Mathematical Statistics
Stockholm University

Modelling the Effect of Inbreeding Among Founders in Linkage Analysis

Ola Hössjer

Research Report 2005:1

ISSN 1650-0377

Postal address:

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

Internet:

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

Mathematical Statistics
Stockholm University
Research Report **2005:1**,
http://www.math.su.se/matstat

Modelling the Effect of Inbreeding Among Founders in Linkage Analysis

Ola Hössjer*

February 2005

Abstract

In this paper, we present a unified mathematical model for linkage analysis that allows for inbreeding among founders in all families. The identical by descent-configuration (IBD) of each pedigree is modeled as a Markov process containing two parameters; the inverse inbreeding and kinship coefficient and a rate parameter proportional to the inverse expected length of chromosome segments shared IBD by two different founder haplotypes. We use Hidden Markov Models and define a forward-backward algorithm for computing the conditional IBD-distribution given marker data, thereby extending the multipoint method of Lander and Green (1987) to situations where founders are inbred.

In principle, our methodology is valid for arbitrary pedigree structures, although for computational and storage reasons, the state space of the Markov process must be further reduced for pedigrees with many founders.

Simulation and theoretical approximations for nonparametric linkage analysis (NPL) based on affected sib pairs reveal that NPL scores are inflated and type 1 errors increased when parents are not genotyped and the inbreeding coefficient or rate parameter is underestimated. This effect is increases when the number of families gets large.

KEY WORDS: Founders, identical by descent, inbreeding, hidden Markov algorithm, nonparametric linkage analysis.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ola@math.su.se. Financial support from the Swedish Research Council, contract nr. 626-2002-6286.

1 Introduction

It is well known that linkage analysis is sensitive to misspecification of the inferred relationship between individuals belonging to the same family. For instance, an incorrect pedigree structure, such as non-paternity and unidentified adoption, for some families in the data set may affect both significance level and power (Ott, 1999, Chapter 11). When parental genotypes are available, these pedigree errors are often detectable by identification of genotypes that cause Mendelian inconsistencies (Boehnke and Guo, 1992, Ott, 1993, Stringham and Boehnke, 1996). When parental genotypes are missing, other methods have been developed for testing an assumed relationship between pairs of individuals. These methods use genomewide marker data, either for estimating the proportion of alleles that are identical-by-state of the pair (Ehn and Wagner, 1998) or for computing multipoint likelihoods for various types of pairs (Boehnke and Cox, 1997, Göring and Ott, 1997, McPeek and Sun, 2000). For large complex pedigrees with multiple inbreeding loops, correct modeling of relationships often increases the computational burden of the multipoint linkage algorithm. However, the cost of simplifying the pedigree structure is decreased power. This was noted by Dyer et al (2001) when using variance component techniques for detecting QTLs in a large Hutterite pedigree.

Even though the pedigree structure is correct, relationships of the founders might be misspecified. Founders are traditionally considered unrelated in linkage analysis, meaning that no founder alleles are identical-by-descent (IBD). However, if families belong to a population with some degree of inbreeding, there are often pairs of founder haplotypes that share long chromosomal segments (Broman and Weber, 1999). Miano et al. (2000) noticed that failure to model inbreeding of founders resulted in inflated lod scores for a sample of three families. As a result, type one errors are inflated if thresholds are not adjusted properly. The same is true for nonparametric linkage (NPL) (Génin and Clerget-Darpoux, 1996, 1998) and MLS scores (Leutenegger et al., 2002) based on affected sib pair families.

For samples of many small pedigrees, the potential effect of misspecifying inbreeding of founders can be much larger than pedigree errors. The reason is that pedigree errors typically affect a small proportion of families whereas misspecified founder relations can be present in all families.

In this paper we introduce a hidden Markov model for the IBD-configuration of a family with possible inbreeding among founders. It is valid for arbitrary family structures and contains i) the inheritance vector v, which specifies inheritance at all meioses in the pedigree (Donnely, 1983) and ii)

a Markov process u that models IBD-configuration of founder alleles. The latter process contains two main parameters; K, an integer which is the inverse of the inbreeding and kinship coefficient, and λ, a rate parameter which is inversely proportional to the average chromosomal length shared IBD by a pair of founder alleles. In the special case of one single founder, u reduces to the Markov model proposed by Stam (1980).

Viewing the IBD-configuration $w = (u, v)$ as the hidden Markov regime, we compute its conditional distribution given marker data. This multipoint approach to linkage analysis can be viewed as a generalization of the one proposed by Kruglyak at al. (1996), and implemented in the Genehunter, Allegro (Gudbjartsson et al., 2000) and Merlin (Abecasis et al., 2002) programs. In these programs founders are unrelated ($K = \infty$) and u uninformative. Then v is a hidden Markov regime under Haldane's map function of no interference and the forward-backward algorithm for Hidden Markov Models (Baum, 1972, Lander and Green, 1987) is used for computing the conditional distribution of v given marker data at all loci of interest. We extend the Lander-Green algorithm to inbreeding situations ($K < \infty$) and compute the conditional distribution of w given data at all loci of interest.

Our algorithm can be applied to a wide range of genetic models and score functions, including parametric and nonparametric linkage analysis as well as quantitative trait locus methods. In this paper, we consider NPL scores. Based on simulated data as well as theoretical approximations for affected sib pair families we show that NPL scores are inflated and significance level and power increased when the degree of inbreeding is underestimated and the average length of pairwise IBD-segments overestimated. When the parents are genotyped, these effects can be alleviated by careful choice of the score function.

The paper is organized as follows: In Section 2 we extend nonparametric linkage analysis to account for inbreeding among founders. The Markov model for the IBD-configuration along one chromosome is defined in Section 3. Simulation and computation of family scores are treated in Section 4, where in particular the HMM algorithm is defined. In Section 5, we extend the HMM algorithm to dense marker maps, and in the next two Sections 6 and 7, the effect of misspecifying K and λ and choosing score function are discussed. Numerical results are presented in Section 8 and a summary and various extensions of our work is provided in Section 9. Finally, some mathematical details are collected in the appendix.

2 Linkage Analysis and IBD-Configurations Allowing for Inbreeding

Consider a collection of N families with occurrence of a certain disease. For each family DNA marker data is collected for as many individuals as possible along one or several chromosomes. In addition, phenotypes are registered in all families. These are variables related to the disease. For instance, it could be quantitative, such as body mass index or glucose concentration, or a binary affection status indicator. Suppose we have wish to test presence of a disease causing gene τ along a genomic region Ω. We formulate this as an hypothesis testing problem, with null hypothesis H_0 that τ is unlinked to Ω, and alternative hypothesis H_1 that it is not:

$$\begin{aligned} H_0: &\quad \tau \in \Omega, \\ H_1: &\quad \tau \notin \Omega. \end{aligned}$$

Alternatively, we may view the testing procedure as one where H_0 is tested against a family of pointwise alternative hypotheses $H_1(x) : \tau = x$.

In NPL, a typical test statistic $Z(x)$ for discriminating between H_0 and $H_1(x)$ compares whether segregation of phenotypes in the N families are compatible with their IBD-configurations at x. For binary phenotypes, $Z(x)$ quantifies the amount of allele sharing IBD at x among affected pedigree members in the N families (Penrose, 1935, Whittemore and Halpern, 1994, Kruglyak et al., 1996, McPeek, 1999). More general NPL scores can be defined for a large class of phenotypes and genetic models by comparing the IBD-configuration at x for all families with their phenotypes, so that $Z(x)$ is large when individuals with concordant (discordant) phenotypes share more (less) alleles IBD than is expected by chance (Whittemore, 1996, Hössjer, 2003b, 2005). When testing H_0 against H_1, one uses maximal NPL score

$$Z_{\max} = \sup_{x \in \Omega} Z(x) \tag{1}$$

as test statistic. We reject H_0 when $Z_{\max}$ exceeds a given threshold z, giving significance level and power

$$\begin{aligned} \alpha(z) &= P_{H_0}(Z_{\max} \geq z) \\ \beta(z) &= P_{H_1}(Z'_{\max} \geq z), \end{aligned} \tag{2}$$

where $Z'_{\max}$ is defined as $Z_{\max}$, but with the maximum restricted to loci x on the same chromosome as τ. A large class of NPL score statistics

$$Z(x) = \sum_{i=1}^{N} \gamma_i Z_{\text{fam},i}(x), \tag{3}$$

are defined as a linear combination of the family scores $Z_{\text{fam},i}(x)$, with the possibility of assigning larger weights γ_i to more informative families. Here $Z_{\text{fam},i}(x)$ quantifies compatibility, at locus x, between phenotypes and IBD-pattern at x among the genotyped pedigree members in family i. See Weeks and Lange (1988), Fimmers et al. (1989), Whittemore and Halpern (1994) and Kruglyak et al. (1996).

In order to define the family scores, we omit family index i and consider a fixed family $\mathcal{P} = \{1, \ldots, n\}$ with n members. These are numbered so that the first f ones, $\mathcal{F} = \{1, \ldots, f\}$, are founders, i.e. have no ancestors in the pedigree, and the remaining $n-f$ ones, $\mathcal{N} = \{f+1, \ldots, n\}$, are nonfounders. Let $Y = (Y_1, \ldots, Y_n)$ be the collection of phenotypes in the family. The k^{th} component Y_k is the phenotype of individual k, which could be quantitative, binary ($Y_k = 0$ and $Y_k = 1$ for affected and unaffected pedigree members) or unknown.

When defining the IBD-configuration at x for a specific family, the founders are traditionally treated as unrelated, meaning that no founder alleles are IBD. With this assumption, two alleles in the pedigree can only be IBD if they receive the same founder alleles during transmission at locus x. This means that the IBD-configuration at x is a function of the inheritance vector $v(x)$ (Donnely, 1983). This is a binary vector of length $m = 2(n-f)$, where m is the number of meioses in the pedigree. It summarizes allele transmission in the pedigree at locus x, such that the s^{th} bit $v_s(x)$ is 0 or 1 depending on whether the s^{th} meiosis corresponds to a grandpaternal or grandmaternal allele being transmitted to the child. We denote the collection of all 2^m inheritance vectors as $\mathcal{V}$.

In this paper, we define the so called allele IBD-configuration at x more generally as

$$w(x) = (u(x), v(x)),$$

where $u(x)$ is the allele IBD-configuration of the f founders at x (Thompson, 1974). It can be written as a list of founder alleles $u = (b_1, b_2, \ldots, b_{2f})$ of a fully polymorphic marker, with b_{2k-1} and b_{2k} the paternal and maternal alleles of founder k. Hence, two alleles j and j' are IBD if and only if $b_j = b_{j'}$. The actual numbering of alleles is not important, any permutation of allele labels will do. Let $|u|$ denote the number of distinct founder alleles IBD of u. If founder alleles are numbered $\mathcal{A} = \{1, \ldots, 2f\}$, u defines a disjoint decomposition $\mathcal{A} = \cup_{l=1}^{|u|} U_l$, where IBD founder alleles belong to the same subset. This gives rise to an equivalent representation $u = \{U_1, \ldots, U_{|u|}\}$. Table 1 shows the set $\mathcal{U} = \{u\}$ of founder allele IBD-configurations when $f = 2$. This is the well known list of IBD-states for a pair of individuals

(Gillois, 1964, Jacquard, 1974). In Table 2, we list $|\mathcal{U}|$ as function of f. The collection $\mathcal{W} = \mathcal{U} \times \mathcal{V}$ of all possible allele IBD-configurations has size $|\mathcal{W}| = |\mathcal{U}| \times |\mathcal{V}| = 2^m |\mathcal{U}|$. It is seen that $|\mathcal{U}|$ and $|\mathcal{W}|$ increase rapidly with f. Later on, we will discuss various ways of reducing the size of the state space for larger pedigrees.

For each family, the NPL score $Z_{\text{fam}}(x)$ is based on i) recovering $w(x)$ from marker data $\boldsymbol{M}$ and ii) quantifying compatibility between $w(x)$ and Y by means of a real-valued score function $S(w) = S(w; Y)$, where $w = w(x)$ is regarded as random and Y as fixed.

For a single chromosome of length L Morgans, assume that $\boldsymbol{M}$ consists of H markers, located at $0 \leq x_1 < \ldots < x_H \leq L$. If the marker at x_h is d_h-allelic, we let $(a_{h,2k-1}a_{h,2k})$ be the marker genotype of $k \in \mathcal{P}$ at x_h, consisting of one maternal $a_{h,2k-1}$ and one paternal allele $a_{h,2k}$, both of which are coded as numbers from $\{0, 1, \ldots, d_h - 1\}$. The phase of the genotype is unknown without further information from previous generations, i.e. no imprinting is assumed. If $\mathcal{T} \subset \mathcal{P}$ is the set of genotyped individuals, $M_h = \{(a_{h,2k-1}a_{h,2k})\}_{k \in \mathcal{T}}$ is the marker data at x_h and $\boldsymbol{M} = (M_1, \ldots, M_H)$ the marker data at all loci along one chromosome.

Following Weeks and Lange (1988), Fimmers et al. (1989), Whittemore and Halpern (1994) and Kruglyak et al. (1996), the NPL score at x for one family is defined as

$$Z_{\text{fam}}(x) = (E(S(w(x)|\boldsymbol{M}) - \nu_S)/\kappa_{S,x}, \tag{4}$$

where standardization by $\nu_S = E_{H_0}(E(S(w(x))|\boldsymbol{M})) = E(S(w))$ and $\kappa^2_{S,x} = \text{Var}_{H_0}(E(S(w(x))|\boldsymbol{M}))$ ensures that $Z_{\text{fam}}(x)$ has zero mean and unit variance under H_0. The variance in the definition of $\kappa^2_{S,x}$ is taken with respect to variations in marker data $\boldsymbol{M}$, whereas the phenotype vector Y is fixed. Since this is often computationally involved, we will use the complete marker data approximation

$$\kappa^2_{S,x} = \kappa^2_S = \text{Var}_{H_0}(S(w(x))). \tag{5}$$

It can be shown that (5) implies $\text{Var}_{H_0}(Z_{\text{fam}}(x)) \leq 1$, with equality for complete marker data (Kruglyak et al., 1996).

By definition of the family scores, the total NPL score (3) satisfies $E_{H_0}(Z(x)) = 0$. It the weights are chosen so that $\sum_{i=1}^N \gamma_i^2 = 1$, we obtain $\text{Var}_{H_0}(Z_{\text{fam}}(x)) \leq 1$ when using the simplified standardization (5), with equality iff marker data at x is complete in all families. By the Central Limit Theorem, $Z(\cdot) = \{Z(x);\, 0 \leq x \leq L\}$ is approximately a zero mean Gaussian process under H_0, whose marginal distribution is standard normal $N(0, 1)$ when marker data is complete.

Another definition of the total NPL score, which is equivalent to (3) for complete marker data, was proposed by Kong and Cox (1997). It involves estimating an auxiliary parameter at each locus x.

A major goal of this paper is to investigate how the significance level and power (2) are affected by inbreeding among founders in all families. In fact, we will find that the distribution of $Z(\cdot)$ may change considerably when the founder allele IBD-configuration is misspecified, and this affects $\alpha(z)$ as well as $\beta(z)$.

3 Stochastic Model for Allele IBD-Configuration

In order to assess the statistical properties of the total NPL score (3) we need to define a stochastic model. For convenience, we consider one chromosome of length L Morgans and assume that the triplets $(w(\cdot), Y, \boldsymbol{M})$, with $w(\cdot) = \{w(x);\ 0 \le x \le L\}$, are independent between families. Thus we only consider one fixed family in this section.

3.1 IBD Configuration Process

We assume that $u(\cdot) = \{u(x);\ 0 \le x \le L\}$ and $v(\cdot) = \{v(x);\ 0 \le x \le L\}$ are two independent processes. Under Haldane's (1919) model of no interference, the m components of the latter process are independent Markov processes with state space $\{0, 1\}$ and transitions occurring according to a Poisson process with intensity 1. Hence, $v(\cdot)$ is a Markov process on $\{0, 1\}^{\otimes m}$ with intensity matrix

$$A_v(v, v') = \begin{cases} 1, & |v - v'| = 1, \\ 0, & |v - v'| = 2, \\ -m, & v = v', \end{cases}$$

and $|v - v'| = \sum_{j=1}^{m} |v_j - v'_j|$ is the Hamming distance between v and v'.

In order to define $u(\cdot)$, the founder genotypes of the males are drawn independently with replacement among the males in a population of size N_0. Similarly, the founder genotypes of the females are drawn from the females in the same population. The ancestral history of this population is defined by tracing allele transmission T generations backward in time. Let N_t be the population size in generation $t = 0, 1, \ldots, T$. We regard generation T as a founder population, and all of its $2N_T$ haplotypes as unrelated (non-IBD). Crossovers but not mutations are allowed for in formation of the successive generations $T - 1, \ldots, 0$. We may represent this by means of an ancestral

recombination graph $\mathcal{G}(\cdot) = \{\mathcal{G}(x);\, 0 \leq x \leq L\}$ (Hudson, 1983, Kaplan and Hudson, 1985, Griffiths, 1991), where $\mathcal{G}(x)$ is the coalescence process T generations back in time (Kingman, 1982). Two alleles from Generation 0 are IBD at x iff they originate from the same haplotype of Generation T at x. Then $u(\cdot)$ is a function of $\mathcal{G}(\cdot)$ and the f randomly picked individuals of Generation 0 that constitute the founders of the pedigree of interest. Let $K(x)$ be the number of haplotypes from Generation T that have survived down to Generation 0 at locus x. We assume that

1. $K(x) \equiv K$ at all x, and the descendants in Generation 0 of each of the K haplotypes of Generation T divide, at each locus x, the $2N_0$ alleles of Generation 0 into K equally large groups. Moreover, at each x, for a randomly chosen individual k among the males or females of Generation 0, the (ordered) pair of ancestral haplotypes transmitted to k at x is uniformly distributed on $\{1, \ldots, K\} \times \{1, \ldots, K\}$.

Assumption 1 gives the marginal distribution of $u(x)$ at all loci x: The $2f$ ancestral founder haplotype numbers at x are drawn independently and uniformly from $\{1, \ldots, K\}$, and alleles with the same number form the equivalence classes U_l of $u(x)$. Table 1 lists the marginal distribution when $f = 2$ for several values of K. It has previously been used in linkage analysis by Génin and Clerget-Darpoux (1996,1998) and Génin et al. (1998). These authors use the condensed list of nine genotype identity states, which they refer to as IBW states, rather than the 15 allele identity states.

Let F be the inbreeding coefficient, i.e. the probability that two alleles of a founder are IBD and ϕ the kinship coefficient, i.e. the probability that two randomly picked alleles, one from each of two distinct founders, are IBD. Then, it follows easily that

$$F = \phi = 1/K \tag{6}$$

independently of the founder or pair of founders chosen.

Even though $\mathcal{G}(\cdot)$ is a Markov process under Haldane's map function, $u(\cdot)$ need not be Markov. However, we will propose a simple model under which $u(\cdot)$ is Markov as well. Let $I_j(x) \in \{1, \ldots, K\}$ be the ancestral haplotype number assigned to the j^{th} founder allele at x and $I(x) = (I_1(x), \ldots, I_{2f}(x))$ be the collection of ancestral haplotype numbers of all founders at x. Then $u(x) = g(I(x))$ is a function of $I(x)$, with each $U_l \in u(x)$ defined as a set founder alleles j with identical $I_j(x)$.

The above discussion implies that $\{I_j(x)\}_{j=1}^{2f}$ are independent and uniformly distributed on $\{1, \ldots, K\}$ at all x. We will further assume that

2. $\{I_j(\cdot)\}_{j=1}^{2f}$ are independent and stationary Markov processes on $\{1,\ldots,K\}$ with the same intensity matrix $A_{I_j} = (A_{I_j}(i,i'))$, satisfying

$$A_{I_j}(i,i') = \begin{cases} \lambda/(K-1), & i \neq i', \\ -\lambda, & i = i', \end{cases}$$

where $\lambda > 0$ is a given constant.

The constant λ can be interpreted by considering the length L_{shared} of a segment shared IBD by a pair of founder alleles j and j'. Then, if chromosome boundary effects are ignored, L_{shared} is exponentially distributed with

$$E(L_{\text{shared}}) = 1/(2\lambda). \tag{7}$$

Similarly, the length of segments not shared IBD is also exponentially distributed, with mean $(1-F)/(2\lambda F)$.

More explicitly, we may reformulate the interpretation K and λ in terms of the ancestral recombination graph as follows: Consider two distinct founder alleles j and j'. Tracing their ancestry backward in time at locus x, let T_{coal} be the generation of coalescence. Hence T_{coal} is a function of j, j' and $\mathcal{G}(x)$. We assume $T_{\text{coal}} = \infty$ whenever $T_{\text{coal}} > T$, reflecting the fact that ancestral founder alleles from Generation T are regarded as non-IBD. Using interpretations (6), (7) of K and λ, it follows that

$$\begin{array}{rcl} K^{-1} & = & P(T_{\text{coal}} < \infty), \\ \lambda^{-1} & \geq & E(T_{\text{coal}}^{-1} | T_{\text{coal}} < \infty), \end{array} \tag{8}$$

where, in the second row, we used the fact that $2T_{\text{coal}}$ meioses are needed to join j and j', each one with intensity 1 to switch state from grandmaternal to grandpaternal transmission when Haldane's map function is used and map distance is measured in Morgans. Since there are switches that don't change the IBD-configuration of j and j' (Fisher, 1954), we have inequality in the second row of (8). In other words, an upper bound for λ is the harmonic mean of T_{coal} between two distinct founder alleles, conditional on the event that they are IBD.

An interesting consequence of (8) is that the same inbreeding coefficient can be obtained if j and j' have one or a few close common ancestors on one hand and many distant common ancestors on the other hand. In the latter case T_{coal} and λ will be larger and the mosaic of IBD and non-IBD segments shorter. See Leutenegger et al. (2003) for a similar discussion. In fact, using (8), the values of K and λ depend on our definition of IBD. If we increase

T, that is, choose to consider a larger population with ancestral founders further back in time, both K and λ will increase.

An immediate consequence of Assumption 2 is that $I(\cdot)$ is a stationary Markov process on $\{1,\ldots,K\}^{\otimes 2f}$ with intensity matrix

$$A_I(I,I') = \begin{cases} \lambda/(K-1), & |I-I'|=1, \\ 0, & |I-I'|\geq 2, \\ -2f\lambda, & I=I', \end{cases} \tag{9}$$

where $|I-I'| = \sum_{j=1}^{2f} 1_{\{I_j \neq I'_j\}}$. The following proposition states that $u(\cdot)$ and $w(\cdot)$ are Markov processes as well:

Proposition 1 *Under Assumption 2, $u(\cdot)$ is a stationary Markov process on $\mathcal{U}$ with intensity matrix $A_u = (A_u(u,u'))$, where*

$$A_u(u,u') = \sum_{I';g(I')=u'} A_I(I,I'), \quad u \neq u', \tag{10}$$

and I is chosen arbitrarily such that $g(I)=u$. Moreover, $w(\cdot)$ is a stationary Markov process on $\mathcal{W}$ with intensity matrix

$$A_w(w,w') = A_w((u,v),(u',v')) = \begin{cases} A_u(u,u'), & u \neq u', v = v', \\ A_v(v,v'), & u = u', v \neq v', \\ 0, & u \neq u' \text{ and } v \neq v', \end{cases} \tag{11}$$

The diagonal elements of both A_u and A_w are chosen so that the row sums $\sum_{u'} A_u(u,u')$ and $\sum_{w'} A_w(w,w')$ are zero.

Summarizing, the Markov model for $u(\cdot)$ and $w(\cdot)$ contains two main parameters, K and λ, with natural interpretations in terms of inverse inbreeding and kinship coefficient (6) and half of inverse average length of shared segments IBD (7). When $f=1$, this model reduces to the one considered by Stam (1980).

3.2 Phenotypes

Let $G = \{(a_{2k-1}a_{2k})\}_{k=1}^{2n}$ be the set of genotypes at the disease locus τ, with a_{2k-1} and a_{2k} the paternal and maternal alleles of k. Notice that G is a function of the founder alleles at the disease locus, $a = \{a_j\}_{j=1}^{2f}$, and $v(\tau)$, which we write as $G = G(a, v(\tau))$. For a d-allelic disease locus, we assume that each $a_j \in \{0,\ldots,d-1\}$. The genetic model consists of penetrance

parameters in $P(Y|G)$ and disease allele frequencies $p_r = P(a_j = r)$. We have that Y and $w(\cdot)$ are independent under H_0, whereas under H_1

$$P(Y|w(\cdot)) = P(Y|w(\tau)) = \sum_a P(Y|G(a, v(\tau)))P(a|u(\tau)). \tag{12}$$

In the last step we assumed that a and v are conditionally independent given u, which holds when there is no segregation distortion. We further assume that the members of the ancestral founder generation T are independent, so that the disease locus alleles are chosen independently among these individuals and then transmitted, without mutations, through T generations to $\mathcal{A}$. Hence

$$P(a|u) = \prod_{j \in \mathcal{A}(u)} p_{a_j} \tag{13}$$

whenever a_j in constant within each $U_l \in u$ and $P(a|u) = 0$ otherwise, and $\mathcal{A}(u) \subset \mathcal{A}$ is the set of founder alleles that contains (say) the smallest element from each $U_l \in u$.

For affected sib pairs, (12)-(13) was used by Génin and Clerget-Darpoux (1996) for calculating power for affected sib pair score functions.

3.3 Marker Data

Assuming markers are in linkage equilibrium (LE) with each other and with the disease locus, we get

$$P(\boldsymbol{M}|w(\cdot), Y) = P(\boldsymbol{M}|w(\cdot)) = \prod_{h=1}^{H} P(M_h|w(x_h)), \tag{14}$$

both under H_0 and H_1. This assumption holds, for instance, if the markers of the ancestral founder population are in LE and then the population sizes $N_{t-1}, \ldots, N_0$ of the next generations are large enough to prevent genetic drift from inducing linkage disequilibrium. Each term on the RHS of (14) is defined by conditioning on the vector $a_h = \{a_{hk}\}_{k=1}^{2f}$ of founder alleles at x_h,

$$\begin{array}{rcl} P(M_h|w(x_h)) & = & \sum_{a_h} P(M_h|a_h, v(x_h))P(a_h|u(x_h)) \\ & = & \sum_{a_h;(a_h,v(x_h))\to M_h} P(a_h|u(x_h)). \end{array} \tag{15}$$

The last equality reflects complete penetrance; each term $P(M_h|a_h, v(x_h))$ is one if M_h is compatible with $(a_h, v(x_h))$ and zero otherwise. The founder allele probability is defined as in (13) for each marker;

$$P(a_h|u) = \prod_{j \in \mathcal{A}(u)} p_{h,a_{hj}}, \tag{16}$$

with $\{p_{hr}\}_{r=0}^{d_h-1}$ are the marker allele frequencies at x_h.

4 Algorithms

4.1 Founder Phase Reduction

In order to decrease the state space $\mathcal{W}$ of $w(\cdot)$, we employ founder phase reduction (Kruglyak et al., 1996). Since the genotype of each founder $k \in \mathcal{F}$ has unknown phase, it is impossible to distinguish two IBD-patterns w and w', when the latter is obtained from the former by switching paternal and maternal alleles of founder k as well as the parental origin of all meioses transmitted from k to its children. We formalize this by letting π_k be the permutation of $\mathcal{A}$ that switches the two founder alleles of $k \in \mathcal{P}$, i.e. $\pi_k(2k-1) = 2k$, $\pi_k(2k) = 2k-1$ and $\pi_(j) = j$ for all other $j \in \mathcal{A}$. Then, π_k induces another mapping $\tilde{\pi}_k : \mathcal{U} \to \mathcal{U}$ by $\tilde{\pi}_k(u) = g(I \circ \pi_k)$ if $u = g(I)$. (It is easy to check that $\tilde{\pi}_k$ is well defined, i.e. not dependent upon the choice of I.) Moreover, let $c_k \in \mathcal{V}$ be the inheritance vector with ones in positions corresponding to children of k and zeros elsewhere. Then, if $w = (u, v)$, we let $w' = (\tilde{\pi}_k(u), v + c_k)$, where addition of inheritance vectors is componentwise modulo 2. Combining founder phase switching for different founders in all 2^f possible ways, we get equivalence classes of IBD states as follows: Given any binary vector $\xi = (\xi_1, \ldots, \xi_f)$ of length f, we define $\tilde{\pi}_\xi = \tilde{\pi}_1^{\xi_1} \circ \ldots \circ \tilde{\pi}_f^{\xi_f}$, $c_\xi = \sum_{k=1}^{f} \xi_k c_k$ and let

$$\bar{w} = \{(\tilde{\pi}_\xi(u), v + c_\xi);\ \xi \in \{0,1\}^{\otimes f}\} \tag{17}$$

be the equivalence class to which $w = (u, v)$ belongs. The resulting state space $\bar{\mathcal{W}} = \{\bar{w}\}$ has reduced size $|\bar{\mathcal{W}}| = 2^{m-f}|\mathcal{U}|$ instead of $|\mathcal{W}| = 2^m|\mathcal{U}|$. For instance, for a nuclear family with two children, we reduce the number of allele IBD-configurations from $16 \cdot 15 = 240$ to $4 \cdot 15 = 60$. Formally, we abbreviate (17) as $\bar{w} = (u, \bar{v})$, where $\bar{v} = \{v + c_\xi\}_\xi$ is the collection of inheritance vectors obtained by founder phase switching. Each $\bar{v} \in \bar{\mathcal{V}}$ can be represented as an inheritance vector of length $m - f$ as follows: There is a unique $v' \in \bar{v}$ which has zeros at the f bits corresponding to the first offspring of each founder. Then $\bar{v}$ is represented as the the remaining $m - f$ bits of v'.

It turns out that the Markov property is not lost during founder phase switching:

Proposition 2 *The founder phase reduced process $\bar{w}(\cdot) = \{\bar{w}(x);\ 0 \leq x \leq L\}$ is Markov with intensity matrix*

$$A_{\bar{w}}(\bar{w}, \bar{w}') = \sum_{w';\overline{w'}=\bar{w}'} A_w(w, w'), \tag{18}$$

where w is any IBD-configuration belonging to $\bar{w}$.

In absence of imprinting, phenotype and marker probabilities are invariant with respect to founder phase switching. We also impose the same (mild) requirement on the score function S and obtain

$$\begin{aligned} S(\bar{w}) &= S(w), \\ P(Y|\bar{w}(\cdot)) &= P(Y|w(\cdot)), \\ P(\boldsymbol{M}|\bar{w}(\cdot)) &= P(\boldsymbol{M}|w(\cdot)). \end{aligned} \tag{19}$$

Proposition 2 and (19) are utilized in the next two subsections to simulate marker data and compute family scores $Z_{\text{fam}}(x)$.

4.2 Simulation of Marker Data

Simulation of linkage scores under the null and alternative hypotheses has been considered by Boehnke (1986), Ploughman and Boehnke (1989), Ott (1989) and Terwilliger et al. (1993). We briefly show how to extend their results to incorporate inbreeding among founders.

When marker data $\boldsymbol{M}$ is simulated under H_0, it is convenient to generate the pair $(\boldsymbol{M}, \bar{w}(\cdot))$ using

$$P(\boldsymbol{M}, \bar{w}(\cdot)) = P(\bar{w}(\cdot)) \prod_{h=1}^{H} P(M_h|\bar{w}(x_h)). \tag{20}$$

The term $\bar{w}(\cdot)$ is simulated in the standard way as a finite state Markov process in continuous time, using Proposition 2. Given $\bar{w}(x_h)$, marker data at x_h is generated by first choosing (arbitrarily) any $w = (u, v)$ such that $\overline{(u, v)} = \bar{w}(x_h)$, then simulating marker founder alleles given u and segregating them to nonfounders according to v, using the third equation of (19) and $P(M_h, a_h|w) = P(a_h|u)P(M_h|a_h, v)$. The term $P(a_h|u)$ is given by (16), and founder alleles are generated independently in each group $U_l \in u$ according to the marker allele frequencies at x_h. The term $P(M_h|a_h, v)$ involves no simulation, since M_h is a deterministic function of a_h and v.

Simulation of $\boldsymbol{M}|Y$ under H_1 is similar, using

$$P(\boldsymbol{M}, \bar{w}(\cdot)|Y) = P(\bar{w}(\cdot)|Y) \prod_{h=1}^{H} P(M_h|\bar{w}(x_h)). \tag{21}$$

instead of (20). The first term on the RHS of (21) is simulated by first generating $\bar{w}(\tau)$ and then propagating $\bar{w}(\cdot)$ independently to the left and right from τ according to the same Markov process as under H_0, i.e.

$$P(\bar{w}(\cdot)|Y) = P(\bar{w}(\tau)|Y)P(\bar{w}(\tau-)|\bar{w}(\tau))P(\bar{w}(\tau+)|\bar{w}(\tau)),$$

where $\bar{w}(\tau-) = \{\bar{w}(x);\, 0 \le x < \tau\}$ and $\bar{w}(\tau+) = \{\bar{w}(x);\, \tau < x \le L\}$. When generating $\bar{w}(\tau)$, we use $P(\bar{w}(\tau)|Y) \propto P(\bar{w}(\tau))P(Y|\bar{w}(\tau))$ and (12).

4.3 HMM Algorithm for Family Score

Using (19), the family score (4) can be written

$$Z_{\text{fam}}(x) = (\sum_{\bar{w}} S(\bar{w})P(\bar{w}(x) = \bar{w}|\boldsymbol{M}) - \nu_S)/\kappa_{S,x}. \tag{22}$$

The inheritance distribution $P(\bar{w}(x) = \bar{w}|\boldsymbol{M})$ is evaluated at a grid of loci x of interest by applying the forward-backward algorithm for Hidden Markov Models. This extends work of Lander and Green (1987) from unrelated to related founder alleles. Our starting point is the Markov property of the hidden regime $\bar{w}(\cdot)$ in Proposition 2, together with the conditional distribution of $\boldsymbol{M}$ given $\bar{w}(\cdot)$, see (14) and (19). Suppose we wish to evaluate $Z(\cdot)$ on a grid $0 \le x^1 < \ldots < x^S \le L$. Let $\boldsymbol{M}_{s-} = \{M_h,\, 0 \le x_h \le x^s\}$, $\boldsymbol{M}_{s+} = \{M_h,\, x^s < x_h \le L\}$ and define forward probabilities $\alpha_s(\bar{w}) = P(\boldsymbol{M}_{s-}, \bar{w}(x_s) = \bar{w})$ and backward probabilities $\beta_s(\bar{w}) = P(\boldsymbol{M}_{s+}|\bar{w}(x_s) = \bar{w})$ for $s = 1, \ldots, S$. Then, because of Proposition 2, (14) and (19),

$$P(\bar{w}(x^s) = \bar{w}|\boldsymbol{M}) = \frac{\alpha_s(\bar{w})\beta_s(\bar{w})}{\sum_{\bar{w}' \in \bar{\mathcal{W}}} \alpha_s(\bar{w}')\beta_s(\bar{w}')}, \quad s = 1, \ldots, S. \tag{23}$$

The forward probabilities are computed recursively from left to right, and the backward probabilities recursively from right to left, see the appendix for details.

5 Dense Marker Map

When the number of markers is high, nearby markers will be in linkage disequilibrium, making the LE formula (14) a bit inaccurate (Schaid et al., 1994). However, the LE approximation is analytically much more tractable, allowing us to consider the limit of dense marker maps ($H \to \infty$, $\max_{1 \le h \le H-1}(x_{h+1} - x_h) \to 0$) theoretically. In fact, in many cases we obtain closed form approximations of the significance level and power (2) as function of the amount and type of inbreeding among founders (K and λ) and the set of genotyped pedigree members $\mathcal{T}$.

For a single chromosome of length L Morgans, we let $\boldsymbol{M} = \{M(x);\, 0 \le x \le L\}$ denote marker data when the marker map is dense. Intuitively, with H large, $M(x)$ can be thought of as a *combination* of M_h for marker loci x_h in

close vicinity of x. Using the LE assumption (14), such a group of marker loci can be regarded as one fully polymorphic marker at x when $H \to \infty$. This means that $M(x)$ is a genotype IBD-configuration of all genotyped family members. The identity $M_h = M(x_h)$ is only valid in the special case when the marker at x_h is fully polymorphic.

A genotype IBD-configuration M can be written $\{(b_{2k-1}b_{2k}),\, k \in \mathcal{T}\}$, where b_{2k-1} and b_{2k} are the maternal and paternal alleles (in unknown order, unless the is information from previous generations) for individual k at a fully polymorphic marker. That is, $b_j = b_{j'}$ iff the two alleles are IBD. Notice that any other representation of M, obtained by permuting allele labels, will do. For a nuclear family with two children, the marker genotype configurations are listed in Tables 3-5 when the children and all family members are genotyped respectively.

It is clear that

$$M = f(\bar{w}) \tag{24}$$

is a function of the founder phase reduced allele IBD-configuration of the whole pedigree. Therefore, the marker penetrance function is

$$P(\boldsymbol{M}|\bar{w}(\cdot)) = 1_{\{f(\bar{w}(x))=M(x) \text{ for all } 0\le x\le L\}}$$

for dense marker maps. To evaluate the inheritance distribution $P(\bar{w}(x) = \bar{w}|\boldsymbol{M})$ at a grid of points $0 \le x^1 < \ldots \le x^S \le L$, we use (23), provided the forward and backward probabilities are redefined as $\alpha_s(\bar{w}) = P(\boldsymbol{M}(x^s-), \bar{w}(x^s) = \bar{w})$ and $\beta_s(\bar{w}) = P(\boldsymbol{M}(x^s+)|\bar{w}(x^s) = \bar{w})$, where $\boldsymbol{M}(x-) = \{M(x'),\, 0 \le x' \le x\}$ and $\boldsymbol{M}(x+) = \{M(x'),\, x < x' \le L\}$. Recursive algorithms for computing forward and backward probabilities are described in the appendix.

6 Misspecifying K and λ

Given marker data, *computation* of family scores depends on on the *assumed* values of K and λ. From now on, we write K and λ for assumed and K_{true} and λ_{true} for true values. Writing $\nu_S = \nu_S(K)$ and $\kappa_S^2 = \kappa_S^2(K)$, the family score (22) with standardization (5) becomes

$$Z_{\text{fam}}(x; K, \lambda) = (\sum_{\bar{w}} P_{K,\lambda}(\bar{w}(x) = \bar{w}|\boldsymbol{M})S(\bar{w}) - \nu_S(K))/\kappa_S(K), \tag{25}$$

both for non-dense and dense marker maps.

When $K = \infty$, it suffices to consider the restriction of $S = S(\bar{w}) = S(u, \bar{v})$ to allele IBD-configurations with no inbreeding among founders (NIF). Since

there is only one founder allele configuration $u_\infty = (1, 2, \ldots, 2f)$ with positive probability (=1) when $K = \infty$, we can restrict the sum in (25) to NIF-configurations $\bar{w} = (u_\infty, \bar{v})$. With

$$S^{\mathrm{NIF}}(\bar{v}) = S(u_\infty, \bar{v}),$$

the family score becomes

$$Z_{\mathrm{fam}}(x; \infty) = (\sum_{\bar{v}} P_\infty(\bar{v}(x) = \bar{v}|\boldsymbol{M}) S^{\mathrm{NIF}}(\bar{v}) - \nu_S(\infty))/\kappa_S(\infty), \tag{26}$$

when $K = \infty$. This is the traditional definition used e.g. by Kruglyak et al. (1996) for non-dense marker maps $\boldsymbol{M} = \{M_h\}$, involving only the (founder phase reduced) inheritance vector $\bar{v}$. Notice however, for *dense* marker maps $\boldsymbol{M} = \{M(x)\}$ there is a positive probability that (26) is not well defined when $K_{\mathrm{true}} < \infty$[1].

When *generating* marker data $\boldsymbol{M}$, the *true* values K_{true} and λ_{true} are used in (20)-(21). Hence the statistical properties of the family scores as well as the total score $Z(\cdot)$ will depend on K, K_{true}, λ and λ_{true}. As a result, we will find that the power and significance level (2) are often quite sensitive to misspecification of K_{true} and λ_{true}.

7 Choosing Score Functions

Traditionally, most score functions used in linkage analysis are functions of the (founder phase reduced) inheritance vector $\bar{v}$. With a slight abuse of notation, write

$$S = S(\bar{v}) \tag{27}$$

to denote the fact that $S(\bar{w}) = S(u, \bar{v})$ is independent of u. We refer to (27) as a *transmission-based* score function. Another possibility is to let

$$S = S(M) \tag{28}$$

depend on the genotype IBD-configuration M among the genotyped pedigree members. By this we mean that $S(\bar{w})$ is constant over all sets $f^{-1}(M)$, with f as in (24). We refer to (28) as an *IBD-based* score function. When there is no inbreeding among founders, (27) and (28) are essentially equivalent, in

[1] More precisely, $P_\infty(\bar{v}(x) = \bar{v}|\boldsymbol{M})$ is not well defined when $P_{K_{\mathrm{true}}}(\boldsymbol{M}) = 0$. This happens when $K_{\mathrm{true}} < \infty$, $K = \infty$ and, for at least one x', $M(x')$ is such that several founder alleles are IBD.

the sense that any IBD-based score function is transmission-based, and any transmission-based score function used in practice is IBD-based.

With inbreeding among founders, (27) and (28) are no longer equivalent. To see this, consider a nuclear family with father ($k = 1$), mother ($k = 2$) and two children ($k = 3, 4$). The IBD-based score function S_{IBD} of Table 3 counts the number of alleles shared IBD by the two sibs, the so called mean sharing score function. However, it is possible to define a transmission-based score function that checks whether the parents transmit the same grandparental alleles or not to the children. Let $v = (v_1, v_2, v_3, v_4)$ be the inheritance vector, with v_1 and v_2 the outcomes of the paternal and maternal meioses of the $k = 3$ child and v_3 and v_4 the outcomes of the paternal and maternal meioses of the $k = 4$ child. Since $c_1 = (1, 0, 1, 0)$ and $c_2 = (0, 1, 0, 1)$, we have $\bar{v} = (\bar{v}_1, \bar{v}_2)$, where $\bar{v}_1$ is zero or one depending on whether the father transmits the same grandpaternal allele to his two children or not. Similarly, $\bar{v}_2$ is zero or one depending on whether the mother transmits the same grandpaternal allele to her children or not. Then

$$S_{\text{tr}}(\bar{v}) = 1_{\{\bar{v}_1=0\}} + 1_{\{\bar{v}_2=0\}} \tag{29}$$

is equivalent to S_{IBD} when there is no inbreeding among founders, i.e. $S_{\text{IBD}}^{\text{NIF}} = S_{\text{tr}}^{\text{NIF}}$. Still, $S_{\text{IBD}} \neq S_{\text{tr}}$, as can be seen by considering $u = (1123)$. Then $\bar{v} = (0, 0)$, $S_{\text{tr}}(\bar{v}) = 2$ and $\bar{v} = (1, 0)$, $S_{\text{tr}}(\bar{v}) = 1$ both correspond to $M = \{(12)(12)\}$, $S_{\text{IBD}}(M) = 2$.

An IBD-based score function is biologically more reasonable, since it is the accumulation of alleles IBD, not the non-uniform transmission, that determines the probability of certain phenotypes Y. In fact, $S(\bar{w}) = P(Y|\bar{w}(\tau) = \bar{w})$ is IBD- but not transmission-based

In (25), the sum $\sum_{\bar{w}} P_{K,\lambda}(\bar{w}(x) = \bar{w}|\boldsymbol{M})S(\bar{w})$ is less sensitive to misspecification of K_{true} and λ_{true} for an IBD-based score function, at least when there are many markers H. In fact, it can be shown that the sum converges to $S(M(x))$ in the limit of a dense marker map. The reason is that $M(x)$ can be recovered for a dense marker map and an IBD-based score function is, by definition, constant over $f^{-1}(M(x))$. On the other hand, the mean $\nu_S(K)$ and variance $\kappa_S^2(K)$ are independent of K for transmission- but not for IBD-based score functions. The reason is that the distribution of $\bar{v}$ is independent of K_{true}, whereas the distribution of M is not. For instance, the score function (29) satisfies $\nu_{S_{\text{tr}}} = 1$ and $\kappa_{S_{\text{tr}}} = 0.7071$. The corresponding values for S_{IBD} depend on K as shown in Table 4.

Ideally, we would prefer an IBD-based score function with little dependence of $\nu_S(K)$ and $\kappa_S^2(K)$ on K. This is possible at least when all family members

are genotyped. Write $M = M_{\mathcal{T}}$ to indicate that M depends on the set of genotyped individuals. Then, when all family members are genotyped, $M = M_{\mathcal{P}} = (M_{\mathcal{F}}, M_{\mathcal{N}})$, with $M_{\mathcal{F}}$ and $M_{\mathcal{N}}$ the genotype IBD-configurations of the founders and nonfounders respectively. It can be shown that the number $|M_{\mathcal{F}}|$ of distinct founder alleles is a minimal sufficient statistic for K (cf. Table 1). Given any IBD-based score function S, we let $\tilde{\nu}_S = E_{H_0}(S(M)||M_{\mathcal{F}}|)$ and $\tilde{\kappa}_S^2 = \mathrm{Var}_{H_0}(S(M)||M_{\mathcal{F}}|)$ and define the robustified version

$$\tilde{S}(M) = \begin{cases} \nu_S(\infty) + \kappa_S(\infty)(S(M) - \tilde{\nu}_S)/\tilde{\kappa}_S, & \tilde{\kappa}_S > 0, \\ \nu_S(\infty), & \tilde{\kappa}_S = 0, \end{cases} \tag{30}$$

of S. It it agrees with S when there is no inbreeding among founders, i.e. $S^{\mathrm{NIF}} = \tilde{S}^{\mathrm{NIF}}$. Moreover, $\nu_{\tilde{S}}(K) = \nu_{\tilde{S}}(\infty)$ for all K and $\kappa_{\tilde{S}}^2(K) \approx \kappa_{\tilde{S}}^2(\infty)$, with a difference of order $P(\tilde{\kappa}_{\tilde{S}}^2 = 0)$ that is often negligible. The robustified mean sharing score function $\tilde{S}_{\mathrm{IBD}}$ is listed in Table 5. It has $\nu_{\tilde{S}_{\mathrm{IBD}}}(K) = 1$ and $\kappa_{\tilde{S}_{\mathrm{IBD}}}(K) \approx 0.7071$ for all K.

8 Numerical Results

We simulated the NPL score (3) of $N = 1000$ affected sib pair families along one chromosome of length 150 cM, using equal weights $\gamma_i = 1/\sqrt{N}$. The three versions of the mean-sharing score function described in Section 7 were used; $S = S_{\mathrm{tr}}$, S_{IBD} or $\tilde{S}_{\mathrm{IBD}}$. Figures 1-2 display the NPL score for one such simulation under H_0 when all four and two family members (the sibs) are genotyped respectively. Figures 3-4 display another simulation under H_1, with a biallelic disease locus ($d = 2$) placed in the middle of the chromosome. The disease allele frequency $p = p_1$ was set to 0.1, and the penetrance parameters to $\psi_0 = \psi_1 = 0.1$ and $\psi_2 = 0.8$. Here ψ_j is the probability that an individual with j disease alleles in his or her genotype becomes affected. For each combination of hypothesis H_i, S and number of genotyped individuals, four marker maps were used; a less informative map with markers of heterozygosity 0.8 every 10 cM, a very informative map with markers of heterozygosity 0.9 every 1 cM, a dense marker map and finally, and ideal dense marker map requiring knowledge of $S(\bar{w}(x))$ at all loci x for all families[2]. The reason for including the ideal dense marker map is that its NPL scores can be analyzed theoretically. In all cases, $K = 10000$, $K_{\mathrm{true}} = 100$ and $\lambda = \lambda_{\mathrm{true}} = 10$.

By construction, $Z(\cdot)$ should be unbiased under H_0 at all loci when K_{true} and λ_{true} are correctly specified. As seen from Figures 1-2, misspecification

[2]In more detail, this means that each family score is given by (A.10).

of K_{true} leads to a strong upward bias of the dense NPL score of S_{IBD}. (That is, the NPL score based on a dense marker map.) A somewhat smaller upward bias can also be noted for the dense NPL scores based on S_{tr} or $\tilde{S}_{\text{IBD}}$ when the sibs are genotyped. When all four pedigree members are genotyped, neither $\tilde{S}_{\text{IBD}}$ nor S_{tr} gives any significant bias in the dense NPL score. Overall, $\tilde{S}_{\text{IBD}}$ is least affected by misspecification of K_{true}, followed by S_{tr} and S_{IBD}. This can be explained by looking at the ideal dense NPL scores: These are unbiased for S_{tr} and $\tilde{S}_{\text{IBD}}$, but has positive bias for S_{IBD}. On the other hand, the ideal dense and dense NPL scores agree only for S_{IBD} and, when all four family members are genotyped, for $\tilde{S}_{\text{IBD}}$. In the remaining cases the dense NPL score is upward biased compared to the ideal dense one. This implies that the dense NPL score is unbiased only in one case, for $\tilde{S}_{\text{IBD}}$ when all family members are genotyped (although it is nearly unbiased also for S_{tr} in this case when K is close to ∞). Similar phenomena can also be seen under H_1 in Figures 3-4. In this case case, the true peak of $Z(\cdot)$ in the middle of the chromosome is dominated by a false peak to the right for S_{IBD}, and, when only the sibs are genotyped, for S_{tr} and $\tilde{S}_{\text{IBD}}$.

Among the finite marker maps, the more informative one gives NPL scores very close to the dense marker map in all cases. The less informative one also gives NPL scores fairly close to the dense marker map in all cases except one; when S_{IBD} is used and all four family members are genotyped. This discrepancy might seem surprising, but we believe it to be caused by founder inbreeding being mixed up with IBS-sharing that is not IBD-sharing.

Next we studied the effect of misspecifying K_{true} on significance level $\alpha(z)$ and power $\beta(z)$ for a genomewide scan of affected sib pair families over all 22 autosomes, with chromosome lengths as in Ott (1999, Table 1.2). We used theoretical approximations defined in the appendix both for $\alpha(z)$ and $\beta(z)$. These are valid for ideal dense NPL scores, and hence also for dense NPL scores when either $S = S_{\text{IBD}}$ or when $S = \tilde{S}_{\text{IBD}}$ and all four pedigree members are genotyped. We initially assumed $K = \infty$[3], $K_{\text{true}} = 500$, $\lambda_{\text{true}} = 10$ and $N = 1000$, and then varied one of K_{true}, λ_{true} and N at a time. Figure 5 shows plots of $\alpha(z)$ against these three variables for S_{IBD} and $\tilde{S}_{\text{IBD}}$, when the threshold z is chosen so that $\alpha(z) = 0.05$ if $K_{\text{true}} = K = \infty$. We find that $\alpha(z)$ is nearly unaffected by misspecification of K_{true} for $\tilde{S}_{\text{IBD}}$, whereas it is dramatically inflated for S_{IBD} when either K_{true} is small or when any of λ_{true} and N are large. Similarly, inflation of the power can be seen in Figure 6 for S_{IBD} but hardly at all for $\tilde{S}_{\text{IBD}}$.

[3]To be exact, we consider the limit $K \to \infty$, since marker data with $K = \infty$-probability zero occurs with positive probability when $K_{\text{true}} < \infty$.

When K_{true} and/or λ_{true} are misspecified, there are three important quantities that may change and affect the significance level; $\mu_0 = E_{H_0}(Z(x))$, $\sigma_0 = \sqrt{\text{Var}_{H_0}(Z(x))}$ and the crossover rate ρ. The latter is defined in the appendix and quantifies the amount of fluctuation of $Z(\cdot)$. The larger ρ is, the larger is the effective number of independent tests along the genome, leading to a larger $\alpha(z)$. For S_{IBD}, choosing K too large implies that μ_0 increases notably from 0 whereas σ_0 decreases very little from 1. For $\tilde{S}_{\text{IBD}}$, both μ_0 and (to a very good approximation) σ_0 are unaffected by the choice of K. From Figure 7 we find that ρ is about the same for S_{IBD} and $\tilde{S}_{\text{IBD}}$. Hence we conclude that μ_0 is most important for explaining why $\alpha(z)$ is much more inflated for S_{IBD} than for $\tilde{S}_{\text{IBD}}$ when K is chosen too large or λ too small.

Génin and Clerget-Darpoux (1996,98) calculated pointwise power and significance level for several IBD-based score functions including S_{IBD}, assuming one fully polymorphic marker at τ. Our results in Figures 5-6 are extensions that i) take multiple testing into account and ii) include the robustified score function $\tilde{S}_{\text{IBD}}$.

9 Discussion

In this paper we have proposed a novel multipoint approach for nonparametric linkage analysis which allows for inbreeding among founders. It is valid for arbitrary pedigree structures and contains two parameters chosen by the user, the inverse inbreeding coefficient K and half the inverse expected length of segments shared IBD, λ. We have illustrated our methodology for affected sib pair families, using simulation of nonparametric linkage scores as well as theoretical approximations of significance levels and power. Our findings show, when founders are not genotyped, that choosing K and λ too large leads to inflated NPL scores Z as well as increased significance level and power. When founders are genotyped, these effects can be corrected for by careful choice of score function S. The inflation of NPL scores increases with sample size, since the cumulative effect of underestimating inbreeding in many families is larger then. Although we have focused on NPL in this paper, we believe our inbreeding multipoint approach can be applied to other kinds of linkage analysis as well. For instance, Abney et al. (2002) have recently developed a QTL mapping technique for large inbred pedigrees, with variance components allowing for inbreeding within individuals. For smaller pedigrees, we believe their approach can be combined with ours.

Our methodology is based on a simple two parameter Markov model for the founder allele IBD-configuration process $u(\cdot)$, which enables us to use HMMs

for calculating linkage scores. It can be generalized in several ways:

Firstly, the number of parameters can be enlarged, allowing e.g. for inbreeding and kinship coefficients to vary between founders and pairs of founders. See Weir (1994), Cannings (1998) and Leutenegger et al. (2002) for extensions along these lines when $f = 2$.

Secondly, time- and memory-constraints could be improved by state space reduction. The most time- and memory-consuming part involves $|\bar{\mathcal{W}}| \times |\bar{\mathcal{W}}|$-matrices in the forward-backward algorithm for HMMs, cf. (A.3)-(A.6). A nuclear family with two parents and $n' = n-2$ children has $m = 2n'$ meioses. Hence $|\bar{\mathcal{W}}| = 15 \cdot 2^{2n'-2}$, which is feasible for most nuclear families of practical interest. However, for larger pedigrees with many founders, the state space quickly becomes too large without further restrictions. To alleviate this, the state space $\mathcal{U}$ of founder allele IBD configurations can be decreased by requiring $|u| \geq r$ for some $2 \leq r \leq 2f$. When $r = 2f$, we are back to usual linkage analysis with no inbreeding among founder alleles. The total probability of all states that are removed is $O(K^{r-1-2f})$. If $\mathcal{U}_r$ and $\bar{\mathcal{W}}_r$ denote the reduced state spaces for u and $\bar{w}$, we have

$$\begin{aligned} |\mathcal{U}_{2f-1}| &= 2f^2 - f + 1, \\ |\mathcal{U}_{2f-2}| &= 2f^4 - 14f^3/3 + 11f^2/2 - 11f/6 + 1 \end{aligned}$$

and $|\bar{\mathcal{W}}_r| = 2^{m-f}|\mathcal{U}_r|$. This makes the multipoint approach computationally feasible not only for small but also for medium sized pedigrees when r equals $2f-1$ or $2f-2$. See Table 2 for numerical values.

Thirdly, the Markov process model for $I(\cdot)$ could be generalized, allowing more than one founder allele to change at a time. This would be more realistic, viewing $I(\cdot)$ as a function of the f randomly picked founder individuals and the ancestral recombination process $\mathcal{G}(\cdot)$.

Fourthly, the Markov assumption itself can be questioned. It leads to exponentially distributed segments of IBD-sharing between pairs of haplotypes, whereas a distribution with larger coefficient of variation, such a mixture of exponential distributions, would be more realistic. See for instance Chapman and Thompson (2003) for results along these lines. However, a more complicated non-Markov model for $u(\cdot)$ and $\bar{w}(\cdot)$ requires another method of computing family scores than presented here.

We have demonstrated that significance level and power are both sensitive to misspecification of K_{true} and λ_{true}. A conservative approach is to choose K and λ too small to avoid underestimating the level of inbreeding and length of shared segments IBD. We may also choose K from prior knowledge of

the inbreeding coefficient F through (6). The latter varies between populations and is often of the order 0.01 and 0.001 for small and large populations respectively (Morton, 1992, 2002, Morton and Teague, 1996). Another possibility is to estimate F and λ jointly from data using maximum likelihood. For such an estimator to be efficient, most or all of the founders have to be genotyped though. Leutenegger et al. (2003) have developed such an estimator for individuals ($N = f = 1$, $m = 0$). One could choose λ in advance and estimate F from data, see e.g. Ayres and Balding (1998). On the other hand, we have shown that careful choice of score function S decreases the effect of misspecifying K_{true} and λ_{true}, especially when all family members are genotyped

Finally, we want to emphasize that the dense marker HMM algorithm of Section 5 is of independent interest for linkage analysis (with or without inbreeding among founders). It does not require specification of marker allele frequencies, and in the future, with very dense marker maps, it will probably be possible to determine the piecewise constant genotype IBD-configuration process $\boldsymbol{M} = \{M(x)\}$ manually. When $K = \infty$, $M(\cdot)$ will often be Markov (as for affected sib pairs), and then family scores based on a dense marker HMM algorithm coincides with the much simpler (A.10), which does not require a HMM algorithm. However, the Markov property is not guaranteed in general (not even when $K = \infty$) and hence the dense marker HMM algorithm provides a general way of computing NPL family scores when marker information is complete given the set of genotyped individuals.

Appendix

In order to prove Propositions 1 and 2, we use the following lemma, which gives explicit conditions under which a function of a Markov process is Markov itself.

Lemma 1 *Let $X(\cdot) = \{X(t);\, 0 \le t \le L\}$ be a stationary Markov process in continuous time with finite state space $\mathcal{X}$ and intensity matrix A_X. Let $Y(t) = g(X(t))$. Then, a sufficient condition for $Y(\cdot) = \{Y(t);\, 0 \le t \le L\}$ to be Markov with intensity matrix A_Y is that*

$$A_Y(y, y') = \sum_{x';\, g(x')=y'} A_X(x, x') \tag{A.1}$$

is well defined, i.e. not dependent upon the choice of $x \in \mathcal{X}$ such that $g(x) = y$. Let $\{\mathcal{X}_l\}_{l=1}^n$ be a decomposition of $\mathcal{X}$ into equivalence classes, defined as

inverse images $g^{-1}(y)$ of g. Then, a sufficient condition for (A.1) to hold is the existence of J one-to-one mappings $h_j : \mathcal{X} \to \mathcal{X}$, $j = 1, \ldots, J$, such that

i) $h_j(\mathcal{X}_l) = \mathcal{X}_l$, $l = 1, \ldots, n$.

ii) *given any l and $x, x' \in \mathcal{X}_l$, there exists a sequence $j_1, \ldots, j_q$, $q = q(x, x')$, such that $h_{j_q} \circ \ldots \circ h_{j_1}(x) = x'$.*

iii) $A_X(x, x') = A_X(h_j(x), h_j(x'))$ *for all x, x'.*

Proof. The sufficiency of (A.1) for $Y(\cdot)$ to be Markov is well known, see e.g. Dudoit and Speed (1999) and Rosenblatt (1974). Hence we only need to prove that the RHS of (A.1) is independent of $x \in g^{-1}(y)$ when i)-iii) hold. Given $x_1 \in \mathcal{X}_l$, assume first that $x_2 = h_j(x_1)$ for some $l \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, J\}$. Then

$$\sum_{x';\, g(x')=y'} A_X(x_1, x') = \sum_{x';\, g(x')=y'} A_X(h_j(x_1), h_j(x')) = \sum_{x';\, g(x')=y'} A_X(x_2, x'), \tag{A.2}$$

using iii) in the first identity and i) in the second. For an arbitrary $x_2 \in \mathcal{X}_l$, we may find a sequence $j_1, \ldots, j_q$ satisfying ii). Repeating the argument in (A.2) q times we find that $\sum_{x';\, g(x')=y'} A_X(x_1, x') = \sum_{x';\, g(x')=y'} A_X(x_2, x')$. Since l and $x_1, x_2 \in \mathcal{X}_l$ were arbitrarily chosen, this proves the lemma. □

Proof of Proposition 1. In order to prove that $u(\cdot)$ is a Markov process, we will use Lemma 1 and verify (A.1) with $X(\cdot) = I(\cdot)$ and $Y(\cdot) = u(\cdot)$. This is equivalent to verifying that the RHS of (10) is independent of $I \in g^{-1}(u)$. In order to do so, we will establish i)-iii) with $J = K!$ and $\{h_j\}_{j=1}^{J}$ the set of permutations on $\{1, \ldots, K\}$. By definition of $u(x) = g(I(x))$, i) holds as well as ii) with $q = 1$. Since $A_I(I, I')$ only depends on $|I' - I|$ and $|h_j(I') - h_j(I)| = |I' - I|$, iii) follows. Finally, the fact that $w(\cdot)$ is a Markov process with intensity matrix (11) follows easily from the fact that $w(\cdot) = (u(\cdot), v(\cdot))$, with $u(\cdot)$ and $v(\cdot)$ two independent Markov processes. □

Proof of Proposition 2. We apply Lemma 1 with $X(\cdot) = w(\cdot)$ and $Y(\cdot) = \bar{w}(\cdot)$. We need to prove that the RHS of (A.1) is independent of $x \in g^{-1}(y)$, or equivalently that the RHS of (18) is independent of $w \in \bar{w}$. To do so, we establish i)-iii) with $J = f$ and $h_k(w) = h_k(u, v) = (\tilde{\pi}_k(u), v + c_k)$, $k = 1, \ldots, J$. Conditions i)-ii) follow directly from the definition of $\bar{w}$. In

order to establish iii), it suffices, in view of (11), to show that a) $A_u(u,u') = A_u(\tilde{\pi}_k(u), \tilde{\pi}_k(u'))$ and b) $A_v(v,v') = A_v(v+c_k, v'+c_k)$ for $k = 1, \ldots, f$. But

$$\begin{aligned} A_u(u,u') &= \textstyle\sum_{I_1';g(I_1')=u'} A_I(I_1, I_1') \\ A_u(\tilde{\pi}_k(u), \tilde{\pi}_k(u')) &= \textstyle\sum_{I_2';g(I_2')=\tilde{\pi}_k(u')} A_I(I_2, I_2'), \end{aligned}$$

for any I_1 and I_2 such that $g(I_1) = u$ and $g(I_2) = \tilde{\pi}_k(u)$. By definition of $\tilde{\pi}_k$, we may choose $I_2 = I_1 \circ \pi_k$. But

$$\sum_{I_1';g(I_1')=u'} A_I(I_1, I_1') = \sum_{I_1';g(I_1')=u'} A_I(I_1 \circ \pi_k, I_1' \circ \pi_k) = \sum_{I_2';g(I_2')=\tilde{\pi}_k(u')} A_I(I_2, I_2'),$$

proving a). To prove b), notice that $A_v(v,v')$ only depends on $|v'-v|$. Since $|(v'+c_k)-(v+c_k)| = |v'-v|$, b) follows. □

Recursion formulas for forward and backward probabilities. Consider first a finite marker map $\{x_h\}_{h=1}^H$. Let Q_h be a diagonal matrix of order $|\bar{\mathcal{W}}|$ with entries $Q_h(\bar{w},\bar{w}) = P(M_h|\bar{w}(x_h) = \bar{w})$ on the diagonal. Introduce artificial 'boundary grid points' $x^0 = 0$ and $x^{S+1} = L$. The forward probabilities are initialized as $\alpha_0(\bar{w}) = P(\bar{w})$ and then updated recursively as

$$\alpha_s(\bar{w}) = \sum_{\bar{w}'} \alpha_{s-1}(\bar{w}') P_s(\bar{w}', w), \quad s = 1, \ldots, S, \tag{A.3}$$

where $P_s = \{P_s(\bar{w},\bar{w}')\}$ is a $|\bar{\mathcal{W}}| \times |\bar{\mathcal{W}}|$-matrix, defined as

$$P_s = \begin{cases} \exp((x^{s-1} - x^s)A_{\bar{w}}), & \text{if } h_{s-1} = h_s, \\ \exp((x^{s-1} - x_{h_{s-1}+1})A_{\bar{w}})Q_{h_{s-1}+1} & \\ \quad \cdot \prod_{h=h_{s-1}+2}^{h_s} \exp((x_{h-1} - x_h)A_{\bar{w}})Q_h & \\ \quad \cdot \exp((x_{h_s} - x^s)A_{\bar{w}}), & \text{if } h_{s-1} < h_s, \end{cases} \tag{A.4}$$

for $s = 1, \ldots, S+1$, where h_s satisfies $h_0 = 0$, $h_{S+1} = H$ and $x_{h_s} \le x^s < x_{h_s+1}$ for $s = 1, \ldots, S$. To make the last definition well defined for all s, we introduce artificial 'boundary markers' $x_0 = 0$ and $x_{H+1} > L$. Similarly, the backward probabilities are initialized as $\beta_{S+1}(\bar{w}) = 1$, and then updated recursively through

$$\beta_s(\bar{w}) = \sum_{\bar{w}'} P_{s+1}(w,\bar{w}')\beta_{s+1}(\bar{w}'), \quad s = S, \ldots, 1. \tag{A.5}$$

For dense marker maps we proceed as follows: Assume there are n possible marker IBD configurations $M = M^1, \ldots, M^n$, and decompose $\bar{\mathcal{W}}$ into n components $\bar{\mathcal{W}}^i = \{\bar{w};\, f(\bar{w}) = M^i\}$, $i = 1, \ldots, n$, with f as in (24). Define $i(x) \in \{1, \ldots, n\}$ through $M(x) = M^{i(x)}$. Initialize the forward probabilities

as $\alpha_0(\bar{w}) \propto P(\bar{w})1_{\{\bar{w}\in\bar{\mathcal{W}}^{i(x^0)}\}}$. Since $\alpha_s(\bar{w}) = 0$ when $\bar{w} \notin \bar{\mathcal{W}}^{i(x^s)}$, we only need to update the forward probabilities using (A.3) when $\bar{w}' \in \bar{\mathcal{W}}^{i(x^{s-1})}$ and $\bar{w} \in \bar{\mathcal{W}}^{i(x^s)}$. Hence P_s is an $|\bar{\mathcal{W}}^{i(x^{s-1})}| \times |\bar{\mathcal{W}}^{i(x^s)}|$-matrix which we define as follows: Introduce sub-matrices $A_{ij} = \{A_{\bar{w}}(\bar{w}, \bar{w}');\ \bar{w} \in \bar{\mathcal{W}}^i \text{ and } \bar{w}' \in \bar{\mathcal{W}}^j\}$ of $A_{\bar{w}}$ for $i, j = 1, \ldots, n$. Since $\bar{w}(\cdot)$ is a piecewise constant Markov process, $M(\cdot)$ will be piecewise constant as well, with jumps at $0 < z_1 < \ldots < z_R < L$. Introduce artificial boundary 'jump points' $z_0 = 0$ and $z_{R+1} > L$, and let i_h be the constant value of $i(x)$ on (z_h, z_{h+1}) for $h = 0, \ldots, R-1$ and i_R the constant value of $i(x)$ on (z_R, L). Define h_s through $h_0 = 0$, $h_{S+1} = H$ and $z_{h_s} \le x^s < z_{h_s+1}$ for $s = 1, \ldots, S$. Then

$$P_s = \begin{cases} \exp((x^{s-1} - x^s)A_{i_{h_s}, i_{h_s}}), & \text{if } h_{s-1} = h_s, \\ \exp((x^{s-1} - z_{h_{s-1}+1})A_{i_{h_s}, i_{h_s}})A_{i_{h_s}, i_{h_s+1}} & \\ \quad \cdot \prod_{h=h_{s-1}+2}^{h_s} \exp((z_{h-1} - z_h)A_{i_{h-1}, i_{h-1}})A_{i_{h-1}, i_h} & \\ \quad \cdot \exp((z_{h_s} - x^s)A_{i_{h_s}, i_{h_s}}), & \text{if } h_{s-1} < h_s. \end{cases} \tag{A.6}$$

The backward probabilities are computed as before in (A.5), using (A.6) instead of (A.4). □

Analytical approximations of significance level and power for dense marker maps. Let $\mu_0 = E_{H_0}(Z(x))$, $\mu_1(x) = E_{H_1}(Z(x))$, $\mu_1 = \mu_1(\tau)$, $\mu_1' = \mu_1'(\tau)$, $\sigma_0^2 = \mathrm{Var}_{H_0}(Z(x))$ and $\sigma_1^2(x) = \mathrm{Var}_{H_1}(Z(x))$. We assume that $Z(\cdot)$ is stationary, with covariance function satisfying

$$r_Z(t) = \mathrm{Cov}_{H_0}(Z(x), Z(x+t)) = \sigma_0^2(1 - 2\rho|t|) + o(|t|) \tag{A.7}$$

for small lags t, where ρ is the crossover rate. This crucial assumption requires a dense set of markers. In the above formulas, expectation and covariance is with respect to K_{true} and λ_{true}, whereas the NPL score Z depends on K and (sometimes also) λ.

For large N, we approximate $Z(\cdot)$ by a Gaussian process. Moreover, assuming a sequence of contiguous alternatives the distribution of $Z(\cdot)$ under H_1 is asymptotically the same as that of $\{\mu_1(x) - \mu_0 + Z(x);\ 0 \le x \le L\}$ under H_0 (Feingold et al., 1993, Hössjer, 2003c). Hence we obtain the H_1-distribution of $Z(\cdot)$ from the H_0-distribution simply by adding the drift function $\mu_1(\cdot) - \mu_0$. In particular, this approximation entails $\sigma_1^2(\cdot) \approx \sigma_0^2$. Using results of Lander and Bolstein (1989), Feingold et al. (1993) and Lander and Kruglyak (1995), we can approximate the significance level by

$$\alpha(z) \approx 1 - \exp\left(-(1 - \Phi(z'))(C + 2\rho L_{\text{total}}(z')^2)\right) \tag{A.8}$$

where $z' = (z - \mu_0)/\sigma_0$, C is the number of chromosomes and L_{total} the total length of Ω. The power is approximated by

$$\beta(z) \approx 1 - \Phi(z' - \eta) + \varphi(z' - \eta)\left(\frac{2}{\eta d} - \frac{1}{\eta(2d-1) + z'}\right) \tag{A.9}$$

where Φ and ϕ are the standard normal distribution and density functions, $\eta = (\mu_1 - \mu_0)/\sigma_0$ is the noncentrality parameter and $d = -\mu_1'/(2\eta\sigma_0\rho)$ a normalized mean slope at the disease locus.

The quantities involved in (A.8)-(A.9) are computed as

$$\begin{aligned} \mu_j &= \textstyle\sum_{i=1}^N \gamma_i \mu_{ji}, \quad j = 0, 1, \\ \mu_1' &= \textstyle\sum_{i=1}^N \gamma_i \mu_{1i}', \\ \sigma_0^2 &= \textstyle\sum_{i=1}^N \gamma_i^2 \sigma_{0i}^2, \\ \rho &= \textstyle\sum_{i=1}^N \gamma_i^2 \sigma_{0i}^2 \rho_i / \sum_{i=1}^N \gamma_i^2 \sigma_{0i}^2, \end{aligned}$$

where μ_{0i}, μ_{1i}, μ_{1i}', σ_{0i}^2 and ρ_i are quantities defined for the i^{th} family score in the same way as the corresponding quantities of Z. In particular, for pedigrees of the same form with identical phenotype vectors Y and equal weighting $\gamma_i = 1/\sqrt{N}$, it follows that $\mu_j = \sqrt{N}\mu_{ji}$, $\mu_j' = \sqrt{N}\mu_{ji}'$, $\sigma_0^2 = \sigma_{0i}^2$ and $\rho = \rho_i$.

We assume familywise NPL scores of the form

$$Z_{\text{fam},i}(x) = R_i(\bar{w}_i(x)). \tag{A.10}$$

where $\bar{w}_i(\cdot)$ is the IBD-configuration process and R_i the standardized score function of family i. It is evident from (25) that (A.10) requires a dense set of markers $\boldsymbol{M} = \{M(x)\}$. However, this is not a sufficient condition. This follows from (4) and the fact that $M(\cdot)$, in general, is not a Markov process and then $P(\bar{w}(x) = \cdot|\boldsymbol{M}) \neq P(\bar{w}(x) = \cdot|M(x))$ is not a function of x alone. However, (A.10) holds if the unstandardized score function S_i of family i is IBD-based, i.e. $S_i(\bar{w}) = S(\bar{w}; Y_i) = S_i(M)$, where Y_i the phenotype vector of family i and $M = f(\bar{w})$ the marker genotype IBD-configuration corresponding to $\bar{w}$. Then (A.10) holds with

$$R_i(\bar{w}) = (S_i(\bar{w}) - \nu_{S_i})/\kappa_{S_i}.$$

Since $\nu_{S_i} = \nu_{S_i}(K)$ and $\kappa_{S_i} = \kappa_{S_i}(K)$, it follows that R_i as well as the family scores (A.10) depend on K but not on λ

Put $P_{0i}(\bar{w}) = P_{H_0}(\bar{w}_i(x) = \bar{w})$, $P_{1i}(\bar{w}) = P_{H_1}(\bar{w}_i(\tau) = \bar{w}|Y_i)$ and let $A_{\bar{w}_i}$ be the intensity matrix of $\bar{w}_i(\cdot)$. These three quantities all depend on K_{true}

and λ_{true}. Then, generalizing results from Hössjer (2003a) and Ängquist and Hössjer (2005), we find that

$$\begin{aligned}
\mu_{0i} &= \textstyle\sum_{\bar{w}} R_i(\bar{w}) P_{0i}(\bar{w}), \\
\mu_{1i} &= \textstyle\sum_{\bar{w}} R_i(\bar{w}) P_{1i}(\bar{w}), \\
\sigma^2_{0i} &= \textstyle\sum_{\bar{w}} (R_i(\bar{w}) - \mu_{0i})^2 P_{0i}(\bar{w}), \\
\mu'_{1i} &= \textstyle\sum \sum_{\bar{w} \neq \bar{w}'} P_{1i}(\bar{w}) A_{\bar{w}_i}(\bar{w}, \bar{w}') (R_i(\bar{w}') - R_i(\bar{w})), \\
\rho_i &= 0.25 \cdot \textstyle\sum \sum_{\bar{w} \neq \bar{w}'} P_{0i}(\bar{w}) A_{\bar{w}_i}(\bar{w}, \bar{w}') (R_i(\bar{w}') - R_i(\bar{w}))^2.
\end{aligned} \tag{A.11}$$

We emphasize that $\mu_{0i} = \mu = 0$ and $\sigma_{0i} = \sigma_0 = 1$ when $K = K_{\text{true}}$, but not necessarily when K_{true} is misspecified.

More refined analytical approximations of power and significance level can be defined, with adjustment for non-dense marker maps and/or non-normality, cf. Feingold et al. (1993), Tu and Siegmund (1999), Tang and Siegmund (2001) and Ängquist and Hössjer (2005). Alternatively, the significance level may be computed by importance sampling for any kind of marker data, as in Ängquist and Hössjer (2004). □

References

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002). Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97-101.

Abney, M., Ober, C. and McPeek, M.S. (2002). Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting, serum-insulin level in the Hutterites. *Am. J. Hum. Genet.*, **70**, 920-934.

Ängquist, L. and Hössjer, O. (2004). Using importance sampling to improve simulation in linkage analysis. *Stat. Appl. Gen. Mol. Biol.* **3**, article 5.

Ängquist, L. and Hössjer, O. (2005). Improving the calculation of statistical significance in genome-wide scans. To appear in *Biostatistics*.

Ayres, K.L. and Balding, D.J. (1998). Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769-777.

Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1-8.

Boehnke, M. (1986). Estimating the power of a proposed linkage study: A practical computer simulation approach. *Am. J. Hum. Genet.*, **39**, 513-527.

Boehnke, M. and Cox, N.J. (1997). Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.*, **61**, 423-429.

Boehnke, M. and Guo, S-W. (1992). Statistical approaches to identify marker typing error in linkage analysis. *Am. J. Hum. Genet.*, **51**(Suppl), A183.

Broman, K.W. and Weber, J.L. (1999). Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. *Am. J. Hum. Genet.*, **65**, 1493-1500.

Cannings, C. (1998). On the probabilities of identity states in permutable populations. *Am. J. Hum. Genet.*, **62**, 698-702.

Chapman, N.H. and Thompson, E.A. (2003). A model for the length of tracts of identity by descent in finite random mating populations. *Theor. Pop. Biol.* **64**, 141-150.

Donnelly, K. (1983). The probability that some related individuals share some section of the genome identical by descent. *Theor. Pop. Biol.*, **23**, 34-64.

Dudoit, S. and Speed, T. (1999). A score test for linkage using identity by descent data from sibships. *Ann. Statist.* **27**, 943-986.

Dyer, T.D., Blangero, J., Williams, J.T., Göring, H.H. and Mahaney, M.C. (2001). The effect of pedigree complexity on quantitative trait linkage analysis. *Genet. Epidemiology* **21**(Suppl 1), S236-243.

Ehm, M.G. and Wagner, M. (1998). A test statistic to detect errors in sib-pair relationships. *Am. J. Hum. Genet.*, **62**, 181-188.

Feingold, E., Brown, P.O. and 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.

Fimmers, R., Seuchter, S.A., Neugebauer, M., Knapp, M. and 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.

Fisher, R.A. (1954). A fuller theory of junctions in inbreeding. *Heredity* **8**, 187-197.

Génin, E. and Clerget-Darpoux, F. (1996). Consanguinity and the sib-pair method: an approach using identity by descent between and within individuals. *Am. J. Hum. Genet.*, **59**, 1149-1162.

Génin, E. and Clerget-Darpoux, F. (1998). Reply to Weeks and Sinsheimer. *Am. J. Hum. Genet.*, **62**, 731-736.

Génin, E., Quesneville, H. and Clerget-Darpoux, F. (1998). On the probability of identity states in permutable populations: Reply to Cannings. *Am. J. Hum. Genet.*, **62**, 726-727.

Gillois, M. (1964). La relation d'indentité en génétique. Thèse Faculté des Sciences de Paris, Paris.

Griffiths, R.C. (1991). The two-locus ancestral graph. In: Basawa I.V. and Taylor, R.I. (eds.) *Selected Proceedings of the Sheffield Symposium on Applied Probability*, pp. 100-117. Lecture Notes-Monograph Series, vol. 18 Hayward, CA:IMS.

Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**, 12-13.

Göring, H.H. and Ott, J. (1997). Relationship estimation in affected sib pair analysis of late-onset diseases. *Eur. J. Hum. Genet.*, **5**, 69-77.

Haldane, J.B.S. (1919). The combination of linkage values and the calculation of distances between loci of unlinked factors. *J. Genetics* **8**, 299-309.

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. Mathematical Statistics, Stockholm University, Report 2003:21.

Hössjer, O. (2005). Conditional likelihood score functions in linkage analysis. To appear in *Biostatistics*.

Hudson, R.R. (1983). Properties of neutral allele model with intragenic recombination. *Theor. Pop. Biol.*, **23**, 183-201.

Jacquard, A. (1974). *The genetic structure of populations*. Springer-Verlag, New York.

Kaplan, N.L. and Hudson, R.R. (1985). The use of sample genealogies for studying a selectively neutral m-loci model with recombination. *Theor. Pop. Biol.*, **28**, 382-396.

Kingman, J.F.C. (1982). On the genealogy of large populations. In Gani J. and Hannan, E.J. (eds.), *Essays in Statistical Science: Papers in Honours of P.A.P. Moran*, pp. 97-112, North-Holland Publishing, Amsterdam.

Kong, A. and Cox, N.J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* **61**, 1179-1188.

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.

Lander, E. and Bolstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199.

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

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

Leutenegger, A-L., Génin, E., Thompson, E. and Clerget-Darpoux, F. (2002). Impact of parental relationship in maximum lod score affected sib pair method. *Genet. Epidemol.* **23**, 413-425.

Leutenegger, A-L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F. and Thompson, E. (2003). Estimation of inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.*, **73**, 516-523.

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

McPeek, M.S. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.*, **66**, 1076-1094.

Miano, M.G. et al. (2000). Pitfalls in homozygosity mapping. *Am. J. Hum. Genet.*, **67**, 1348-1351.

Morton, N.E. (1992). Genetic structure of forensic populations. *Proc. Natl. Acad. Sci. USA* **89**, 2556-2560.

Morton, N.E. (2002). Applications and extensions of Malecot's work in human genetics. In *Modern Developments in Theoretical Populations Genetics, the legacy of Gustave Malecót*, eds. Slatkin, M. and Veuille, M., Oxford University Press, Oxford, 20-36.

Morton, N.E. and Teague, J.W. (1996). Kinship, inbreeding and matching probabilities. *Molecular biology and human diversity, Symposium of the Society for the Study of Human Biology* **38**, Cambridge University Press, Cambridge, 51-62.

Ott, J. (1989). Computer-simulation methods in human linkage analysis. *Proc. Natl. Acad. Sci. USA* **86**(11), 4175-4178.

Ott, J. (1993). Detecting marker inconsistencies in human gene mapping. *Human Heredity* **43**, 25-30.

Ott, J. (1999). *Analysis of human genetic linkage*, 3rd ed. John Hopkins University Press, Baltimore.

Penrose, L.S. (1935). The detection of autosomal linkage in data which consists of brothers and sisters of unspecified parentage. *Annals of Eugenics* **6**, 133-138.

Ploughman, L.M. and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex trait. *Am. J. Hum. Genet.*, **44**, 543-551.

Rosenblatt, M. (1974). *Random Processes. Graduate Texts in Mathematics* **17**, 2nd ed. Springer, New York.

Schaid, D.J. et al. (2004). Comparison of microsatellites versus single-nucleotid polymorphisms in a genome linkage screen for prostate cancer-susceptibiliby loci. *Am. J. Hum. Genet.*, **75**, 948-965.

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genet. Res. Camb.* **35**, 131-155.

Stringhman, H.M. and Boehnke, M. (1996). Identifying marker typing incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, **59**, 946-950.

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

Terwilliger, J.D., Speer, M. and Ott, J. (1993). Chromosome-based method for rapid computer simulation in human linkage analysis. *Genet. Epidimiol.* **10**, 217-224.

Thompson, E.A. (1974). Gene identities and multiple relationships. *Biometrics*, **30**, 667-680.

Tu, I.P. and Siegmund, D. (1999). The maximum of a function of a Markov chain and applications to linkage analysis. *Adv. Appl. Prob.* **31**, 510-531.

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

Weir, B. (1994). The effect of inbreeding on forensic calculations. *Annu. Rev. Genet.* **28**, 597-621.

Whittemore, A. and 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 overview. *Biometrics* **59**, 704-716.

u	$\{U_l\}$	$M_{\mathcal{F}}$	$P(u)$			
			K	$K=10$	$K=100$	$K=1000$
(1111)	(1234)	{(11),(11)}	K^{-3}	0.0010	10^{-6}	10^{-9}
(1112)	(123)(4)	{(11),(12)}	$(K-1)K^{-3}$	0.0090	$9.9 \cdot 10^{-5}$	10^{-6}
(1121)	(124)(3)	{(11),(12)}	$(K-1)K^{-3}$	0.0090	$9.9 \cdot 10^{-5}$	10^{-6}
(1211)	(134)(2)	{(12),(11)}	$(K-1)K^{-3}$	0.0090	$9.9 \cdot 10^{-5}$	10^{-6}
(2111)	(1)(234)	{(12),(11)}	$(K-1)K^{-3}$	0.0090	$9.9 \cdot 10^{-5}$	10^{-6}
(1122)	(12)(34)	{(11),(22)}	$(K-1)K^{-3}$	0.0090	$9.9 \cdot 10^{-5}$	10^{-6}
(1212)	(13)(24)	{(12),(12)}	$(K-1)K^{-3}$	0.0090	$9.9 \cdot 10^{-5}$	10^{-6}
(1221)	(14)(23)	{(12),(12)}	$(K-1)K^{-3}$	0.0090	$9.9 \cdot 10^{-5}$	10^{-6}
(1123)	(12)(3)(4)	{(11),(23)}	$(K-1)(K-2)K^{-3}$	0.072	0.0097	0.0010
(1213)	(13)(2)(4)	{(12),(13)}	$(K-1)(K-2)K^{-3}$	0.072	0.0097	0.0010
(1231)	(14)(2)(3)	{(12),(13)}	$(K-1)(K-2)K^{-3}$	0.072	0.0097	0.0010
(1233)	(1)(2)(34)	{(12),(33)}	$(K-1)(K-2)K^{-3}$	0.072	0.0097	0.0010
(2131)	(1)(24)(3)	{(12),(13)}	$(K-1)(K-2)K^{-3}$	0.072	0.0097	0.0010
(1231)	(1)(23)(4)	{(12),(13)}	$(K-1)(K-2)K^{-3}$	0.072	0.0097	0.0010
(1234)	(1)(2)(3)(4)	{(12),(34)}	$(K-1)(K-2)(K-3)K^{-3}$	0.5040	0.9411	0.9940

Table 1: Founder allele IBD-configurations u and founder genotype IBD-configurations $M_{\mathcal{F}} = M_{\mathcal{F}}(u)$ when $f = 2$, as well as probabilities $P(u)$ for various values of K.

f	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{U}_{2f-2}\rvert$	$\lvert\mathcal{U}_{2f-1}\rvert$
1	2	2	2
2	15	14	7
3	203	81	16
4	4140	295	29
5	115 975	796	46
6	4 213 597	1772	67
7	190 899 322	3459	92

Table 2: The number of possible allele IBD-configurations among f founders when no restrictions are imposed ($|\mathcal{U}|$) and with at least r different alleles IBD ($|\mathcal{U}_r|$), $r = 2f-2$, $2f-1$.

M	$S_{\mathrm{IBD}}(M)$
$\{(12),(34)\}$	0
$\{(12),(13)\}$	1
$\{(12),(12)\}$	2
$\{(11),(22)\}$	0
$\{(11),(23)\}$	0
$\{(12),(33)\}$	0
$\{(11),(12)\}$	1
$\{(12),(11)\}$	1
$\{(11),(11)\}$	2

Table 3: List of all 9 marker IBD configurations $M = M_{\mathcal{N}} = \{(b_5 b_6),(b_7 b_8)\}$ for a nuclear family when the two children ($k = 3, 4$) are genotyped, as well as the mean sharing score function $S_{\mathrm{IBD}}(M)$

K	K_{true}	$\nu_{S_{\text{IBD}}}(K)$	$\kappa_{S_{\text{IBD}}}(K)$	N	μ_0	σ_0
∞	10	1	0.7071	100	1.9870	0.9479
	100				0.2107	0.9950
	1000				0.0212	0.9995
	∞				0	1
	10			1000	6.2834	0.9479
	100				0.6664	0.9950
	1000				0.0670	0.9995
	∞				0	1
100	10	1.0149	0.7035	100	1.7852	0.9527
	100				0	1
	1000				-0.1905	1.0046
	∞				-0.2118	1.0051
	10			1000	5.6454	0.9527
	100				0	1
	1000				-0.6024	1.0046
	∞				-0.6697	1.0051

Table 4: Standardizing constants ν_S and κ_S for the IBD-based mean sharing score function $S = S_{\text{IBD}}$, as well as the mean and standard deviations $\mu_0 = E_{H_0}(Z(x))$ and $\sigma_0 = \sqrt{\text{Var}_{H_0}(Z(x))}$ for the NPL score based on N affected sib pairs when a dense marker map is used. Formulas for μ_0 and σ_0 are given in the appendix. The former simplifies to $\mu_0 = \sqrt{N}(\nu_S(K_{\text{true}}) - \nu_S(K))/\kappa_S(K))$ in this case.

$\lvert M_{\mathcal{F}}\rvert$	M	$\tilde{S}_{\mathrm{IBD}}(M)$	$\lvert M_{\mathcal{F}}\rvert$	M	$\tilde{S}_{\mathrm{IBD}}(M)$
4	{(12),(34),(13),(24)}	0	3	{(11),(23),(12),(12)}	1.8911
	{(12),(34),(13),(14)}	1		{(11),(23),(12),(13)}	0.7030
	{(12),(34),(13),(23)}	1		{(12),(33),(13),(13)}	1.8911
	{(12),(34),(13),(13)}	2		{(12),(33),(13),(23)}	0.7030
3	{(12),(13),(11),(11)}	1.8911	2	{(12),(12),(11),(11)}	1.6236
	{(12),(13),(11),(12)}	0.7030		{(12),(12),(11),(12)}	0.3764
	{(12),(13),(12),(11)}	0.7030		{(12),(12),(12),(11)}	0.3764
	{(12),(13),(11),(13)}	0.7030		{(12),(12),(12),(12)}	1.6236
	{(12),(13),(13),(11)}	0.7030		{(12),(12),(11),(22)}	-0.8708
	{(12),(13),(12),(12)}	1.8911		{(11),(22),(12),(12)}	1.6236
	{(12),(13),(13),(13)}	1.8911		{(11),(12),(11),(11)}	1.6236
	{(12),(13),(11),(23)}	-0.4852		{(11),(12),(11),(12)}	0.3764
	{(12),(13),(23),(11)}	-0.4852		{(11),(12),(12),(11)}	0.3764
	{(12),(13),(12),(13)}	0.7030		{(11),(12),(12),(12)}	1.6236
	{(12),(13),(13),(12)}	0.7030		{(12),(11),(11),(11)}	1.6236
	{(12),(13),(12),(23)}	0.7030		{(12),(11),(11),(12)}	0.3764
	{(12),(13),(23),(12)}	0.7030		{(12),(11),(12),(11)}	0.3764
	{(12),(13),(13),(23)}	0.7030		{(12),(11),(12),(12)}	1.6236
	{(12),(13),(23),(13)}	0.7030	1	{(11),(11),(11),(11)}	1
	{(12),(13),(23),(23)}	1.8911			

Table 5: List of all 39 marker IBD configurations $M = M_{\mathcal{P}} = \{(b_1b_2),(b_3b_4),(b_5b_6),(b_7b_8)\}$ of a nuclear family when all family members ($k = 1,2,3,4$) are genotyped, as well as the robustified mean sharing score function $\tilde{S}_{\mathrm{IBD}}(M)$.

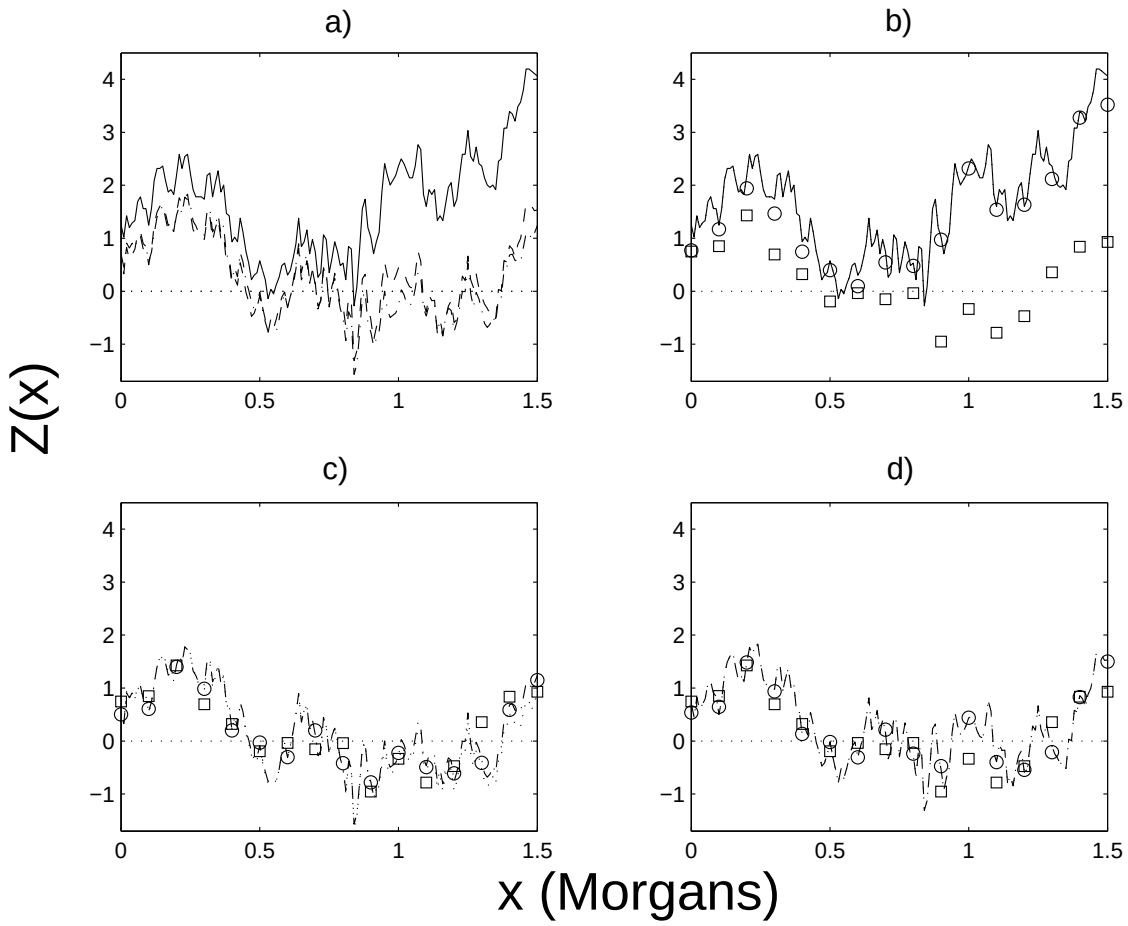

Figure 1: Plot of NPL score $Z(x)$ along one chromosome of length 150 cM under H_0 for 1000 affected sib pairs when all four pedigree members are genotyped. Three different score functions are used; S_{IBD} (b), S_{tr} (c) and $\tilde{S}_{\text{IBD}}$ (d); and four marker maps; ideal dense (dotted), highly informative (circles), less informative (squares) and dense (solid for S_{IBD}, dash-dotted for S_{tr} and dashed for $\tilde{S}_{\text{IBD}}$). The highly informative map has markers at positions $0, 1, \ldots, 150$ cM, each one with 10 equally frequent alleles. The less informative map has markers at positions $0, 10, \ldots, 150$ cM, each one with 5 equally frequent alleles. Panel a) shows the dense marker NPL scores for all three score functions. Marker data for all combinations of score functions and maps are based on the same allele IBD-configuration processes $\bar{w}(\cdot)$ for all families. $Z(\cdot)$ is computed at grid points $0, 1, \ldots, 150$ cM for all maps. Only NPL scores at positions $0, 10, \ldots, 150$ cM are shown for the two non-dense maps.

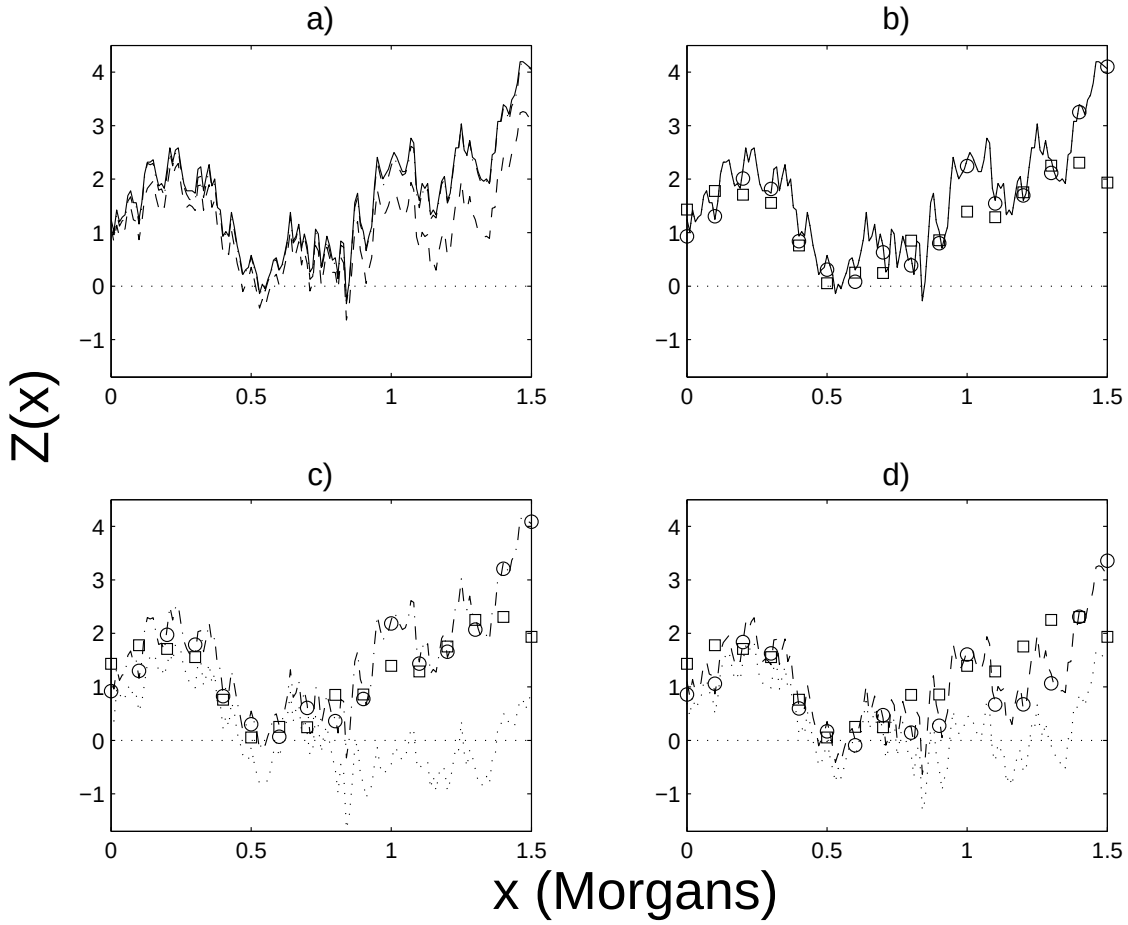

Figure 2: Plot of NPL score $Z(x)$ along one chromosome of length 150 cM under H_0 for 1000 affected sib pairs when the two sibs are genotyped. Marker data for all combinations of score functions and maps are based on the same allele IBD-configuration processes $\bar{w}(\cdot)$ as in Figure 1 for all families. For details on score functions and marker maps, see Figure 1.

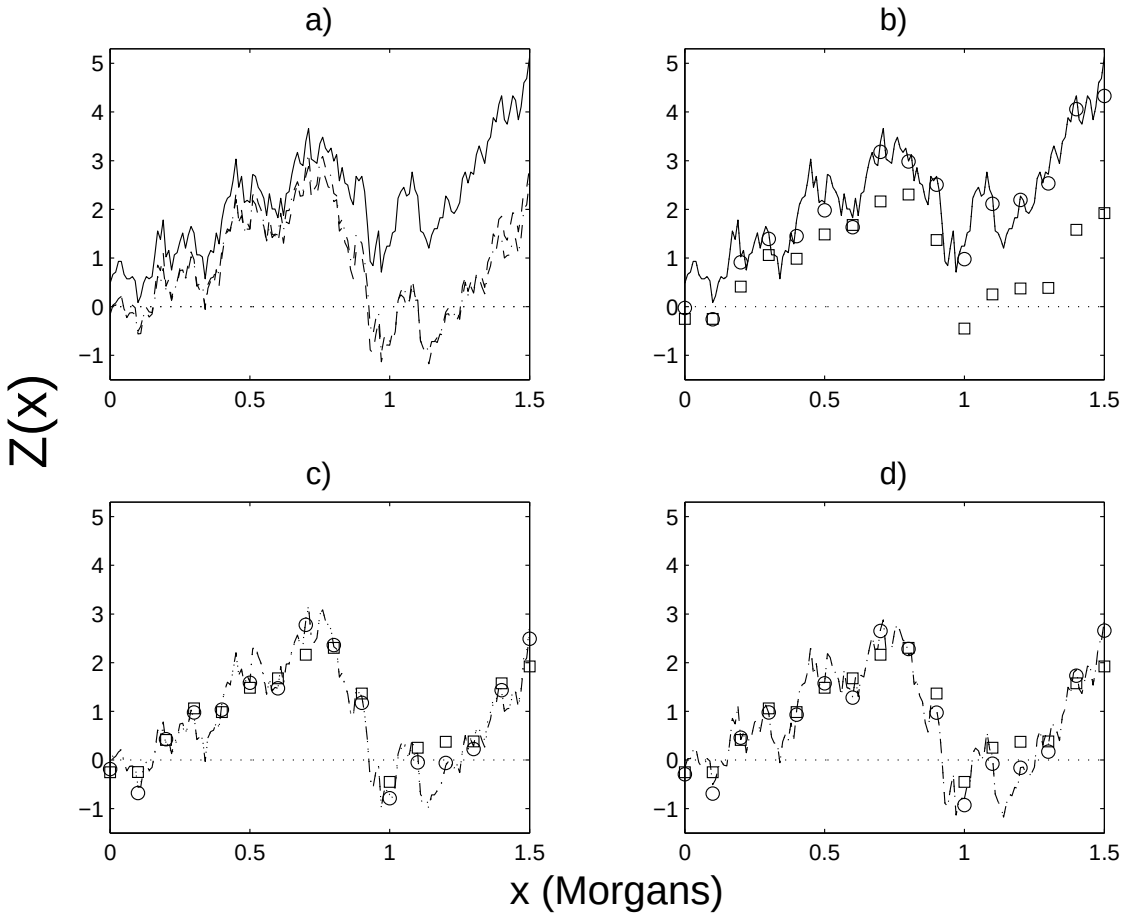

Figure 3: Plot of NPL score $Z(x)$ along one chromosome of length 150 cM under H_1 for 1000 affected sib pairs when all four family members are genotyped. The disease locus is positioned at 75 cM. It is biallelic, with disease allele frequency 0.1 and penetrance parameters $\psi_0 = \psi_1 = 0.1$ and $\psi_2 = 0.8$. For details on score functions and marker maps, see Figure 1.

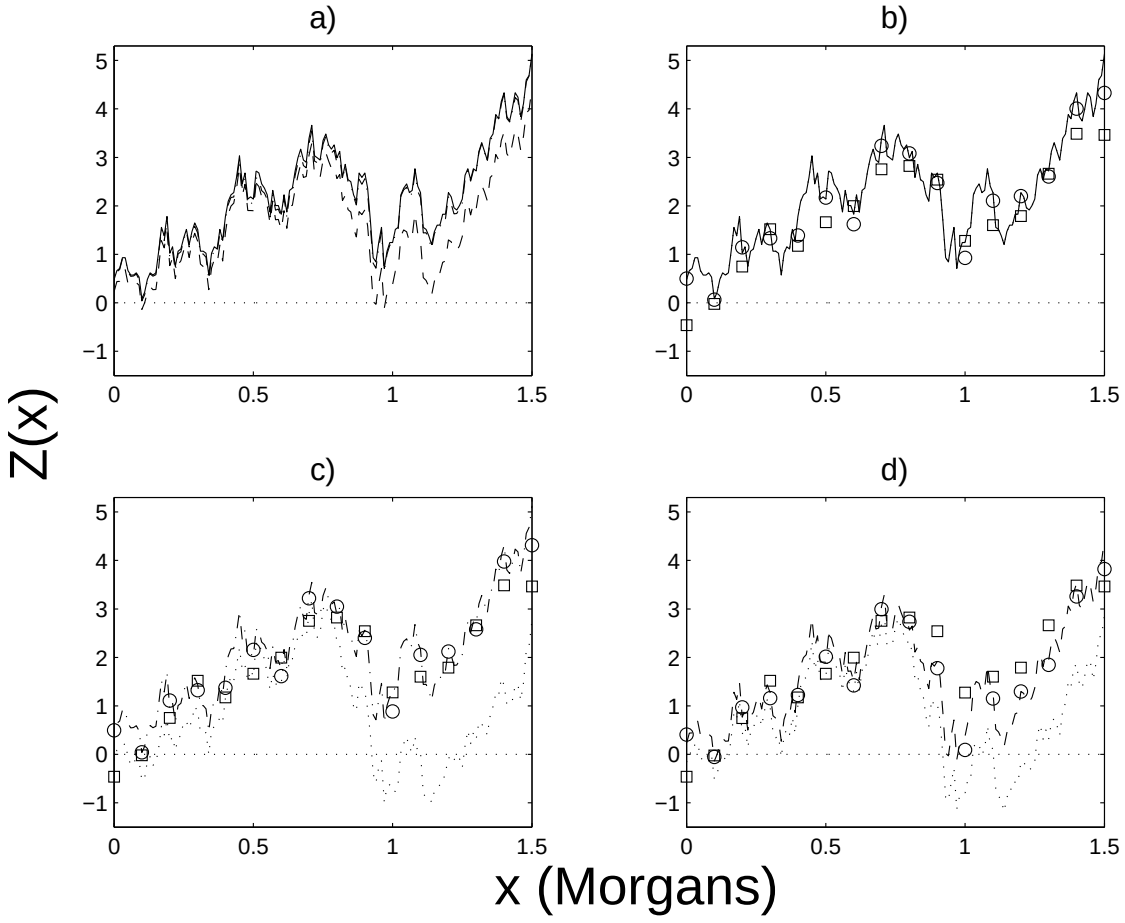

Figure 4: Plot of NPL score $Z(x)$ along one chromosome of length 150 cM under H_1 for 1000 affected sib pairs when the two sibs are genotyped. The disease locus position, genetic model and allele IBD-configuration processes $\bar{w}(\cdot)$ for all families are the same as in Figure 3. For details on score functions and marker maps, see Figure 1.

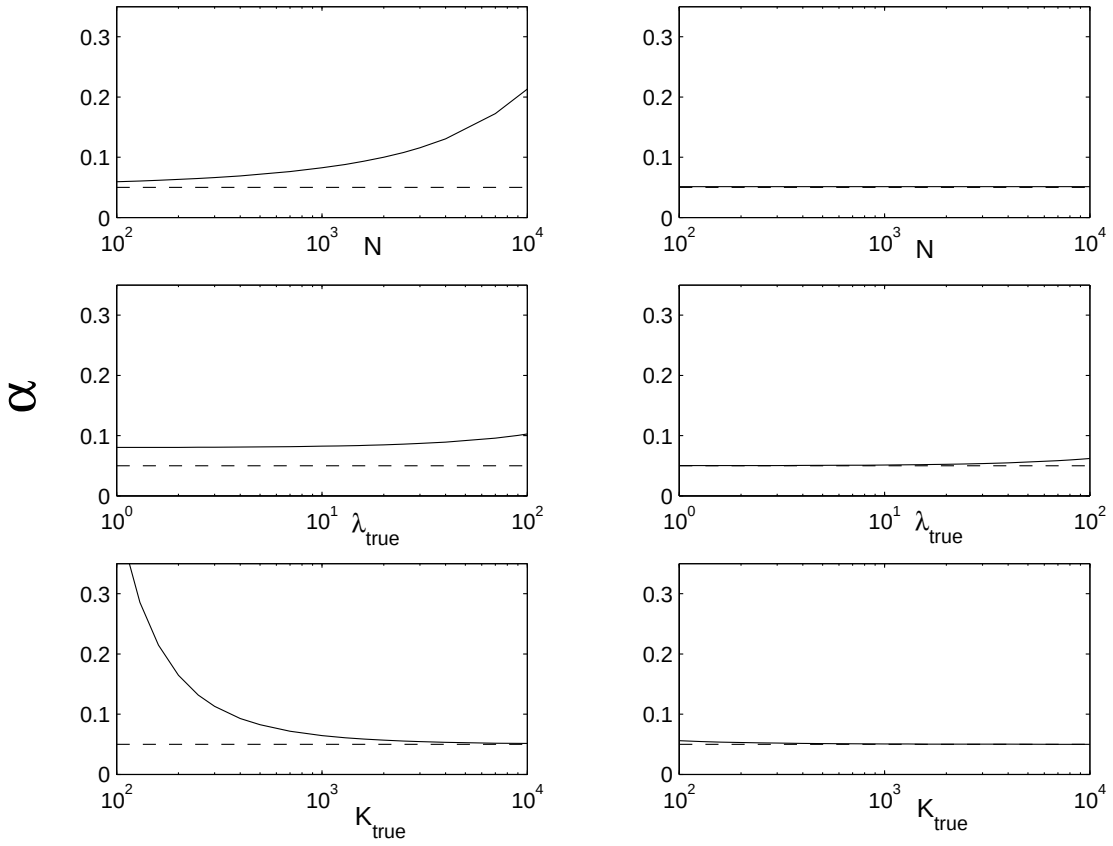

Figure 5: Plot of nominal (dashed) and actual (solid) significance levels for a genomewide scan of N affected sib pair families as function of K_{true}, λ_{true} and N. The marker map is ideal dense and the score functions are S_{IBD} (left panels) and $\tilde{S}_{\text{IBD}}$ (right panels). No inbreeding is assumed ($K = \infty$). The parameter values not varied in each panel are fixed to $K_{\text{true}} = 500$, $\lambda_{\text{true}} = 10$ and $N = 1000$ respectively.

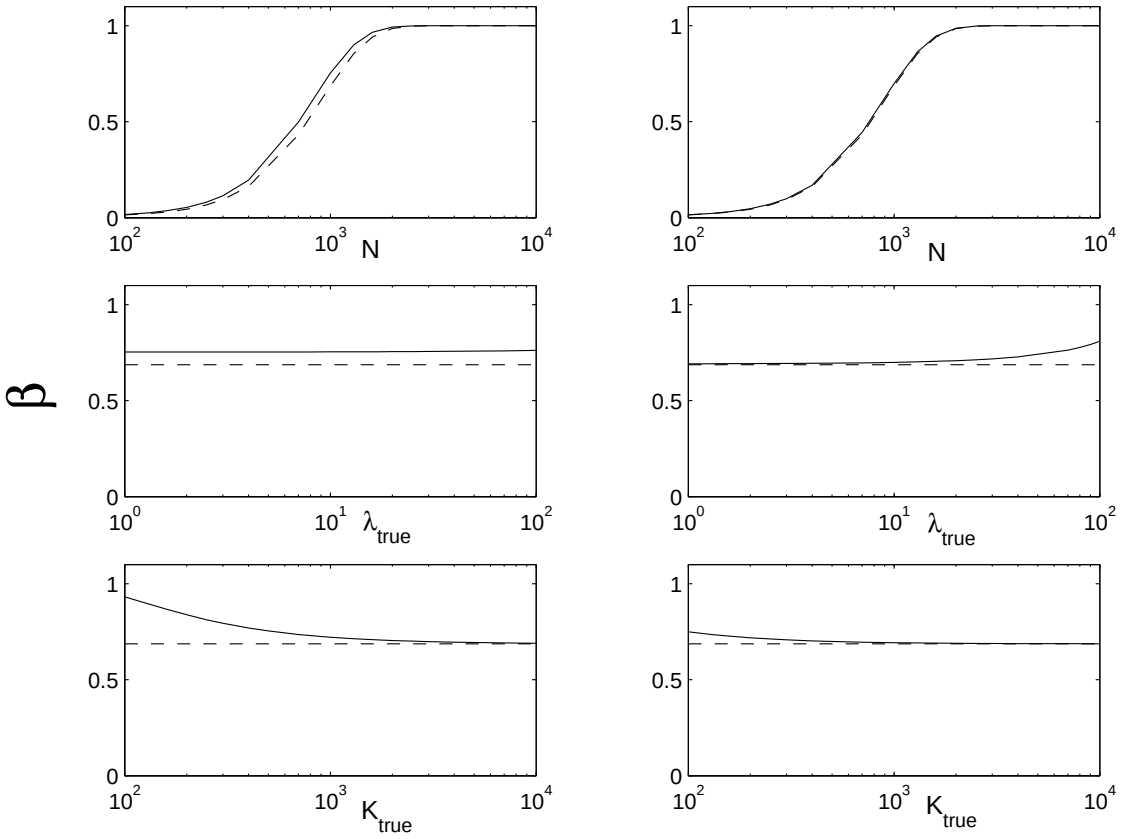

Figure 6: Plot of nominal (dashed) and actual (solid) power for a genomewide scan of N affected sib pair families as function of K_{true}, λ_{true} and N. The disease gene is biallelic, with disease allele frequency 0.1 and penetrance parameters $\psi_0 = \psi_1 = 0.1$ and $\psi_2 = 0.8$. For further details, see Figure 5.

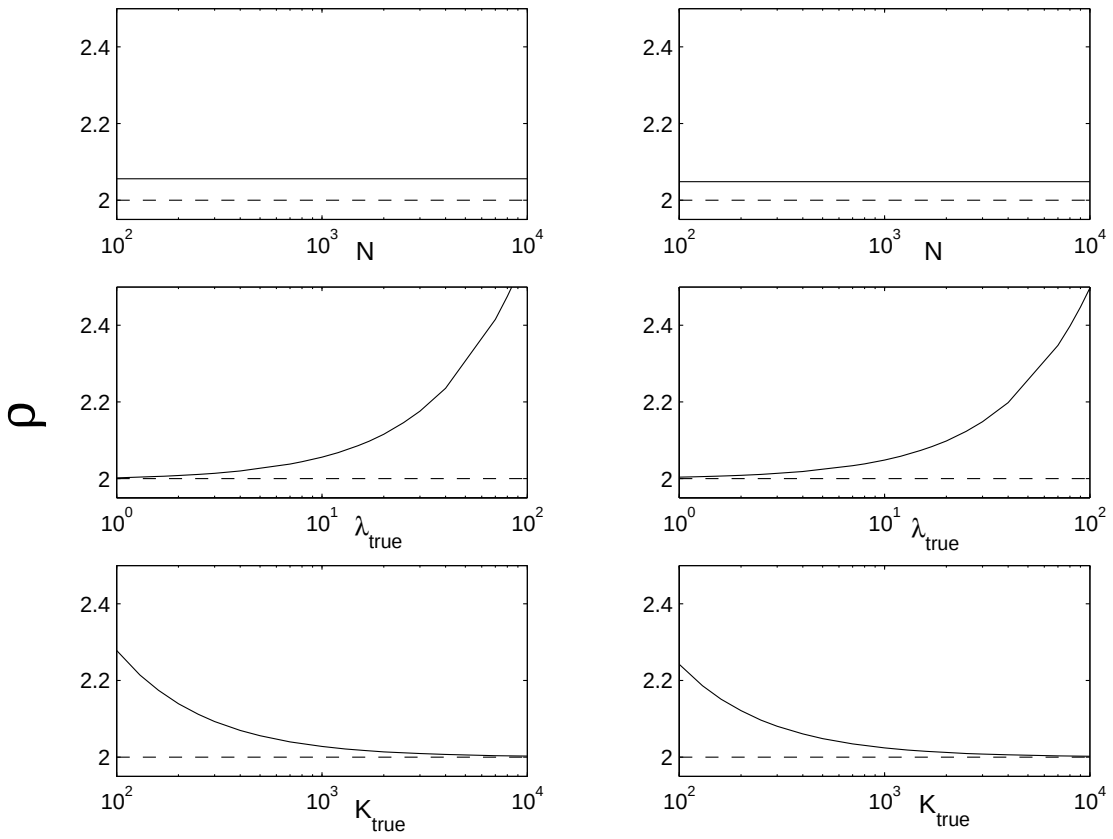

Figure 7: Plot of nominal (dashed) and actual (solid) crossover rate for a genomewide scan of N affected sib pair families as function of $K_{\rm true}$, $\lambda_{\rm true}$ and N. For further details, see Figure 5.