,

$$f(Y_k) = E(f(Y_k|G_k)) = q^2\psi_0^{(k)} + 2pq\,\psi_1^{(k)} + p^2\psi_2^{(k)},$$

where Y_k is kept fixed and expectation is with respect to G_k, the genotype of k. Similarly,

$$\begin{aligned}\sigma_g^{(k)} &:= \mathrm{Var}(f(Y_k|G_k)) \\ &= q^2\big(\psi_0^{(k)} - f(Y_k)\big)^2 + 2pq\big(\psi_1^{(k)} - f(Y_k)\big)^2 \\ &\quad + p^2\big(\psi_2^{(k)} - f(Y_k)\big)^2\end{aligned}$$

for $k = 1, 2$, which we refer to as the conditional genetic variance of the penetrance for k. It can be split into conditional additive and dominance genetic variance components $\sigma_g^{(k)} = \sigma_a^{(k)} + \sigma_d^{(k)}$, where

$$\begin{aligned}\sigma_a^{(k)} &= 2pq\big(p\big(\psi_2^{(k)} - \psi_1^{(k)}\big) + q\big(\psi_1^{(k)} - \psi_0^{(k)}\big)\big)^2 \\ \sigma_d^{(k)} &= (pq)^2\big(\psi_2^{(k)} - 2\psi_1^{(k)} + \psi_0^{(k)}\big)^2\end{aligned}$$

for $k = 1, 2$, c.f. Elston *et al.* (2002).

Introduce the 3×3 matrix $\psi = \{\psi_{jl}\}_{j,l=0}^2$ of joint penetrances, i.e. ψ_{jl} is the value of $f(Y_1, Y_2|G_1, G_2)$ when G_1, 1's genotype, has j disease alleles and G_2, 2's genotype, has l disease alleles. Define

$$\sigma_g^{(12)} := E(f(Y_1, Y_2|G, G)) - E(f(Y_1, Y_2|G, G')), \quad (1)$$

where Y_1 and Y_2 are fixed and expectation is with respect to two *independent* genotypes G and G'. It represents the difference in joint probability of phenotypes for two individuals having different and identical genotypes, respectively. We refer to $\sigma_g^{(12)}$ as the conditional genetic covariance of the joint penetrances. This name is motivated by considering the special case where there is no contribution to the trait from other genes, polygenes or shared environment, i.e. $\psi_{jl} = \psi_j^{(1)}\psi_l^{(2)}$. Then $\sigma_g^{(12)} = \mathrm{Cov}(f(Y_1|G_1), f(Y_2|G_2))$ for a monozygotic twin pair (1,2) and $\sigma_g^{(kk)} = \sigma_g^{(k)}$. It is shown in the Appendix that a decomposition $\sigma_g^{(12)} = \sigma_a^{(12)} + \sigma_d^{(12)}$ into additive and dominant conditional genetic covariances is possible, with

$$\begin{aligned}\sigma_a^{(12)} &= 2pq \cdot u_a\psi u_a', \\ \sigma_d^{(12)} &= (pq)^2 \cdot u_d\psi u_d',\end{aligned} \quad (2)$$

$u_a = (-q, q - p, p)$, $u_d = (1, -2, 1)$ and u_a' and u_d' the transpose of u_a and u_d, respectively.

Relative Risks

We indicate the type of relationship between 1 and 2 with R. For a pair R, the relative risk ratio

$$\lambda_R = \frac{f_R(Y_1|Y_2)}{f(Y_1)} = \frac{f_R(Y_2|Y_1)}{f(Y_2)} = \frac{f_R(Y_1, Y_2)}{f(Y_1)f(Y_2)},$$

quantifies the relative change in density for Y_1 after observing Y_2 or vice versa. Let $I \in \{0, 1, 2\}$ be the number of alleles shared identical by descent by (1,2) at the trait locus and put $f_{Ri}(Y_1, Y_2) = f_R(Y_1, Y_2|I = i)$[1]. Then decompose the relative risk as

$$\lambda_R = \lambda_R^{\text{other}} \cdot \lambda_R^{\text{main}},$$

where $\lambda_R^{\text{main}} = f_R(Y_1, Y_2)/f_{R0}(Y_1, Y_2)$ is the contribution to relative risk because of allele sharing identical by descent at the trait locus and $\lambda_R^{\text{other}} = f_{R0}(Y_1, Y_2)/(f(Y_1)f(Y_2))$ represents relative risk due to other sources of covariation, such as other genes, polygenes and shared environment. Notice that

$$f_{R0}(Y_1, Y_2) = u\psi u', \tag{3}$$

with $u = (q^2, 2pq, p^2)$, equals the second term on the RHS of (1).

Let $\alpha_{Ri} = P(I = i)$ be the prior probability that the pair R shares i alleles IBD and $r_R = 0.5\alpha_{R1} + \alpha_{R2}$ the coefficient of relationship, (Haseman & Elston, 1972). It is shown in the Appendix that

$$\lambda_R^{\text{main}} = 1 + \frac{r_R\sigma_a^{(1,2)} + \alpha_{R2}\sigma_d^{(1,2)}}{f_{R0}(Y_1, Y_2)}, \tag{4}$$

which generalizes expressions obtained by James (1971) and Risch (1990a) for binary trait recurrence and relative risks. In fact, consider a binary trait without polygenic and shared environmental effects, where phenotypes one and zero mean affected and unaffected, respectively. Let $Y_1 = Y_2 = 1$ be the phenotypes of an affected relative pair R. Drop superscript k for penetrance, so that ψ_i is the probability that an individual with i copies of the disease causing allele becomes affected. Then $\sigma_a^{(k)} = \sigma_a^{(1,2)} = \sigma_a^2$ and $\sigma_d^{(k)} = \sigma_d^{(1,2)} = \sigma_d^2$ are the additive and dominant variance of the trait. Let $K_p = P(Y_k = 1)$ be the prevalence. Since $f_{R0}(Y_1, Y_2) = K_p^2$, (4) becomes

$$\lambda_R^{\text{main}} = 1 + \frac{r_R\sigma_a^2 + \alpha_{R2}\sigma_d^2}{K_p^2},$$

in agreement with Risch (1990a).

When $f_{R0}(Y_1, Y_2)$, $\sigma_a^{(1,2)}$ and $\sigma_d^{(1,2)}$ are independent of R, λ_R^{main} will depend on the degree of relationship R only through r_R and α_{R2}, as shown by Risch (1990a). This happens when the penetrance matrix ψ is independent of R, as for one-locus models with no shared environmental or polygenic effects, but also for multiplicative multilocus models, since then the common R-dependent factor in $f_{R0}(Y_1, Y_2)$, $\sigma_a^{(1,2)}$ and $\sigma_d^{(1,2)}$ cancels out in (4). For instance, if there are no dominance effects (p small) and $R = n$ denotes a relationship of degree n, ($n = 0$: MZ twins, $n = 1$: parent-offspring, siblings, $n = 2$: half-sibs, uncle-nephew, grandparent-grandchild, $n = 3$: first cousins), then $\lambda_n - 1 = 2^{-n}(\lambda_0 - 1)$.

Linkage Analysis for Relative Pairs

IBD Probabilities and Effective Number of Meioses

Let $z_{Ri} = P(I = i|Y_1, Y_2)$ be the posterior probability that R shares i alleles IBD. Put $\lambda_{Ri}^{\text{main}} = f_{Ri}(Y_1, Y_2)/f_{R0}(Y_1, Y_2)$. Then, applying Bayes' rule, as in Risch (1990b), we get

$$z_{Ri} = \frac{\alpha_{Ri}f_{Ri}(Y_1, Y_2)}{f_R(Y_1, Y_2)} = \frac{\alpha_{Ri}\lambda_{Ri}^{\text{main}}}{\lambda_R^{\text{main}}}, \quad i = 0, 1, 2. \tag{5}$$

The amount of information contained in (Y_1, Y_2) for testing linkage between the trait locus and genetic markers is to a large extent determined by how much $\{z_{Ri}\}_{i=0}^2$ depart from $\{\alpha_{Ri}\}_{i=0}^2$. The effective number of meioses m^{test} for testing is introduced in Hössjer (2005a) for general pedigrees, phenotypes and genetic models. It quantifies the equivalent number of fully observed meiotic events contained in (Y_1, Y_2). It is shown in the Appendix that

$$\text{m}_R^{\text{test}} = \log_2\left(\frac{z_{R0}^2}{\alpha_{R0}} + \frac{z_{R1}^2}{\alpha_{R1}} + \frac{z_{R2}^2}{\alpha_{R2}}\right), \tag{6}$$

[1]When R is a monozygotic twin pair or a parent-offspring pair the event '$I = 0$' has probability zero and $f_{R0}(Y_1, Y_2|I = 0)$ is not defined, but we can still use formula (3). For the monozygotic twin pair, when '$I = 1$' the probability is zero and $f_{R1}(Y_1, Y_2)$ is defined as for an ordinary sib pair. For a unilineal relationship '$I = 2$' has probability zero, and then $f_{R2}(Y_1, Y_2)$ need not be defined.

with the convention $0/0 = 0$ whenever $\alpha_{Ri} = 0$. For a data set with several relative pairs (of the same or a different kind), $\mathrm{m}^{\mathrm{test}}$ is obtained by simply adding (6) over all pairs. For a unilineal relationship of degree $n(\alpha_{R1} = 2^{-(n-1)}, \alpha_{R2} = 0)$, the maximal possible value $m_R^{\mathrm{test}} = n - 1$ is obtained when the genetic model and (Y_1, Y_2) is such that $z_{R1} = 1$. In other words, when the genetic component at the main locus is strong, distant relationships are more informative than close ones.

Power Approximation

We will now motivate the relevance of m^{test} to linkage analysis. Suppose we wish to test

$$H_0 : \tau \notin \Omega$$

$$H_1 : \tau \in \Omega,$$

where Ω is the genomic region of interest, consisting of n_Ω chromosomes of total length L_Ω Morgans, and τ is the unknown position of the disease gene. Consider a set of N unrelated relative pairs, $i = 1, \ldots, N$, whose relationships R_i and phenotypes $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$ may vary. Let m_i be the number of meiosis in the pedigree corresponding to R_i and $v_i(t)$ the inheritance vector of Pedigree i at locus t. This is a binary vector of length m_i, each bit of which corresponds to a meioses, with value 0 or 1 depending on whether a grandpaternal or grandmaternal allele was transmitted (Donnelly, 1983). Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ be the collection of all phenotypes and $\boldsymbol{v}(t) = (v_1(t), \ldots, v_N(t))$ the collection of inheritance vectors at locus t. The latter is a binary vector of length $m_{\mathrm{total}} = m_1 + \cdots + m_N$. With complete marker information, a wide class of test statistics for testing H_0 against the pointwise alternative $\tau = t, t \in \Omega$, is

$$Z(t) = S(\boldsymbol{v}(t)), \tag{7}$$

where $S(\boldsymbol{w}) = S(\boldsymbol{w}; \boldsymbol{Y})$ is a score function defined for all binary vectors $\boldsymbol{w} = (w_1, \ldots, w_N)$ of length m_{total}. We assume that S is standardized so that $\sum_{\boldsymbol{w}} S(\boldsymbol{w}) = 0$ and $2^{-m\mathrm{total}} \sum_{\boldsymbol{w}} S^2(\boldsymbol{w}) = 1$, ensuring $E_{H_0}(Z(t)) = 0$ and $\mathrm{Var}_{H_0} Z(t)) = 1$ under complete marker information. Notice that (7) contains as special case

$$Z(t) = \sum_{i=1}^{N} \gamma_i S_i(v_i(t)), \tag{8}$$

which is a linear combinations of family scores $S_i(v_i(t))$ with weights γ_i satisfying $\sum_{i=1}^N \gamma_i^2 = 1$, and with S_i a standardized score function of Pedigree i. The class (8) includes the affected pedigree method, see e.g. Weeks & Lange (1988), Fimmers *et al.* (1989), Whittemore & Halpern (1994) and Kruglyak *et al.* (1996). As test statistics for testing H_0 against H_1 we use

$$Z_{\max} = \sup_{t \in \Omega} Z(t),$$

and H_0 is rejected as soon as $Z_{\max}$ exceeds a given threshold z. Since $Z(t)$ can be seen as a stationary process we can use results from extreme value theory to approximate power and significance level; see Appendix for details. The power approximation (A.5) is an increasing function of the noncentrality parameter

$$\eta = E(Z(\tau)) = \sum_{\boldsymbol{w}} S(\boldsymbol{w}) P(\boldsymbol{w}), \tag{9}$$

where $P(\boldsymbol{w}) = P(\boldsymbol{v}(\tau) = \boldsymbol{w} | \boldsymbol{Y}) = \prod_{i=1}^N P(v_i(\tau) = w_i | \boldsymbol{Y}_i)$ is the joint conditional distribution of all inheritance vectors at the trait locus. The maximal noncentrality parameter for the class (7) of test statistics

$$\eta_{\max} = \sqrt{2^{m_{\mathrm{total}}^{\mathrm{test}}} - 1} \tag{10}$$

is attained for a score function $S_{\mathrm{opt}}(\boldsymbol{w}) = (P(\boldsymbol{w}) - \mu)/\sigma$, where μ and σ are standardization constants, see Hössjer (2005a) for details. Notice that S_{opt} requires knowledge of the genetic model, i.e. penetrance parameters and disease allele frequency. Since the genetic model is rarely known for complex diseases, $\eta_{\max}$ is an upper bound of the noncentrality parameter, and $m_{\mathrm{total}}^{\mathrm{test}} = m_{R_1}^{\mathrm{test}} + \cdots + m_{R_N}^{\mathrm{test}}$ is related to this upper bound through a strictly monotone transformation. But since the power to detect linkage is approximately a monotone function of η (cf. (A.4)), the maximal power, obtained by putting $\eta = \eta_{\max}$ in (A.4), depends (approximately) monotonically on $m_{\mathrm{total}}^{\mathrm{test}}$ as well.

From the above discussion we can interpret $m_{\mathrm{total}}^{\mathrm{test}}$ as quantifying the amount of information in the data set for detecting linkage. It is ideal, since it requires knowledge of the genetic model. In particular, for affected relative pairs $m_{\mathrm{total}}^{\mathrm{test}}$ depends on the genetic model through relative risks at the main locus, see (5) and (6). In practice, when the genetic model is unknown we may choose a suboptimal score function S and achieve lower power

than predicted by $m^{\text{test}}_{\text{total}}$ for the true genetic model. A possibility in this case is to compute the noncentrality parameter η in (9) and the associated power in (A.5) for a number of possible P, each one corresponding to a hypothesized genetic model. Then, the sample size should be chosen to be so large that a summary statistic of all power values (e.g. the mean) exceeds a predetermined lower bound. Such an approach would then be robust against model misspecification.

Figure 1 shows the power β of detecting the significant linkage on a chromosome of length 2.985 Morgans. In calculation we used the crossover rate $\rho_Z = 2$ and normalized slope $d = 1$ (both defined in the Appendix), although they can vary a little depending on the genetic model and the pedigree structure (Lander & Kruglyak, 1995; Hössjer, 2005c). We calculated thresholds required to reach three 'standard' significance levels of 0.05, 0.01 and 0.001. In all three cases we needed a value of $m^{\text{test}}_{\text{total}}$ between four and five to achieve a power of 0.9. As we will show below, the number of relative pairs required to reach a high power depends on the genetic model and the phenotypes. In Figure 2, when the threshold z is the one required for genomewide significance the values of $m^{\text{test}}_{\text{total}}$ needed for reaching power 0.9 increase, but not remarkably.

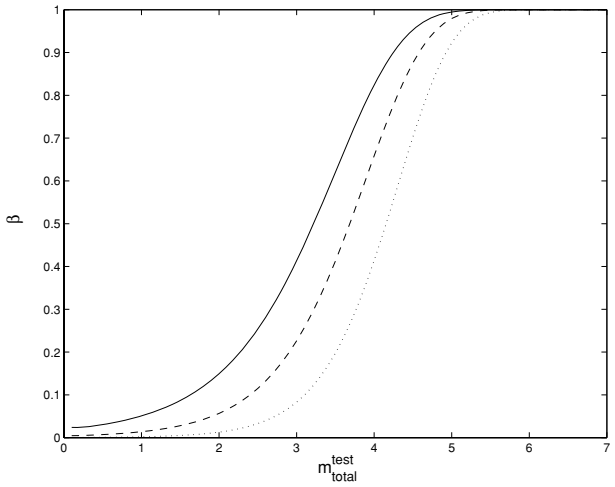

Figure 1 Approximation of the power to detect linkage β as a function of the effective number of meioses $m^{\text{test}}_{\text{total}}$ for one chromosome of length $L_\Omega = 2.985$ Morgans. The threshold z and significance level α are chosen as $z = 3.375$ and $\alpha = 0.05$ (solid), $z = 3.863$ and $\alpha = 0.01(--)$, $z = 4.455$ and $\alpha = 0.001(\cdots)$.

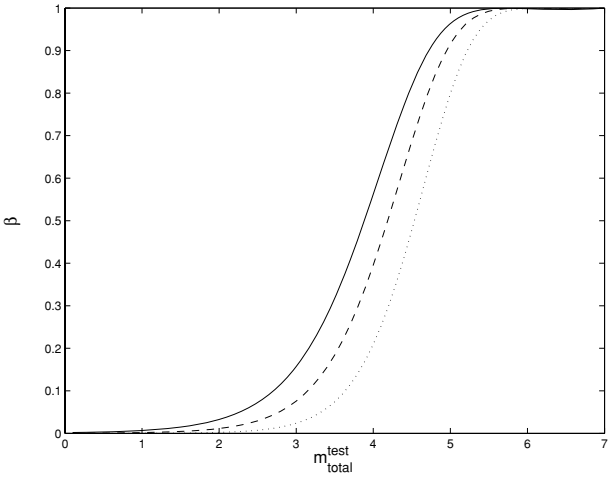

Figure 2 Approximation of the power to detect linkage β as a function of the effective number of meioses $m^{\text{test}}_{\text{total}}$. The threshold z is calculated for a genomewide significance level α, i.e. $n_\Omega = 22$ and $L_\Omega = 33.5$ Morgans. In the Figure $z = 4.1$ and $\alpha = 0.05$ (solid), $z = 4.5$ and $\alpha = 0.01(--)$, $z = 5.02$ and $\alpha = 0.001(\cdots)$.

Genetic Models

Gaussian Phenotypes

We studied relative risks for a class of genetic models where the genetic influence is a mixture of a major gene and a number of polygenes but the major gene and polygenes are unlinked. The goal of the analysis is to map the major gene, and to increase statistical efficiency we take the polygenes into account. If G_k is the major gene's genotype for individual k and Y_k the phenotype, we assume $Y_k|G_k \in N(\mu_{|G_k|}, \sigma^2)$. Here $|G_k|$ is the number of disease alleles of G_k and μ_0, μ_1 and μ_2 are the mean phenotype values for an individual with 0, 1, and 2 disease alleles, respectively. Residual variance σ^2 is the variance caused by polygenic and/or environmental effects. If large values of the phenotype indicate disease a natural constraint is $\mu_0 \le \mu_1 \le \mu_2$. For a relative pair we have $(Y_1, Y_2)|(G_1, G_2) \in N(\mu, \sigma^2\Sigma)$, where $\mu = (\mu_{|G_1|}, \mu_{|G_2|})$, Σ is a 2×2 correlation matrix with ones on the diagonal and correlation coefficient $\rho_Y = r_R h_a^2 + \alpha_{R2} h_d^2$ with h_a^2 and h_d^2 additive and dominant polygenic heritability, respectively (i.e. the fraction of σ^2 due to additive and dominance effects). Then we have that

$$f(Y_k|G_k) = \phi((Y_k - \mu_{|G_k|})/\sigma)/\sigma$$

Model	Disp	Dom	h^2	h_a^2	p	μ_0	μ_1	μ_2	σ
1	2	0	0.1525	0	0.1	−0.1841	0.7365	1.6570	0.9206
2	2	0	0.1525	0.1	0.1	0.1841	0.7365	1.6570	0.9206
3	2	0	0.1525	0.5	0.1	0.1841	0.7365	1.6570	0.9206
4	2	0	0.1525	0.8	0.1	0.1841	0.7365	1.6570	0.9206
5	2	0	0.0020	0.5	0.001	−0.0020	0.9970	1.9960	0.9990
6	2	0	0.1525	0.5	0.1	−0.1841	0.7365	1.6570	0.9206
7	2	0	0.3334	0.5	0.5	−0.81656	0	0.8165	0.8165
8	2	1	0.3810	0.5	0.1	−0.2990	1.2745	1.2745	0.7867
9	2	−1	0.0381	0.5	0.1	−0.0196	−0.0196	1.9419	0.9808
10	2	0.5	0.2652	0.5	0.1	−0.2486	1.0372	1.4658	0.8572
11	2	−0.5	0.0679	0.5	0.1	−0.1062	0.3765	1.8247	0.9654
12	1	0	0.0432	0.5	0.1	−0.0978	0.3913	0.8804	0.9782
13	3	0	0.2883	0.5	0.1	−0.2531	1.0124	2.2779	0.8436
14	4	0	0.4186	0.5	0.1	−0.3050	1.2200	2.7450	0.7625

Table 1 Overview of the parameters in the study with Gaussian phenotypes. Note that $E(Y_k) = 0$, $\mathrm{Var}(Y_k) = 1$, and h^2 is the main locus heritability, i.e. $\mathrm{Var}(\mu_{|G_k|})/\mathrm{Var}(Y_k)$.

and

$$f(Y_1, Y_2|\, G_1, G_2) = \phi_2((Y_1 - \mu_{|G_1|})/\sigma, ((Y_2 - \mu_{|G_2|})/\sigma; \rho_Y)/\sigma^2,$$

where ϕ and $\phi_2(.,.,\rho_Y)$ are the univariate and bivariate standard normal densities, the latter with correlation coefficient ρ_Y. See Lynch & Walsh (1998) and Hössjer (2005b) for more details about Gaussian phenotypes.

Liability Threshold Model

In the liability threshold model the observed phenotypes are binary ($Y_k = 1$ affected and $Y_k = 0$ unaffected) but there is no simple Mendelian inheritance pattern, and the probability of expressing the disorder i.e. the distribution of Y_k is modelled as a function of an underlying quantitative variable X_k, the same kind of variable as in Section 5.1. The phenotype is then defined as $Y_k = 1_{\{X_k \geq T\}}$ where T is a given threshold. For more details about liability threshold models see Todorov & Suarez (2002).

We assume $X_k|\, G_k \in N(\mu_{|G_k|}, 1)$ and hence $X|\, G \in N(\mu, \Sigma)$. We then have penetrance parameters

$$\psi_i = P(Y_k = 1||\, G_k| = i) = 1 - \Phi(T - \mu_i),$$

for $i = 0, 1, 2$, where Φ is the distribution function of a standard normal variable. This yields

$$f(Y_k|\, G_k) = \psi_{|G_k|}^{Y_i} \cdot (1 - \psi_{|G_k|})^{1-Y_i}$$

and

$$f(Y_1, Y_2|\, G_1, G_2) = \int_A \phi_2(x_1, x_2; \rho_X) d\, x$$

where $A = (T - \mu_{|G_1|}, \infty) \times (T - \mu_{|G_2|}, \infty)$ if $Y_1 = Y_2 = 1$, $A = (T - \mu_{|G_1|}, \infty) \times (-\infty, T - \mu_{|G_2|})$ if $Y_1 = 1$ and $Y_2 = 0$, $A = (-\infty, T - \mu_{|G_1|}) \times (T - \mu_{|G_2|}, \infty)$ if $Y_1 = 0$ and $Y_2 = 1$, and $A = (-\infty, T - \mu_{|G_1|}) \times (-\infty, T - \mu_{|G_2|})$ if $Y_1 = Y_2 = 0$. When there are no polygenic effects, i.e. when $h_a^2 = h_d^2 = 0$, Σ is diagonal. Then we obtain an ordinary one-locus model for binary traits, with $f(Y_1, Y_2|\, G_1, G_2) = f(Y_1|\, G_1) f(Y_2|\, G_2)$.

Results

For the Gaussian mixed model we have standardized phenotypes, so that $E(Y_k) = q^2\mu_0 + 2pq\mu_1 + p^2\mu_2 = 0$ and $\mathrm{Var}(Y_k) = 1$. In practice both $E(Y_k)$ and $V(Y_k)$ have to be estimated, for instance as the sample mean and sample variance of a randomly drawn subset of the population. Alternatively one could use the given data set, but then a more sophisticated estimation procedure that includes modelling of the ascertainment scheme is necessary. After standardization there are four essential genetic parameters when $h_d^2 = 0$, namely p, h_a^2, the displacement $\mathrm{Disp} = (\mu_2 - \mu_0)/\sigma$ that quantifies the strength, and $\mathrm{Dom} = (2\mu_1 - \mu_0 - \mu_2)/(\mu_1 - \mu_0)$ that quantifies the degree of dominance of the main locus genetic component. We studied the relation between the relative risk and different parameters, but also investigated the informativity of the different relative pairs by looking at

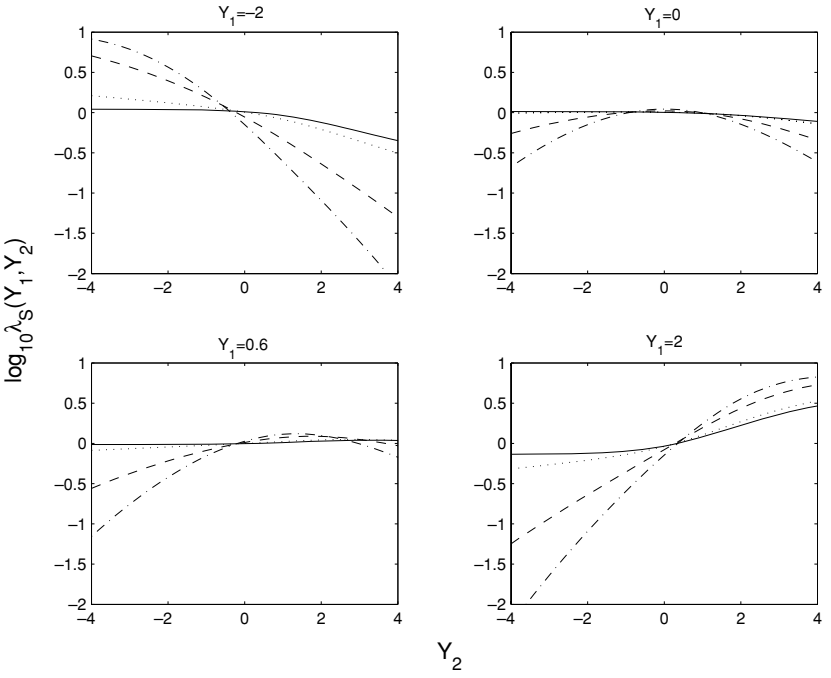

Figure 3 Relative risk λ_S as a function of the values of the trait, Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of additive polygenic heritability; $h_a^2 = 0$ (solid), $h_a^2 = 0.1(\cdots)$, $h_a^2 = 0.5(--)$, and $h_a^2 = 0.8$, $(- \cdot -)$.

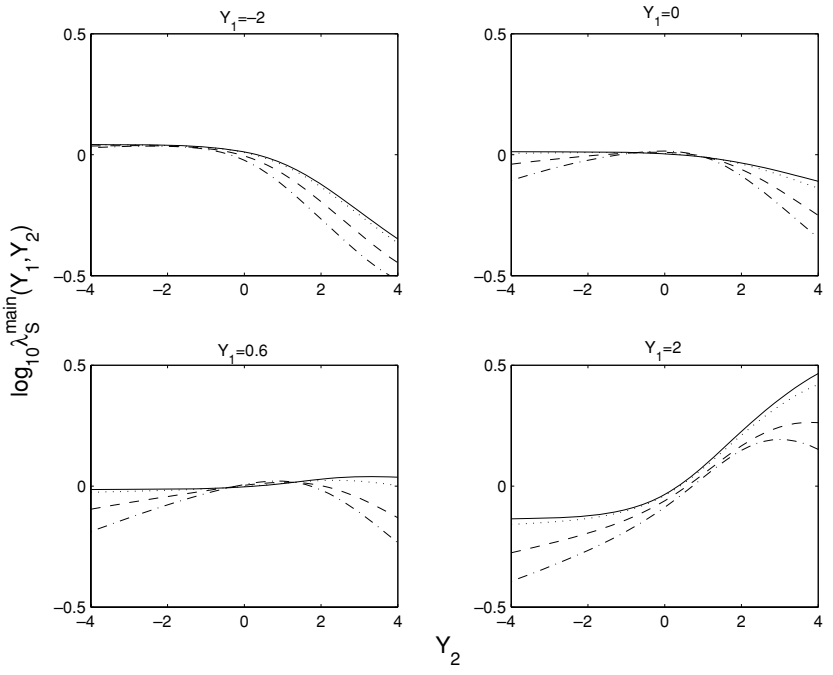

Figure 4 Relative risk due to trait locus λ_S^{main} as a function of the values of the trait, Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of additive polygenic heritability; $h_a^2 = 0$ (solid), $h_a^2 = 0.1(\cdots)$, $h_a^2 = 0.5(--)$, and $h_a^2 = 0.8$, $(- \cdot -)$.

the value of m_R^{test}. A list of genetic models is provided in Table 1.

We have that $\lambda_R = \lambda_R^{\text{main}}$ when $h_a^2 = 0$, see for example Figures 3 and 4, which is a result of no interaction between the main locus and the other loci. We define the pairs as being concordant if they have the same or similar phenotypes and discordant when they have opposite phenotypes. The informativity increases with h_a^2 for discordant pairs and decreases with h_a^2 for concordant pairs, see Figure 5. The most informative are discordant sib pairs with extreme phenotype values, large h_a^2, and large h^2 (main locus heritability). Relatives with phenotypes close or equal to zero are non-informative. These are the same kind of conclusions as in Risch & Zhang

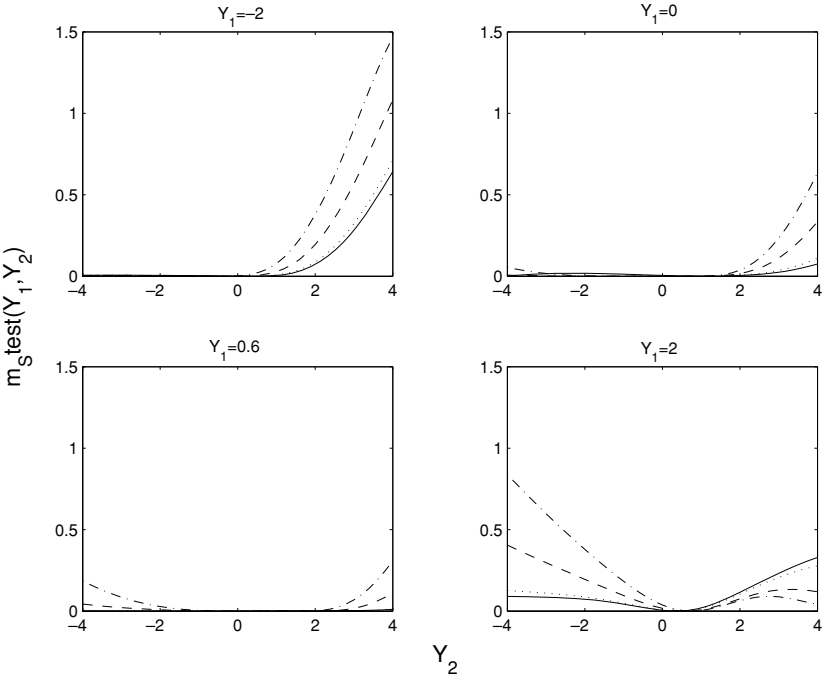

Figure 5 Effective number of meioses m_S^{test} as a function of the values of the trait, Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of additive polygenic heritability; $h_a^2 = 0$ (solid), $h_a^2 = 0.1(\cdots)$, $h_a^2 = 0.5(--)$, and $h_a^2 = 0.8$, $(- \cdot -)$.

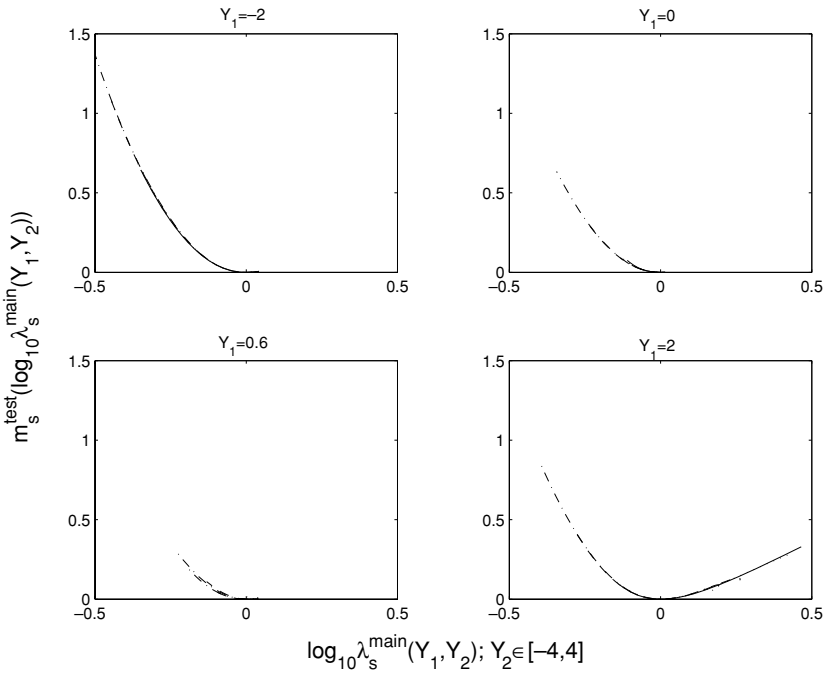

Figure 6 Effective number of meioses m_S^{test} as a function of the values of the relative risk due to the trait locus λ_S^{main}. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of additive polygenic heritability; $h_a^2 = 0$ (solid), $h_a^2 = 0.1(\cdots)$, $h_a^2 = 0.5(--)$, and $h_a^2 = 0.8$, $(- \cdot -)$.

(1995, 1996). Concordant pairs for high positive and low negative values are also informative. If $p < 0.5$ the concordant positive phenotypes are more informative than the concordant negative phenotypes, see Figures 5-10. For discordant phenotypes the close relatives are most informative, but for concordant positive phenotypes the distant relatives are more informative, since α_{R_1} is small, see Figure 7. Further, the parent-offspring pair is not informative since $\alpha_{R_1} = z_{R_1} = 1$. Informativity increases also with $p(0 < p \leq 0.5)$. When p is small, often the dominant model (Dom $= 1$) is most informative and the recessive one (Dom $= -1$) least informative. Informativity increases with Disp, even when h_a^2 increases. Figure 6 displays the relationship between

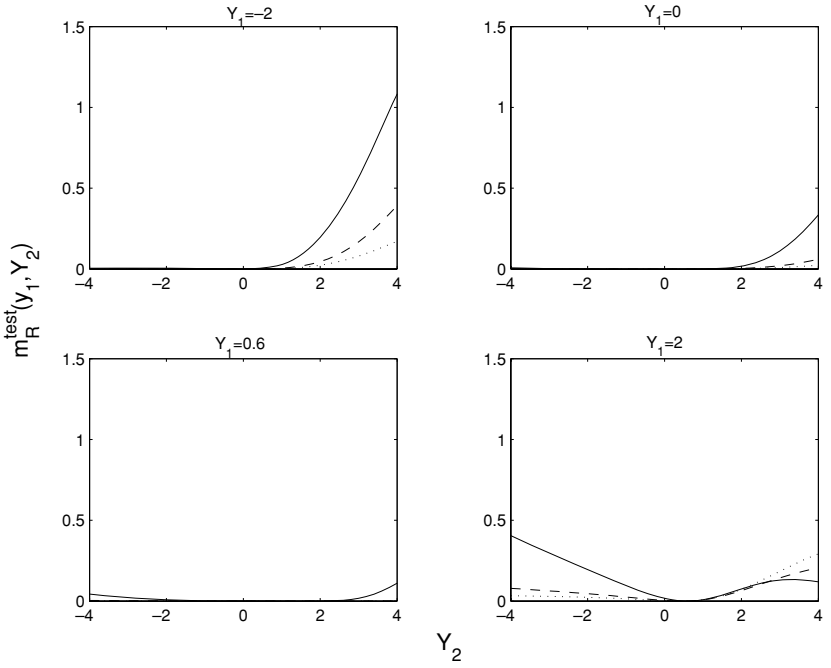

Figure 7 Effective number of meioses m_R^{test} as a function of the values of the trait, Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, additive polygenic heritability $h_a^2 = 0.5$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for three different relative pairs: sib pair (solid), grandparent-offspring (- -), and first cousins ($\cdots$). Note: For a parent-offspring pair $m^{\text{test}} = 0$ since $\alpha_{R_1} = z_{R_1}$.

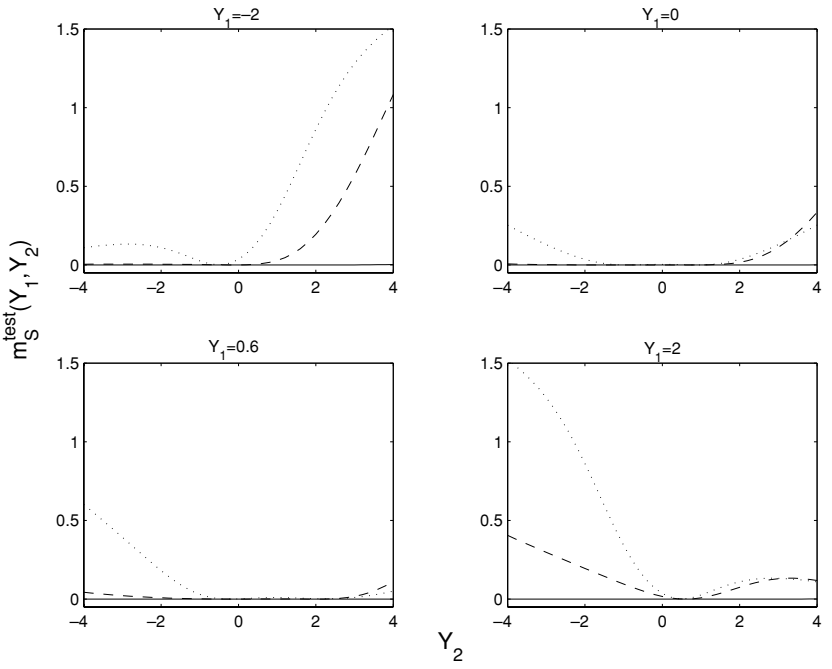

Figure 8 Effective number of meioses m_S^{test} as a function of the values of the trait, Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Additive polygenic heritability $h_a^2 = 0.5$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for three different values of disease allele frequencies, $p = 0.001$ (solid), $p = 0.1$(− −), and $p = 0.5(\cdots)$. When $p = 0.5$ the curve for $Y_1 = 2$ is just a mirror image of the curve when $Y_1 = -2$ because of symmetry.

the relative risk due to trait locus λ_S^{main} and the effective number of meioses m_S^{test}. It can be seen as a composition of Figures 4 and 5, where both λ_S^{main} and m_S^{test} vary as functions of the same phenotypes and genetic models (h_a^2). Interestingly, m_S^{test} is almost the same function of λ_S^{main} for different genetic models and phenotype combinations, although the range of admissible values differs. Large values of $|\log_{10}\lambda_S^{\text{main}}|$ indicate high informativity (large m_S^{test}), although negative values of λ_S^{main} are slightly more informative than positive ones.

For the liability threshold model we studied risk for different pairs of relatives and different values of h_a^2, p, and (ψ_0, ψ_1, ψ_2). Figures 11-15 show some examples from which we can see that both the relative risk and

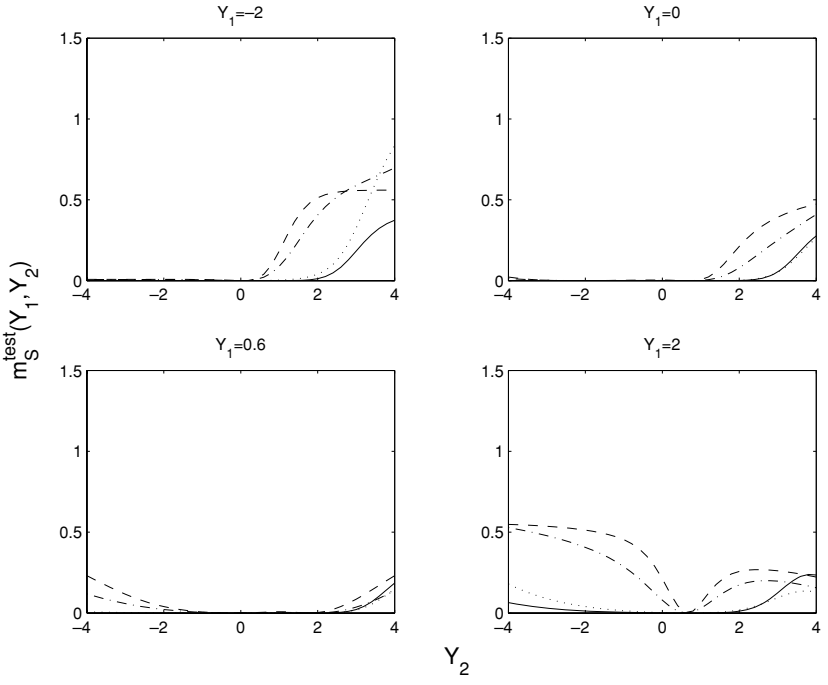

Figure 9 Effective number of meioses m_S^{test} as a function of the values of the trait, Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, additive polygenic heritability $h_a^2 = 0.5$, and displacement Disp $= 2$. Relative risk is calculated for four different values of dominance of the main locus genetic component; Dom $= -1$ (solid), Dom $= 1(--)$, Dom $= 0.5(--)$, and Dom $= -0.5(\cdots)$.

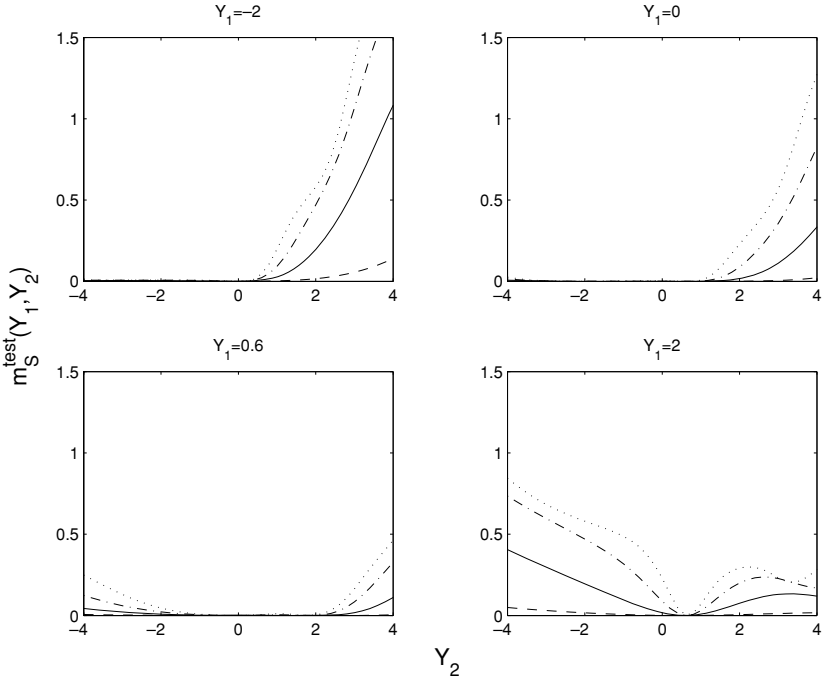

Figure 10 Effective number of meioses m_S^{test} as a function of the values of the trait, Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, additive polygenic heritability $h_a^2 = 0.5$, and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of displacement; Disp $= 1(--)$, Disp $= 2$ (solid), Disp $= 3(- \cdot -)$, and Disp $= 4(\cdots)$.

the effective number of meioses are almost independent of h_a^2. Only when the disease allele frequency p is low in relation to ψ_0 can we observe some dependence. The relative risk increases (slightly) with h_a^2 when $Y_1 = Y_2$ and decreases with h_a^2 when $Y_1 \neq Y_2$. In the case where $Y_1 = Y_2 = 0$ the relative risk is zero or very close to zero. On the other hand, the relative risk at the main locus decreases with h_a^2 in the case where $Y_1 \neq Y_2$ and $Y_1 = Y_2 = 1$. Again, when $Y_1 = Y_2 = 0$ it is very close to zero. From the Figures with the effective number of meioses we can observe that m_R^{test} slightly increases when $Y_1 \neq Y_2$ (discordant pair) and decreases when

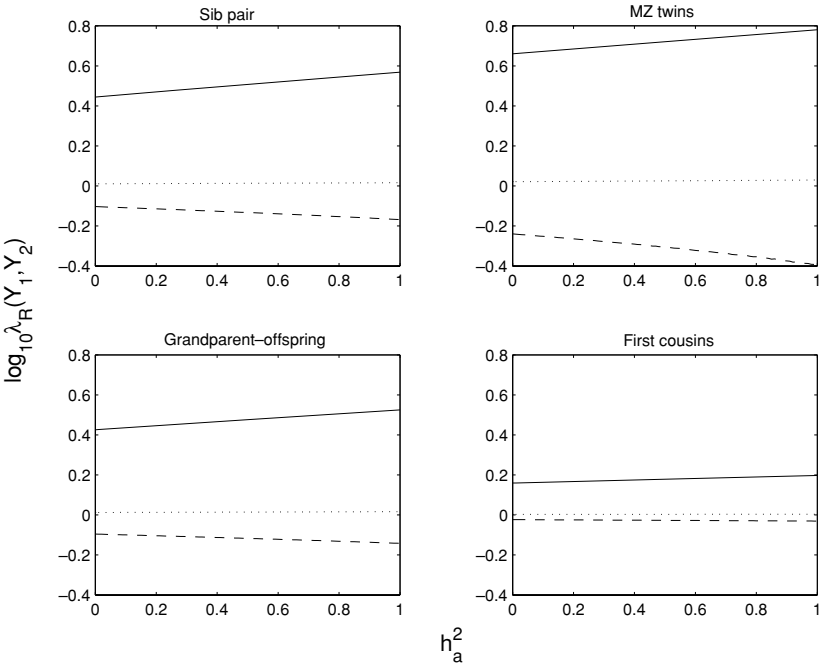

Figure 11 Relative risk λ_R for different relative pairs as a function of the additive polygenic heritability h_a^2 and values of the trait, Y_1 and Y_2. It is calculated for three different combinations of Y_1 and Y_2; $Y_1 = Y_2 = 1$ (solid), $Y_1 = 0$ and $Y_2 = 1(--)$, and $Y_1 = Y_2 = 0(\cdots)$. Disease allele frequency p $= 0.1$. The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

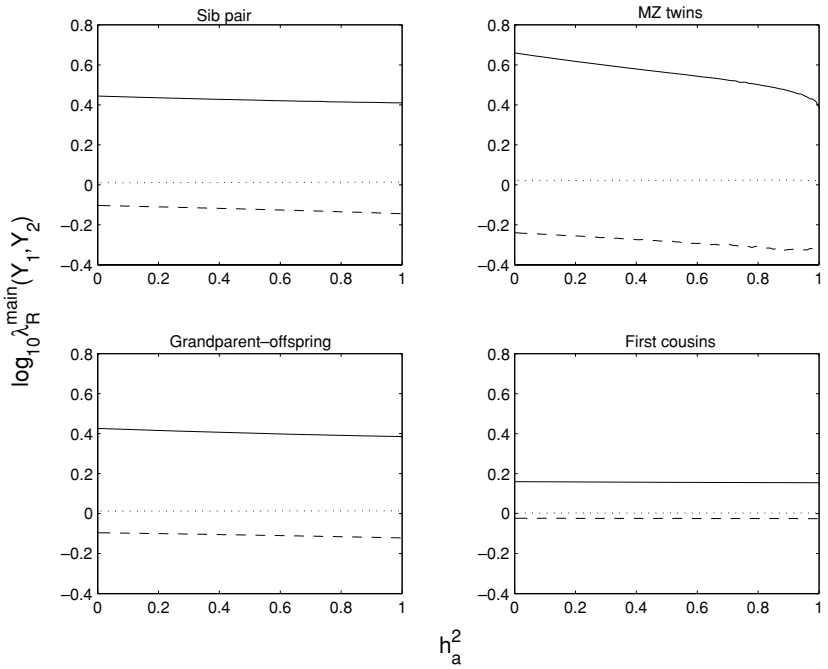

Figure 12 Relative risk at the main locus λ_R^{main} for different relative pairs as a function of the additive polygenic heritability h_a^2 and values of the trait, Y_1 and Y_2. It is calculated for three different combinations of Y_1 and Y_2; $Y_1 = Y_2 = 1$ (solid), $Y_1 = 0$ and $Y_2 = 1(--)$, and $Y_1 = Y_2 = 0(\cdots)$. Disease allele frequency p $= 0.1$. The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

$Y_1 = Y_2 = 1$ (concordant pair). The MZ twin pair is noninformative, since $\alpha_{R2} = z_{R_2} = 1$, as well as a relative pair with $Y_1 = Y_2 = 0$. Distant relationships are more informative than close ones when $Y_1 = Y_2 = 1$, and vice versa when $Y_1 \neq Y_2$. We can also observe that the values of ψ_1 and ψ_2 affect m_R^{test}, but not as much as the value of ψ_0.

Discussion

In this paper we have generalized relative risk ratios to arbitrary genetic models and shown how to split it into two terms, one due to effects at the main locus and one due to polygenic and shared environmental effects. We further extended the results of Risch (1990a,b) and showed that IBD-sharing probabilities given phenotypes

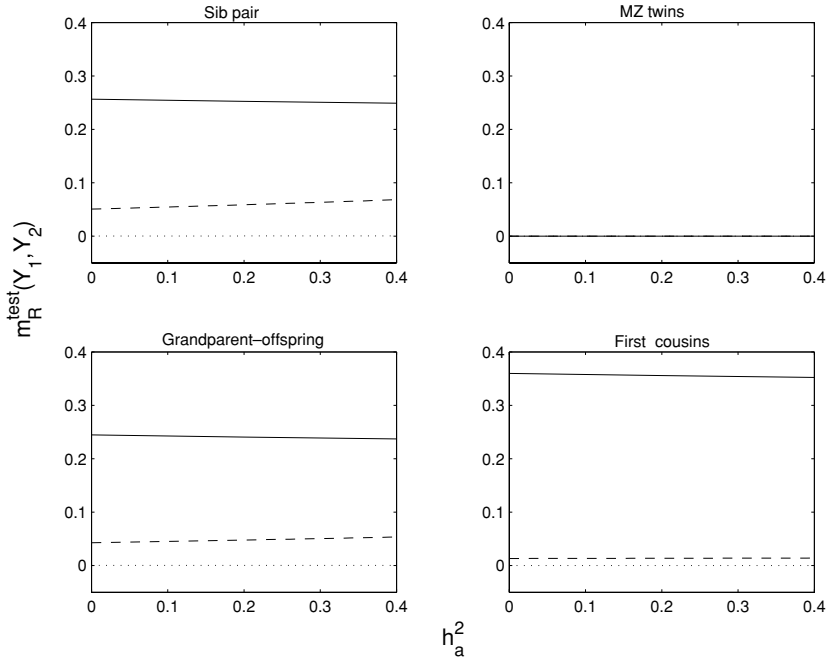

Figure 13 Effective number of meioses m_R^{test} for different relative pairs as a function of the values of the additive polygenic heritability h_a^2 and values of the trait, Y_1 and Y_2. It is calculated for three different combinations of Y_1 and Y_2; $Y_1 = Y_2 = 1$ (solid), $Y_1 = 0$ and $Y_2 = 1(--)$, and $Y_1 = Y_2 = 0(\cdots)$. Disease allele frequency p $= 0.1$. The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

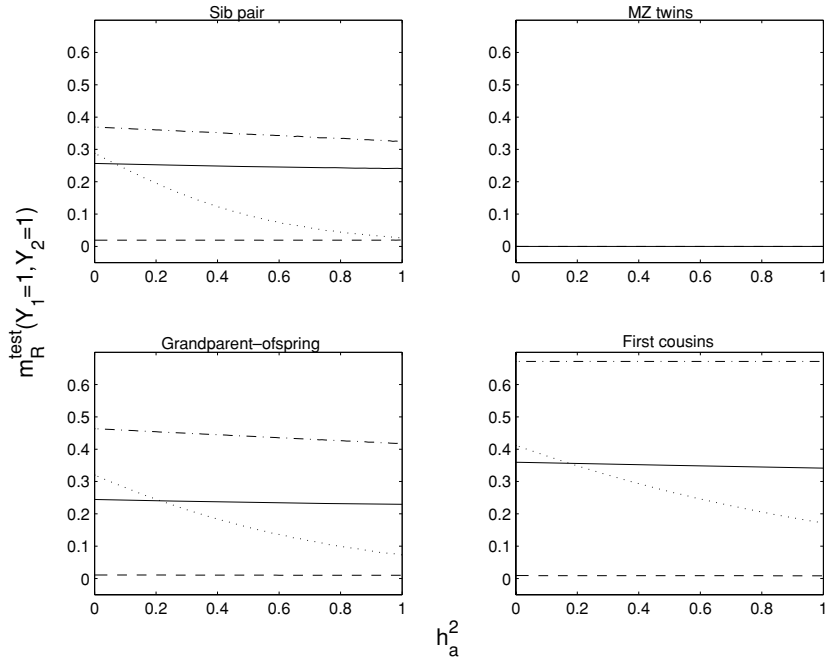

Figure 14 Effective number of meioses m_R^{test} for different relative pairs as a function of the values of the additive polygenic heritability h_a^2 when $Y_1 = Y_2 = 1$. It is calculated for four different values of disease allele frequency, $p = 0.001(\cdots)$, $p = 0.05(- \cdot -)$, $p = 0.1$ (solid) and $p = 0.5(--)$. The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

depend on relative risks at the main locus. Finally, we showed that the power to detect linkage of an optimal test statistic is closely related to these probabilities, by summing the effective number of meioses for testing over all relative pairs. We have shown numerically how relative risks and effective number of meioses (and hence also the number of relative pairs needed for detecting linkage) depend on phenotypes for two classes of genetic models: Gaussian and liability threshold models. For Gaussian phenotypes extreme discordant sib pairs are most powerful. This is because these pairs are unlikely to share alleles IBD for any genetic model. Sib pairs concordant for extreme values can also be useful, whereas sib pairs with intermediate values are only informative when the genetic component at the disease locus is strong. The two major determinants of the

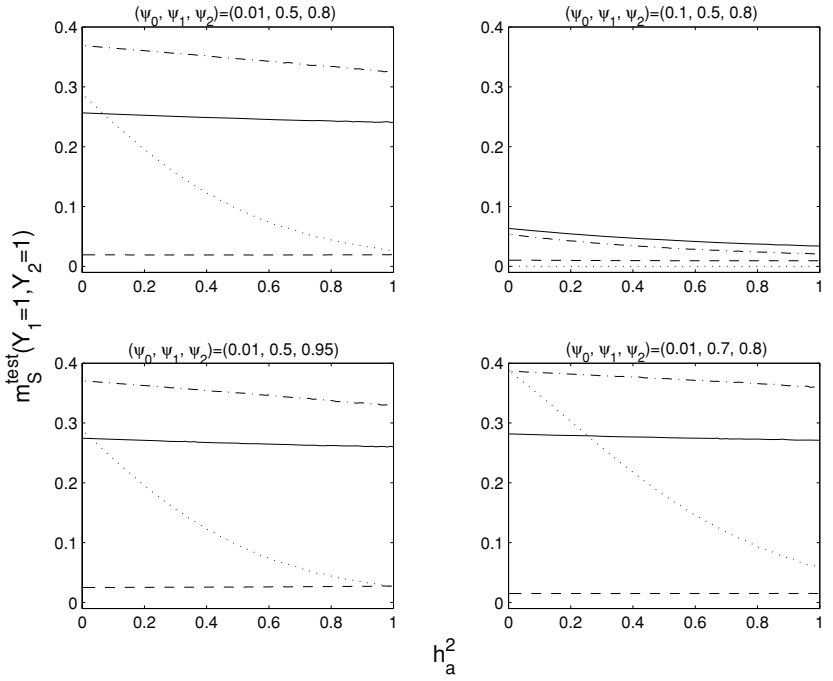

Figure 15 Effective number of meioses m_S^{test} for a sib pair as a function of the values of the additive polygenic heritability h_a^2 when $Y_1 = Y_2 = 1$. It is calculated for four different values of disease allele frequency, $p = 0.001(\cdots)$, $p = 0.05(- \cdot -)$, $p = 0.1$ (solid) and $p = 0.5(--)$ and four different penetrance vectors (ψ_0, ψ_1, ψ_2).

power to detect linkage for a locus contributing to a quantitative trait are the heritability at that locus and the additive polygenic heritability. Additive polygenic heritability also increases statistical efficiency and compensates for power loss when the heritability at the main locus is low.

The liability threshold model seems intuitively reasonable for complex diseases, including one major gene and polygenes. However, we have shown that polygenic heritability affects the effective number of meioses and power very little when the penetrance parameters ψ_i are kept fixed. This indicates that the penetrance parameters themselves (without any polygenic components) provide a good description of a wide range of genetic models for binary traits. Another possibility is to consider relative risks and effective number of meioses for oligogenic multilocus models, as in Risch (1990a).

Our approach is conditional on observed phenotypes. Many authors have realized that treating the number of alleles shared IBD at the marker locus (loci) as dependent, and the sib trait values as an independent variable has several advantages, since sample selection is often through trait values but almost never through marker genotypes (see (Risch & Zhang 1995, 1996; Dudoit & Speed 2000; Kraft & Thomas, 2000; Sham *et al.* 2002). However, conditioning on phenotypes is relevant to most linkage analyses. Although the classical lod score of Morton (1955) is formulated as the tenth logarithm of a likelihood ratio of the joint probability of marker data and phenotypes, this likelihood ratio is essentially equivalent to the conditional probability of marker data given phenotypes. Likewise, the mod score of Risch (1984) and Clerget-Darpoux *et al.* (1986) is equivalent to maximizing (the tenth logarithm of) the conditional probability of marker data given genotypes with respect to genetic model parameters. Ewens & Shute (1986) demonstrated the equivalence of an LR-statistic to a mod score adjusted for ascertainment. On the other hand, Vieland & Hodge (1995, 1996) found that, in general, conditioning on phenotypes only gives an approximate correction for ascertainment. The method of Haseman & Elston (1972) is based on regressing squared phenotype difference on IBD-sharing. However, by standardizing with a nonparametric variance estimator this statistic can be regarded as quantifying IBD-sharing given phenotypes, and hence is robust against ascertainment. Variance components techniques (Amos, 1994; Almasy & Blangero, 1998) are less robust against ascertainment, in that parameters are maximized separately in the numerator and denominator of the likelihood ratio. Likelihood score statistics, on the other hand, can be formulated in terms of marker data given phenotypes,

see Whittemore (1996), Tang & Siegmund (2001) and Hössjer (2005b) for details.

We have assumed complete marker data throughout this paper. Since the cost of genotyping very dense marker maps constantly decreases, it suffices that all relative pairs R_i are genotyped to make this assumption a reasonable theoretical simplification. The fact that we only considered relative pairs in this paper does not mean that we necessarily advocate such a design. In fact, nuclear families with more than two children are often much more informative for linkage than sib pairs (Tang & Siegmund, 2001; Hössjer, 2005a). Nevertheless, sib pair studies are commonly used, and we hope that our results will be useful when designing such studies. Another important aspect of this paper is the usefulness of generalizing relative risk ratios to a larger class of genetic models.

Appendix

Derivation of (3) and (4). Conditioning on I we have

$$f_R(Y_1, Y_2) = \sum_{i=0}^{2} P(I = i) f_R(Y_1, Y_2 | I = i) = \sum_{i=0}^{2} \alpha_{Ri} f_{Ri}(Y_1, Y_2). \quad \text{(A.1)}$$

Notice that (4) follows from (A.1) if we establish

$$f_{R1}(Y_1, Y_2) = f_{R0}(Y_1, Y_2) + 0.5\sigma_a^{(1,2)},$$
$$f_{R2}(Y_1, Y_2) = f_{R0}(Y_1, Y_2) + \sigma_a^{(1,2)} + \sigma_d^{(1,2)}. \quad \text{(A.2)}$$

Assume there are f founders in the pedigree to which the relative pair (1,2) belongs, and let $(a_1, \ldots, a_{2f})$ be a binary vector of length $2f$ containing the founder alleles. Under random mating $\{a_i\}$ are independent random variables with $P(a_i = 1) = p$ and $P(a_i = 0) = q$. Write $G_1 = (a_j a_k)$ and $G_2 = (a_l a_n)$ for the two genotypes of the relative pair, where $j, k, l, n \in \{1, \ldots, 2f\}$ and $I = 1_{\{j=l\}} + 1_{\{j=n\}} + 1_{\{k=l\}} + 1_{\{k=n\}}$. Let $a = (a_j, a_k, a_l, a_n)$ and write

$$f_R(Y_1, Y_2 | G_1, G_2) = h(a), \quad \text{(A.3)}$$

regarding (Y_1, Y_2) as fixed and (G_1, G_2) as varying. Assuming no imprinting, the order of the alleles within each genotype is immaterial, and hence h is determined by the 3×3 joint penetrance matrix

$$\begin{pmatrix} h(0,0,0,0) & h(0,0,0,1) & h(0,0,1,1) \\ h(0,1,0,0) & h(0,1,0,1) & h(0,1,1,1) \\ h(1,1,0,0) & h(1,1,0,1) & h(1,1,1,1) \end{pmatrix} = \begin{pmatrix} \psi_{00} & \psi_{01} & \psi_{02} \\ \psi_{10} & \psi_{11} & \psi_{12} \\ \psi_{20} & \psi_{21} & \psi_{22} \end{pmatrix} = \psi.$$

Let $U = \Re^3$ and $V = U \times U$ be the spaces of 1×3 vectors and 3×3 matrices respectively. Introduce the scalar product $< u, w >= q^2 \cdot u_0 w_0 + 2pq \cdot u_1 w_1 + p^2 \cdot u_2 w_2$ on U and

$$\begin{aligned}(\psi, \theta) = {} & q^2 \cdot q^2 \cdot \psi_{00}\theta_{00} + q^2 \cdot 2pq \cdot \psi_{01}\theta_{01} \\ & + q^2 \cdot p^2 \cdot \psi_{02}\theta_{02} + 2pq \cdot q^2 \cdot \psi_{10}\theta_{10} \\ & + 2pq \cdot 2pq \cdot \psi_{11}\theta_{11} + 2pq \cdot p^2 \cdot \psi_{12}\theta_{12} \\ & + p^2 \cdot q^2 \cdot \psi_{20}\theta_{20} + p^2 \cdot 2pq \cdot \psi_{21}\theta_{21} \\ & + p^2 \cdot p^2 \cdot \psi_{22}\theta_{22}\end{aligned}$$

on V, respectively. An orthonormal basis on U is

$$e_1 = (1, 1, 1),$$
$$e_2 = \frac{1}{\sqrt{2pq}}(-2p, q - p, 2q),$$
$$e_3 = (1/q - 1, -1, 1/p - 1),$$

see Hössjer (2003). Similarly, an orthonormal basis on V consists of the nine matrices $\{e_{ij} = e_i' e_j;\ i, j = 1, 2, 3\}$, so that for instance

$$e_{12} = \frac{1}{\sqrt{2pq}} \begin{pmatrix} -2p & q - p & 2q \\ -2p & q - p & 2q \\ -2p & q - p & 2q \end{pmatrix}$$

Let $\xi_i = (a_i - p)/\sqrt{pq}$, so that $\{\xi_i\}$ are i.i.d. random variables with zero mean and unit variance. The RHS

of (A.3) can be expanded into a sum of uncorrelated terms

$$\begin{aligned} h(a) = (\psi, e_{11}) &+ \frac{1}{\sqrt{2}}(\psi, e_{12})(\xi_l + \xi_n) + (\psi, e_{13})\xi_l\xi_n \\ &+ \frac{1}{\sqrt{2}}(\psi, e_{21})(\xi_j + \xi_k) \\ &+ \frac{1}{2}(\psi, e_{22})(\xi_j + \xi_k)(\xi_l + \xi_n) \\ &+ \frac{1}{\sqrt{2}}(\psi, e_{23})(\xi_j + \xi_k)\xi_l\xi_n + (\psi, e_{31})\xi_j\xi_k \\ &+ \tfrac{1}{\sqrt{2}}(\psi, e_{32})\xi_j\xi_k(\xi_l + \xi_n) + (\psi, e_{33})\xi_j\xi_k\xi_l\xi_n \end{aligned}$$

generalizing the corresponding expansion for U in Hössjer (2003, Lemma 1). See also the supplementary material of Hössjer (2005b) for expansions involving more than two individuals.

When $I = 0$, the indices j, k, l and n are all different. Using the zero mean, unit variance and independence of $\{\xi_i\}$,

$$\begin{aligned} f_{R0}(Y_1, Y_2) &= E(h(a)|\, I = 0) = E(h(a_j, a_k, a_l, a_n)) \\ &= (\psi, e_{11}) = u\psi u', \end{aligned}$$

proving (3). If $I = 1$ we may without loss of generality assume $j = l$ and $j \neq k \neq n \neq j$. Hence

$$\begin{aligned} f_{R1}(Y_1, Y_2) &= E(h(a)|\, I = 1) = E(h(a_j, a_k, a_j, a_n)) \\ &= (\psi, e_{11}) + 0.5(\psi, e_{22}) \\ &= f_{R0}(Y_1, Y_2) + 0.5u_a\psi u_a' \\ &= f_{R0}(Y_1, Y_2) + 0.5\sigma_a^{(1,2)}. \end{aligned}$$

Similarly, if $I = 2$ we assume $j = l$, $k = n$ and $j \neq k$ and obtain

$$\begin{aligned} f_{R2}(Y_1, Y_2) &= E(h(a)|\, I = 2) = E(h(a_j, a_k, a_j, a_k)) \\ &= (\psi, e_{11}) + (\psi, e_{22}) + (\psi, e_{33}) \\ &= f_{R0}(Y_1, Y_2) + u_a\psi u_a' + u_d\psi u_d' \\ &= f_{R0}(Y_1, Y_2) + \sigma_a^{(1,2)} + \sigma_d^{(1,2)}. \end{aligned}$$

The last two displayed equations prove (A.2). Finally notice that $\sigma_g^{(1,2)} = f_{R2}(Y_1, Y_2) - f_{R0}(Y_1, Y_2) = \sigma_a^{(1,2)} + \sigma_d^{(1,2)}$.

Derivation of (6). Let m be the number of meioses in the pedigree to which the relative pair (1,2) belongs and $v = (v_1, \ldots, v_m)$ the corresponding binary inheritance vector at the disease trait locus. Define $P(w) = P(v = w|\, Y_1, Y_2)$ for all 2^m binary vectors w of length m. This gives the posterior distribution, given phenotypes of the inheritance vector at the trait locus. The general expression for the effective number of meioses for testing in Hössjer (2005a) is

$$\mathrm{m}^{\mathrm{test}} = \log_2\left(2^m \sum_w P^2(w)\right) \tag{A.4}$$

for one pedigree. We will show that this expression coincides with (6) for a relative pair R. Let n_i be the number of inheritance vectors that give $I = i$ alleles IBD for the relative pair (1,2), so that $\alpha_{Ri} = n_i/2^m$, since the total number of possible inheritance vectors is 2^m. Further, let C_i be the set inheritance vectors corresponding to $I = i$ (hence $|C_i| = n_i$). Then $P(w) = 2^{-m}z_{Ri}/\alpha_{Ri}$ when $w \in C_i$, so that

$$\begin{aligned} \textstyle\sum_w P^2(w) &= 2^{-2m}\Big(n_0(z_{R0}/\alpha_{R0})^2 + n_1(z_{R1}/\alpha_{R1})^2 \\ &\qquad + n_2(z_{R2}/\alpha_{R2})^2\Big) \\ &= 2^{-m}\left(z_{R0}^2/\alpha_{R0} + z_{R1}^2/\alpha_{R1} + z_{R2}^2/\alpha_{R2}\right), \end{aligned}$$

Insertion of this expression into (A.4) gives (6).

Power Approximation Formula. According to Feingold *et al.* (1993) and Hössjer (2005c), the power $\beta = P_{H_1}(Z_{\max} \geq z)$ can be approximated as a function of the noncentrality parameter $\eta = E(Z(\tau))$ by

$$\beta \approx 1 - \Phi(z - \eta) + \phi(z - \eta)\left(\frac{2}{\eta d} - \frac{1}{\eta(2d - 1) + z}\right), \tag{A.5}$$

where $d = (-E'(Z(x)|_{x=\tau})/(2\rho_z\eta)$ is a normalized mean slope at the disease locus and ρ_Z is the crossover rate that measures the amount of fluctuations in the process Z. The threshold z is calculated so that the significance level $\alpha = P_{H_0}(Z_{\max} \geq z)$ attains a given value. To this end, we use the approximation

$$\alpha \approx 1 - \exp\left(-(1 - \Phi(z))\left(n_\Omega + 2\rho_Z L_\Omega z^2\right)\right). \tag{A.6}$$

defined by Lander & Kruglyak (1995).

Acknowledgement

The first author was sponsored by the National Research School in Genomics and Bioinformatics and the second by the Swedish Research Council, contract nr 626-2002-6286. We are grateful for the reviewers' helpful comments.

References

Almasy, L. & Blangero, J. (1998) Multipoint quantitative trait linkage analysis in general pedigrees. *Am J Hum Genet*, **62**, 1198–1211.

Amos, C. I. (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet*, **54**, 535–543.

Cardon, L. R. & Fulker, D. W. (1994) The power of interval mapping of quantitative trait loci using selected sib pairs. *Am J Hum Genet*, **55**, 825–833.

Carey, G. & Williamson, J. A. (1991) Linkage analysis of quantitative traits-increased power by using selected samples. *Am J Hum Genet*, **49**, 786–796.

Clerget-Darpoux, F., Bonaïti-Pellié, C., & Hochez, J. (1986) Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*, **42**, 393–399.

Donnelly, P. (1983) The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology*, **3**, 34–64.

Dudoit, S. & Speed, T. P. (2000) A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. *Biostatistics*, **1**, 1–26.

Elston, R., Olson, J. & Palmer, L. (2002) Biostatistical genetics and genetic epidemiology. Wiley reference in series in biostatistics.

Ewens, W. J. & Shute, N.C.E. (1986) A resolution of the ascertainment sampling problem. *I. Theory. Theor Popul Biol*, **30**, 388–412.

Fimmers, R., Seuchter, S. A., Neugebauer, M., Knapp, M. & Baur, M. P. (1989) Identity-by-descent analysis using complete high-resolutions. In Multipoint mapping and linkage based on affected pedigree members, eds.Elston, R. C. *et al*. Genetic Analysis Workshop, **6**, Liss, New York, 123–128.

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

Haseman, J. K. & Elston, R. C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet*, **2**, 2–19.

Hössjer, O. (2003) Determining inheritance distributions via stochastic penetrances. *J Amer Statist Assoc*, **98**, 1035–1051.

Hössjer, O. (2005a) Information and effective number of meioses in linkage analysis. *J Math Biol*, **50**(2), 208–232.

Hössjer, O. (2005b) Conditional likelihood score functions in linkage analysis. *Biostatistics*, **6**(2), 313–332.

Hössjer, O. (2005c) Spectral decomposition of score functions in linkage analysis. *Bernoulli* **11**(6), 1093–1113.

James, J. W. (1971) Frequency in relatives for an all-or-none trait. *Ann Hum Genet*, **35**, 47–48.

Kraft, P. & Thomas, D. C. (2000) Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet*, **66**, 1119–1131.

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

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

Lynch, M. & Walsh, B. (1998) Genetics and Analysis of quantitative traits. Sinauer Associates Inc.

Morton, N. E. (1955) Sequential tests for the detection of linkage. *Am J Hum Genet*, **7**(3), 7–318.

Risch, N. (1984) Segregation analysis incorporating genetic markers I. Single-locus models with an application to type I diabetes. *Am J Hum Genet*, **36**, 363–386.

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

Risch, N. (1990b) Linkage strategies for genetically complex traits II. The power of affected relative pairs. *Am J Hum Genet*, **46**, 229–241.

Risch, N. & Zhang, H. (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, **268**, 1584–1589.

Risch, N. & Zhang, H. (1996) Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *Am J Hum Genet*, **58**, 836–843.

Scham, P. C., Purcell, S., Cherny, S. S. & Abecasis, G. R. (2002) Powerful regression based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet*, **71**, 238–251.

Tang, H-K. & Siegmund, D. (2001) Mapping quantitative trait loci in oligogenic models. *Biostatistics*, **2**, 147–162.

Todorov, A. A. & Suarez, B. K. (2002) Liability model. In biostatistical Genetics and Genetic Epidemiology,Elston, R., Olson, J. & Palmer, L. (eds.), Wiley, 430–435.

Vieland, V. J. & Hodge, S. E. (1995) Inherent intractability of the ascertainment problem for pedigree data: A general likelihood framework. *Am J Hum Genet*, **56**, 33–43.

Vieland, V. J. & Hodge, S. E. (1996) The problem of ascertainment in linkage analysis. *Am J Hum Genet*, **58**, 1072–1084.

Weeks, D. & Lange, L. (1988) The affected-pedigree-member method of linkage analysis. *Am J Hum Genet*, **42**, 315–326.

Whittemore, A. & Halpern, J. (1994) A class of tests for linkage using affected pedigree members. *Biometrics*, **50**, 118–127.

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

Received: 9 June 2005
Accepted: 21 November 2005