Original Paper

Human Heredity

Hum Hered 2008;66:138–156
DOI: 10.1159/000126049

Received: June 14, 2007
Accepted after revision: September 6, 2007
Published online: April 14, 2008

Strategies for Conditional Two-Locus Nonparametric Linkage Analysis

Lars Ängquist[a] Ola Hössjer[b] Leif Groop[c]

[a]Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University, Lund,
[b]Department of Mathematics, Division of Mathematical Statistics, Stockholm University, Stockholm, and
[c]Clinical Sciences, Diabetes and Endocrinology, Lund University, Malmö, Sweden

Key Words
Nonparametric linkage analysis · Two-locus linkage analysis · Conditional linkage analysis · Score functions · Conditioning loci · Two-step procedure · Noncentrality parameter · Genome-wide significance and power calculations · ROC curves · Monte Carlo simulation

Abstract
In this article we deal with two-locus nonparametric linkage (NPL) analysis, mainly in the context of conditional analysis. This means that one incorporates single-locus analysis information through conditioning when performing a two-locus analysis. Here we describe different strategies for using this approach. Cox et al. [Nat Genet 1999;21:213–215] implemented this as follows: (i) Calculate the one-locus NPL process over the included genome region(s). (ii) Weight the individual pedigree NPL scores using a weighting function depending on the NPL scores for the corresponding pedigrees at specific conditioning loci. We generalize this by conditioning with respect to the inheritance vector rather than the NPL score and by separating between the case of known (predefined) and unknown (estimated) conditioning loci. In the latter case we choose conditioning locus, or loci, according to predefined criteria. The most general approach results in a random number of selected loci, depending on the results from the previous one-locus analysis. Major topics in this article include discussions on optimal score functions with respect to the noncentrality parameter (NCP), and how to calculate adequate p values and perform power calculations. We also discuss issues related to multiple tests which arise from the two-step procedure with several conditioning loci as well as from the genome-wide tests.

1 Introduction

Nonparametric linkage analysis is commonly used in studies with binary traits for which the genetic model is unknown or hard to estimate, such as for complex diseases. This can be done in several ways. One possibility is to use MLS scores, where the likelihood of data is maximized jointly over disease locus and allele sharing probabilities [2, 3]. Alternatively, for MOD scores, the LOD score may be maximized jointly over disease locus and genetic model parameters, as suggested by [4–6]. A third approach is to choose an allele sharing statistic, here referred to as *score function*, which quantifies compatibility between the inheritance pattern of a pedigree with its phenotypes [7–15].

Two-locus linkage analysis is motivated by the possibility to take advantage of *gene-gene interaction* [16–18]. This is a challenging task, since the collection of genetic models increases dramatically, see e.g. [19] for a complete list of distinct two-locus fully penetrant models for bi-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/08/0663–0138$24.50/0

Accessible online at:
www.karger.com/hhe

Lars Ängquist
Mathematical Statistics, Centre for Mathematical Sciences
Box 118
SE–221 00 Lund (Sweden)
Tel. +46 46 222 4443, Fax +46 46 222 4623, E-Mail Lars.Angquist@matstat.lu.se

nary (biallelic disease loci) traits. In addition to this comes the increased amount of *multiple testing* inherent in searching for two rather than one disease locus. In spite of this, two-locus linkage analysis can often be worthwile; see the review articles [20–22].

When conducting nonparametric two-locus linkage analysis, one possibility is to perform a search for the two disease loci simultaneously or unconditionally. This can be done for the MLS approach [23–25], the score function approach [26, 27], by means of regression analysis [28] or by specifying the allele-sharing probabilities for affected relative pairs in advance [29]. Computation of the relevant NPL scores is facilitated by joint specification of the inheritance vector given marker data at two (unlinked) loci, as implemented in the GENEHUNTER-TWOLOCUS program [27]. Simultaneous search for two loci may also be achieved for quantitative traits using variance components methods, either exactly [30] or by means of Markov Chain Monte Carlo approximation [31].

An alternative strategy, which has greatest potential if one of the two disease loci has a strong marginal effect, is to proceed sequentially and first detect one disease locus y by one-locus linkage analysis, and then perform a conditional two-locus NPL search, where y is kept fixed and a second locus x is varied. This strategy was first proposed by [16] for heterogeneous traits, and has later been investigated in [29] for allele sharing statistics, in [1] for general multiplicative nonparametric score functions and in [32, 33] for linkage analysis based on generalized estimating equations. It is also possible to proceed sequentially for MLS scores [34] as well as for variance components models [30, 31]. Applications of the conditional method can be found e.g. in [35–37].

The work [1] aimed at detecting epistasis or heterogeneity when two disease loci are nonsyntenic, i.e. located on different chromosomes. In this paper we generalize their conditional NPL approach, based on score functions, in various ways. Firstly, we condition on the inheritance vector rather than the one-locus NPL score, so that more general score functions are allowed for. Secondly, we consider conditional search both when the conditional locus is fixed in advance or determined by an initial one-locus scan (the sequential method referred to above). To distinguish these two cases, the relevant null hypotheses H_0 of no linkage and alternative hypotheses H_1 of linkage are defined rigorously. In the sequential case, H_0 is the same as for one-locus and unconditional two-locus NPL analysis; enabling power comparisons. Thirdly, we derive optimal score functions that maximize the noncentrality parameter. This has previously been done for one-locus analysis [38–40]. We generalize these findings and obtain optimal two-locus score functions for conditional as well as for unconditional analyses. This provides a very general way of comparing one-locus, conditional and unconditional two-locus analyses analytically for general pedigree structures. See also [41, 42] for related approaches. Since noncentrality parameters are related to, but still distinct from, power (multliple testing is ignored), we also perform power comparisons between different score functions for various genetic models based on Monte Carlo simulations.

The paper is organized as follows: In *Section 2* we present some basic theory and introduce basic notation, including a description of two-locus genetic disease models. *Section 3* is devoted to nonparametric linkage analysis in the form of one-locus analysis, general (unconditional) two-locus analysis and various versions of conditional two-locus analysis. Subsequently, in *Section 4* the attention is brought to noncentrality parameters and its implication in the form of NCP-optimal score functions. In *Section 5* actual NCP-and power calculations are performed, whereas a concluding discussion is given in *Section 6*. Finally, technical details are referred to the four appendices *A–D*.

2 Two-Locus Genetic Disease Models

Each NPL investigation is done in relation to a predefined autosomal genomic region Ω.[1] If C is the total number of chromosomes we define $\Omega = \cup_{i=1}^{C} c_i$, where c_i is the i-th chromosome which is of genetic map length $|c_i|$. The total length of Ω is therefore $|\Omega| = \Sigma_{i=1}^{C} |c_i|$. Given a specific locus x the corresponding chromosome where it is located is denoted $c(x)$. We assume bi-allelic disease loci, i.e. only the disease and normal allelic variants, D and d respectively, are possible.

A two-locus genetic model consists of three parts: (i) The disease allele-frequencies $p_1 = P(D_1)$ and $p_2 = P(D_2)$, where p_i is the probability of the disease allelic variant with respect to the i-th disease locus ($i = 1$ or 2). (ii) The penetrance matrix,

$$f = \begin{bmatrix} f_{00} & f_{01} & f_{02} \\ f_{10} & f_{11} & f_{12} \\ f_{20} & f_{21} & f_{22} \end{bmatrix}, \tag{1}$$

where f_{ij} refers to the probability of being affected if the corresponding genotypes contain i and j copies of D_1 and D_2 respectively. (iii) The two disease loci, l_1 and l_2. Usually, they are assumed to be located on different chromosomes, i.e. $c(l_1) \neq c(l_2)$.

We will briefly describe four classes of two-locus genetic disease models. For $0 \leq i, j \leq 2$: (i) If the one-locus penetrances are

[1] This is subsequently referred to as our genome.

combined in an additive fashion, $f_{ij} = g_i + h_j$, we speak of an *additive* two-locus genetic disease model. (ii) If the two-locus penetrances instead are defined through products, $f_{ij} = g_i h_j$, we speak of a *multiplicative* two-locus genetic disease model. (iii) If we consider a dual version of the multiplicative model, i.e. where $(1 - f_{ij}) = (1 - g_i)(1 - h_j)$, we end up with $f_{ij} = g_i + h_j - g_i h_j$, which is a *heterogeneity* two-locus genetic disease model. (iv) If the two-locus penetrances depend only on the total number of disease alleles, $f_{ij} = k_{i+j}$, for some function k_l, we speak of a *threshold* two-locus genetic disease model.

3 Nonparametric Linkage Analysis

3.1 One-Locus Analysis

Assume a pedigree set consisting of N pedigrees. The inheritance pattern of the k-th pedigree at locus x is determined by means of the *inheritance vector*,

$$v_k(x) = [p_1(x), m_1(x), p_2(x), m_2(x), \ldots, p_{(n_k - f_k)}(x), m_{(n_k - f_k)}(x)], \quad (2)$$

see [43]. In (2) n_k is the number of individuals, f_k the number of founders and $(n_k - f_k)$ the number of nonfounders of pedigree k. Moreover, $p_i(x)$ and $m_i(x)$ equal 0 if the i-th nonfounder's paternal and maternal allele respectively, at locus x, originate from a grandfather and 1 if they originate from a grandmother. The number of vector positions equals the number of meioses $m_k = 2(n_k - f_k)$.

Define the *pedigree-specific NPL score* for the k-th pedigree at locus x as

$$Z_k(x) = E\left(S_k\left[v_k(x)\right]\right) = \sum_{w \in \mathbb{V}_k} P_{v_k(x)}(w) S_k(w), \quad (3)$$

where $E(X)$ denotes the expected value taken with respect to the stochastic variable X, $P_{v_k(x)}(w) = P(v_k(x) = w \mid \mathrm{MD})$ is the conditional probability of $v_k(x)$ given marker data MD, V_k is the full set of 2^{m_k} inheritance vectors and S_k is the one-locus score function for pedigree k.

We assume that the score function S is a priori standardized to have zero mean and unit variance under the null hypothesis of no linkage, i.e. given that w is uniformly distributed over $\mathbb{V}_k$, we have

$$E_{H_0}(S) = 2^{-m_k} \sum_{w \in \mathbb{V}_k} S_k(w) = 0 \text{ and } V_{H_0}(S) = 2^{-m_k} \sum_{w \in \mathbb{V}_k} S_k^2(w) = 1, \quad (4)$$

where E_{H_0} and V_{H_0} correspond to the expected value and variance of S_k under the null hypothesis H_0. This leads to $V_{H_0}[Z_k(x)] \leq 1$, i.e. an upper bound 1 for the variance of the family score [44].

The *NPL score* [44] for the pedigree set is then a weighted linear combination,

$$Z(x) = \sum_{k=1}^{N} \gamma_k Z_k(x), \quad (5)$$

of the pedigree-specific NPL scores with weights γ_k satisfying

$$\sum_{k=1}^{N} \gamma_k^2 = 1, \quad (6)$$

so that $V_{H_0}[Z(x)] \leq 1$. The weights may be chosen to depend on pedigree structure, size and phenotypes. A different maximum likelihood-based NPL score was introduced by [45]. It coincides with (5) for perfect marker data.

Using the maximum of the *NPL score process* along Ω,

$$Z_{\max} = \sup_{x \in \Omega} Z(x) \quad (7)$$

as a test statistic makes it possible to test for the presence of any disease locus. The natural test hypotheses are

$$\begin{cases} H_0: & \text{No disease locus on } \Omega, \\ H_1: & \text{At least one disease locus on } \Omega, \end{cases} \quad (8)$$

which leads to the genome-wide significance level and power

$$\begin{cases} \alpha(z) = P_{H_0}\left(Z_{\max} \geq z\right), \\ \beta(z) = P_{H_1}\left(Z_{\max} \geq z\right), \end{cases}$$

for a test that rejects H_0 when $Z_{\max} \geq z$.

3.2 Two-Locus Analysis

In the two-locus case we define an *unconditional* two-locus score function $S_k(w_1, w_2)$ for the k-th pedigree, which depends on inheritance vectors $w_1, w_2 \in \mathbb{V}_k$. In analogy with the one-locus case above, the scores are normalized using a two-locus generalization of (4) leading to

$$\sum_{w_1, w_2} S_k\left(w_1, w_2\right) = 0 \text{ and } 2^{-2m_k} \sum_{w_1, w_2} S_k^2\left(w_1, w_2\right) = 1. \quad (9)$$

Moreover, the pedigree-specific NPL score (3) is now generalized, being defined with respect to pairs of loci $(x,y) \in \Omega$, as

$$Z_k\left(x, y\right) = \sum_{w_1, w_2} P_{v_k(x,y)}\left(w_1, w_2\right) S_k\left(w_1, w_2\right), \quad (10)$$

where $P_{v_k(x,y)}(w_1, w_2) = P(v_k(x) = w_1, v_k(y) = w_2 \mid \mathrm{MD})$ is the joint inheritance distribution at loci x and y. We will restrict ourselves to unlinked loci, $c(x) \neq c(y)$, and then inheritance at x and y is independent, i.e. the product of the one-dimensional inheritance distributions appearing in (3) [27].[2] Note that the same restriction is used in the following sections as well. Moreover, the generalization to linked loci can be found in [46].

Now, the next step is to combine pedigree-specific scores (10) into a total NPL score

$$Z\left(x, y\right) = \sum_{k=1}^{N} \gamma_k Z_k\left(x, y\right); \; c(x) \neq c\left(y\right), \quad (11)$$

where γ_k are pedigree weights satisfying (6), so that $V_{H_0}[Z(x)] \leq 1$. As in the one-locus case, they may depend on, for instance, pedigree structure and phenotypes.

Using the null hypothesis in (8) we may, in analogy with (7), present the two-locus maximum NPL score by maximizing $Z(x,y)$ over all loci x and y from different chromosomes, i.e.

$$Z_{\max,\mathrm{tl}} = \sup_{\substack{x, y \in \Omega \\ c(x) \neq c(y)}} Z\left(x, y\right), \quad (12)$$

which leads to genome-wide significance level and power,

[2] Formally, $P_{v_k(x,y)}(w_1, w_2) = P(v_k(x) = w_1 \mid \mathrm{MD}) P(v_k(y) = w_2 \mid \mathrm{MD}) = P_{v_k(x)}(w_1)\, P_{v_k(y)}(w_2)$.

$$\begin{cases}\alpha_{\mathrm{tl}}(z)=P_{H_0}\left(Z_{\max,\mathrm{tl}}\geq z\right),\\ \beta_{\mathrm{tl}}(z)=P_{H_1}\left(Z_{\max,\mathrm{tl}}\geq z\right),\end{cases}$$

where 'tl' is an abbreviation for 'two-locus'.

3.3 Conditional Two-Locus Analysis: Known Conditioning Locus

If we fix a *conditioning locus y* on chromosome $c(y)$ and use (10) and (11) with x varying through $\Omega^{c(y)} = \Omega\backslash c(y)$, we have a *conditional two-locus NPL analysis*. In order to later on define appropriate significance levels, we split our null hypothesis (8) into two parts as

$$\begin{cases}H_0^{c(y)}: \text{ No disease locus on Chromosome } c(y),\\ \bar{H}_0^{c(y)}: \text{ No disease locus outside Chromosome } c(y).\end{cases}\tag{13}$$

This setting leads to a more complicated and involved version of the normalization procedure in (9). Firstly, for the k-th pedigree, the centering is performed using the conditional constraints,

$$\sum_{w_1} S_k(w_1,w_2)=0 \quad (\forall w_2),\tag{14}$$

where w_2 is associated with the conditioning locus y. Defining

$$\overline{S}_k^2(w_2)=2^{-m_k}\sum_{w_1} S_k^2(w_1,w_2)$$

we get, instead of (6), the constraint

$$\sum_{k=1}^{N}\gamma_k^2\sum_{w_2} P_{v_k(y)}(w_2)\overline{S}_k^2(w_2)=1$$

for the pedigree weights γ_k. A reasonable approach is to set γ_k identical for pedigrees with equal structure and phenotypes. The conditional variance $\overline{S}_k^2(w_2)$ quantifies how variable S_k is at the first locus given inheritance vector w_2 at the second locus. Notice that $P_{v_k(y)}(w_2)$ allows for imperfect data at y. It can be shown that (15) implies

$$V_{\bar{H}_0^{c(y)}}\left[Z(x,y)\,|\,MD^{c(y)}\right]\leq 1,\tag{15}$$

where MD^c is marker data from chromosome c.

Assuming perfect marker data, the conditional procedure described in [1] may be recognized as a special case of the general approach. (We refer to it as the *Cox-approach*.) The major difference is that in [1] conditioning is on the one-locus score $S_k[v_k(\mathrm{y})]$ at y, whereas we condition on $v_k(y)$. Thus the multiplicative Cox-approach corresponds to a two-locus score function $S_k(w_1,w_2) = S_k(w_1)f[S_k(w_2)]$, for which (11) can be written

$$Z(x,y)=\sum_{k=1}^{N}\gamma_k Z_k(x)f\left[Z_k(y)\right],\tag{16}$$

with normalization

$$\sum_{k=1}^{N}\gamma_k^2 f\left[Z_k(y)\right]^2=1.$$

This is a multiplicative two-locus score which is analogous to a one-locus NPL score process along $\Omega^{c(y)}$, weighting pedigrees according to a combination of pedigree-specific weights γ_k and a function $f(Z)$ of one-locus NPL scores Z at locus y.[3] The function f in (16) may be chosen in different ways, depending on the assumed genetic model, see e.g. [1, 47, 48].

The maximum conditional two-locus NPL score is based on maximizing $Z(x,y)$, keeping the conditioning locus y fixed and varying x over all chromosomes but $c(y)$. This corresponds to

$$Z_{\max,y}=\sup_{x\in\Omega^{c(y)}} Z(x,y).\tag{17}$$

The conditional significance level and power, given marker data $\mathrm{MD}^{c(y)}$ on chromosome $c(y)$, are

$$\begin{cases}\alpha_y(z)=P_{\bar{H}_0^{c(y)}}\left[Z_{\max,y}\geq z\,|\,MD^{c(y)}\right],\\ \beta_y(z)=P_{\bar{H}_1^{c(y)}}\left[Z_{\max,y}\geq z\,|\,MD^{c(y)}\right],\end{cases}\tag{18}$$

where $\bar{H}_1^{c(y)}$ is the alternative hypothesis,

$\bar{H}_1^{c(y)}$: At least one disease locus outside Chromosome $c(y)$,

corresponding to the lower part of (13).

In (13), we have replaced H_0 by the less restrictive null hypothesis $\bar{H}_0^{c(y)}$. In this setting y is allowed to be, or being linked to, a disease locus. The underlying assumption for this argument can be formalized as:

Assumption 1

$MD^{c(y)}$ is independent of phenotypes under $H_0^{c(y)}$ and $MD^{c(y_1)}$, $MD^{c(y_2)}$, ..., $MD^{c(y_k)}$ are conditionally independent given phenotypes if $k \geq 2$, all $c(y_j)$ are different, and at most one $H^{c(y_j)}$ is not valid.

If $H_0^{c(x)}$ holds, but not necessarily $H_0^{c(y)}$, it follows from Assumption 1 that the conditional distribution of $\mathrm{MD}^{c(x)}$ given phenotypes and $\mathrm{MD}^{c(y)}$ equals the unconditional distribution of $\mathrm{MD}^{c(x)}$. As a consequence, $\alpha_y(z)$ above can be calculated in the same way as a one-locus significance level $\alpha(z)$ along $\Omega^{c(y)}$.

3.4 Conditional Two-Locus Analysis: Unknown Conditioning Loci

When we do not have, or assume, explicit knowledge of an obvious conditioning locus, such as a known disease locus, we may randomly select interesting loci according to some predefined one-locus criterion. This is the motivation for the two-step procedure described below.

Since we possibly deal with multiple conditioning loci, we replace the less restrictive null hypothesis (13) by (8).

Selecting Conditioning Loci

Define the chromosome-wise NPL score maximum,

$$Z_{\max}^c=\sup_{x\in c} Z(x),$$

and the corresponding chromosomal significance level,

$$\alpha^c(z)=P_{H_0^c}\left(Z_{\max}^c\geq z\right),$$

[3] For imperfect data, (16) is not directly equivalent to using (11) with $S_k(w_1,w_2) = S_k(w_1)f[S_k(w_2)]$.

where H_0^c is the upper part of (13). A general one-locus NPL score-dependent selection criterion for conditioning chromosomes is

$$\mathbb{C} = \left\{c;\, Z^c_{\max} \geq z^c\right\}, \tag{19}$$

where z^c is a given threshold for chromosome c.

Further, denote the random positions of the chromosome-wise NPL score maxima as

$$y^c = \arg\max_{x \in c} Z(x). \tag{20}$$

Using (19) and (20) the conditioning loci are selected as the members of the set[4]

$$\mathbb{Y} = \left\{y^c;\, c \in \mathbb{C}\right\},$$

which guarantees that they are all located on different chromosomes.[5]

We will mostly assume *equal* thresholds $z^c = z$ in (19). The tuning constant z reflects the number of conditioning loci that the investigator is willing to use. The choice is a compromise between finding true interactions on one hand and avoiding severe multiple testing on the other hand. One may note that an alternative may be to use *genetic length-dependent* thresholds, e.g. $z^c = (\alpha^c)^{-1}(\delta)$. This may be motivated by giving each chromosome the same probability δ under H_0 of producing a conditioning locus.

Combining Conditional NPL Scores

The most straightforward generalization of (17) to several conditioning loci is to consider a test statistic which maximizes $Z(x,y)$ over all pairs of loci such that $y \in \mathbb{Y}$ is a conditioning locus and x varies freely over all chromosomes outside $c(y)$, i.e.

$$Z_{\max,\mathbb{Y}} = \max_{y \in \mathbb{Y}} \max_{x \in \Omega^{c(y)}} Z(x,y) = \max_{c \in \mathbb{C}} Z_{\max,y^c}. \tag{21}$$

However, we will use a more refined approach based on conditional two-locus p values,

$$p^c = \alpha_{y^c}(Z_{\max,y^c}), \tag{22}$$

with α_y and y^c as defined in (18) and (20). Instead of (21), we then use the minimum p value [49],

$$p_{\min} = \begin{cases} \min_{c \in \mathbb{C}} p^c \text{ if } & \mathbb{C} \neq \emptyset \\ 1 \quad \text{if } & \mathbb{C} = \emptyset \end{cases} \tag{23}$$

as test statistic and reject H_0 whenever $p_{\min}$ is smaller than or equal to a given threshold u. In words, p^c is the p value associated with a conditional maximal NPL score with conditioning locus from chromosome c and $p_{\min}$ is the minimum p value obtained from all conditioning loci in $\mathbb{C}$. As opposed to (21), (23) takes into account varying chromosome lengths and the actual inheritance vectors at $y \in \mathbb{Y}$.[6]

Using (23) we get a genome-wide, or global, significance level and power as

$$\begin{cases} \alpha_{\mathbb{Y}}(u) &= P_{H_0}\left(p_{\min} \leq u\right) \\ \beta_{\mathbb{Y}}(u) &= P_{H_1}\left(p_{\min} \leq u\right) \end{cases}. \tag{24}$$

Under H_0, the probability of including a conditioning locus from chromosome c in $\mathbb{Y}$ is $\lambda^c = \alpha^c(z^c)$. Since each p^c has a uniform distribution on (0,1), ignoring discreteness effects of the null distribution of $Z_{\max,c}$, a simple Bonferroni upper bound for the significance level is

$$\alpha_{\mathbb{Y}}(u) \leq \left(\sum_{c=1}^{C} \lambda^c\right) u.$$

Hence, $\Sigma_c \lambda^c$ can be viewed as a crude measure of the *effective number* of conditioning loci used. In particular, if $z^c = (\alpha^c)^{-1}(\delta)$, each chromosome has the same probability δ of being included as conditioning locus, and the effective number of conditioning loci is $C * \delta$.

The null hypothesis in (24) is H_0, i.e. no disease locus at all. This does not imply that single disease loci are easily detected by this test. On the contrary, $p_{\min}$ is designed to take advantage of interaction between disease loci from different chromosomes and has high power primarily if there are at least two disease loci, of which at least one has strong marginal effect.

4 Noncentrality Parameters

A quantity of great importance is the *noncentrality parameter (NCP)* which measures the expected NPL score, at the disease locus or loci, under an alternative hypothesis. For one-, two- and conditional two-locus NPL scores we define,

$$\begin{aligned} &\mathrm{NCP}_{l_1} = E_{H_1}[Z(l_1)], \\ &\mathrm{NCP}_{l_1,l_2} = E_{H_1}[Z(l_1,l_2)], \\ &\mathrm{NCP}_{l_1|l_2} = E_{H_1}[Z(l_1,l_2) \mid \mathrm{MD}^{c(l_2)}]. \end{aligned} \tag{25}$$

where l_1 and l_2 are the two disease loci. Notice that the first two quantities in (25) are constants, whereas the third is a random variable since we condition on marker data from chromosome $c(l_2)$.

We define a *homogeneous* pedigree set as a genotypes-phenotypes data set constituted only of information on pedigrees with equivalent pedigree structure and phenotypes. For a homogeneous pedigree set consisting of N pedigrees, perfect data, and equal pedigree weights ($\gamma_k = 1/\sqrt{N}$),

$$\begin{aligned} &\mathrm{NCP}_{l_1} = A\sqrt{N}, \\ &\mathrm{NCP}_{l_1,l_2} = B\sqrt{N}, \\ &\frac{1}{\sqrt{N}}\mathrm{NCP}_{l_1|l_2} \xrightarrow{\mathbb{P}} D \;\; as \;\; N \to \infty, \end{aligned} \tag{26}$$

where A and B are the NCPs for a single pedigree and $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. A natural interpretation of D is an average NCP per family for conditional two-locus analysis.

To derive expressions for the NCPs in (26) we define the joint inheritance vector distribution $P(w_1,w_2) = P(v(l_1) = w_1, v(l_2) = w_2 | Y, H_1)$ at the two disease loci, with corresponding marginal distributions, $P_1(w_1)$ and $P_2(w_2)$, and common phenotype vector Y.

[4] If there is prior evidence that a disease locus exists somewhere along chromosome c one may use $\mathbb{Y} = \arg\max_{x \in c} Z(x)$, and H_0 may in this case be replaced by the weaker $\overline{H}_0^c$.

[5] This restriction may be relaxed, requiring only a certain *minimum map distance L* between all pairs of syntenic conditioning loci.

[6] Given a pedigree set, for each single case the collection of inheritance vectors at the corresponding conditioning locus gives rise to, possibly and probably, different NPL score distributions and hence significance distributions.

(For a corresponding computational algorithm, see Appendix A of [48].) The following theorem presents maximal NCPs and corresponding optimal score functions. A proof is given in Appendix A.

Theorem 1 (NCP-Optimal Score Functions)
For a homogeneous pedigree set, the maximum NCPs are

$$A^2 = 2^m \sum_{w_1} P_1^2(w_1) - 1,$$
$$B^2 = 2^{2m} \sum_{w_1, w_2} P^2(w_1, w_2) - 1, \quad (27)$$
$$D^2 = 2^m \sum_{w_1, w_2} \frac{P^2(w_1, w_2)}{P_2(w_2)} - 1.$$

The maxima in (27) are attained for the NCP-optimal score functions,

$$S(w_1) \propto P_1(w_1) - 2^{-m},$$
$$S(w_1,w_2) \propto P(w_1,w_2) - 2^{-2m}, \quad (28)$$
$$S(w_1,w_2) \propto P(w_1|w_2) - 2^{-m}.$$

Evidently $A \leq B$ and $A \leq D$, but there is no simple order relation between B and D. Often $B \geq D$.

We emphasize that all three optimal noncentrality parameters A, B and D require knowledge of the genetic model. They should be interpreted as the best possible values of the NCP that the investigator might expect when the genetic model is correctly specified. In practice though, the genetic model is often unknown. Then the relevant NCPs may be computed, once a score function S has been chosen, for a range of different genetic models. By comparing these with the optimal NCPs, the loss of information is quantified. Here the quantification might come from either one or both of (i) choosing the wrong genetic model or (ii) using a non-optimal score function.

5 Results

Note that the score functions $S = S_{opt}$ in (28) which maximize NCPs do not have to maximize power, since NPL score distributions deviate from the standard normal and multiple testing is ignored in the NCP criterion. However, the score function that actually maximizes power should in most cases be close to S_{opt}. See for instance [53] on how the NCP is related to an analytical approximation of the power in the one-locus case. For this reason we discuss both NCPs and power in this section. A more direct way of comparing optimal power would be to generalize the optimal unconditional tests of [26] to (i) more general pedigrees than affected sib-pairs and (ii) conditional linkage analysis.

Moreover, throughout this section, we will use homogeneous pedigree sets taken from figure 1. To reduce the computational complexity and increase interpretability we also assume perfect data throughout all calculations.

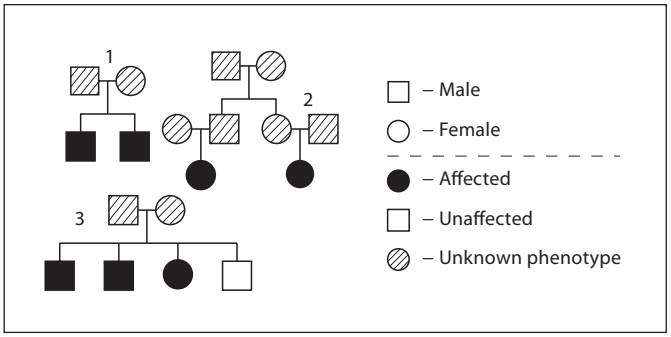

Fig. 1. The pedigrees used in NCP and power calculations.

Some comments on imperfect data are given at the end of Section 5.

5.1 NCP Calculations

We calculate maximal NCP parameters in (27) for several distinct genetic models under the constraint of a constant disease prevalence $K = 0.01$. For more details on our selection of genetic models, see Appendix B. In figures 2–5, and later on, all the results are displayed for these various types of disease models.

In all simulations we report NCP as function of the *displacement*,

$$d = \max_{i,j} f_{ij} - \min_{i,j} f_{ij} = f_{22} - f_{00},$$

which quantifies the strength of the genetic model. We have $d = d(x)$, where x is the penetrance parameter defined in Appendix B.

From figures 2–5, we notice that: (i) In all cases, $B \geq D \geq A$ in (27). (ii) For multiplicative models, as expected, $A = D$. (iii) For additive and heterogeneity models, D is consistently larger than A, though closer to A than to B. (iv) For threshold models, D tends to be closer to B than to A.

This implies that among our set of models the threshold-type (followed by the additive- and heterogeneity-type) seems to be most suitable for conditional two-locus analyses. One may also note that in this setting the performance of our additive and heterogeneity models is quite similar; see [50] for a motivation. Further, the NCPs seem to be heavily dependent on the prevalence K. We have performed similar analyses using $K = 0.1$ showing, for instance, that for additive and heterogeneity models $A \approx D$. (Results not shown; see [48].) For further discussion, see Section 6.

Exploring the influence of allele frequencies, setting $p = p_1 = p_2$ we have noticed that the NCPs are increasing

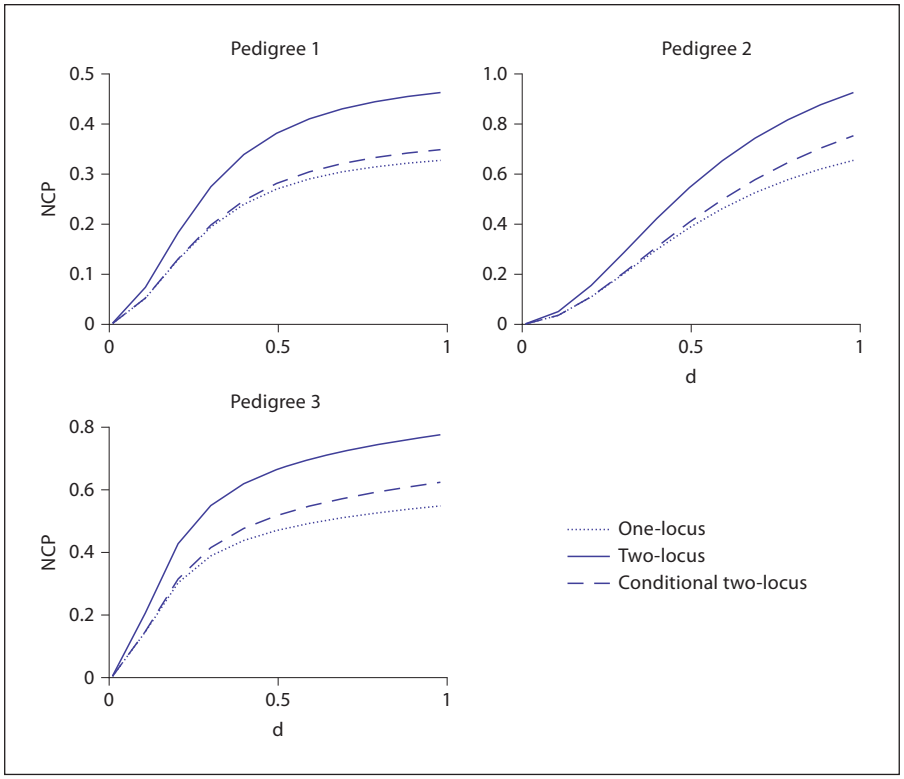

Fig. 2. Maximum one-, two- and conditional two-locus NCP calculations using pedigrees 1–3, equal disease allele frequency at both disease loci ($p = p_1 = p_2 = 0.01$), a constant disease prevalence $K = 0.01$ and symmetric *additive* two-locus disease penetrance models (f^1) and displacement $d = d(x)$ with $x = 0{:}0.024{:}0.24$. $N = 1$, i.e. NCP = A, B and D in (26) respectively.

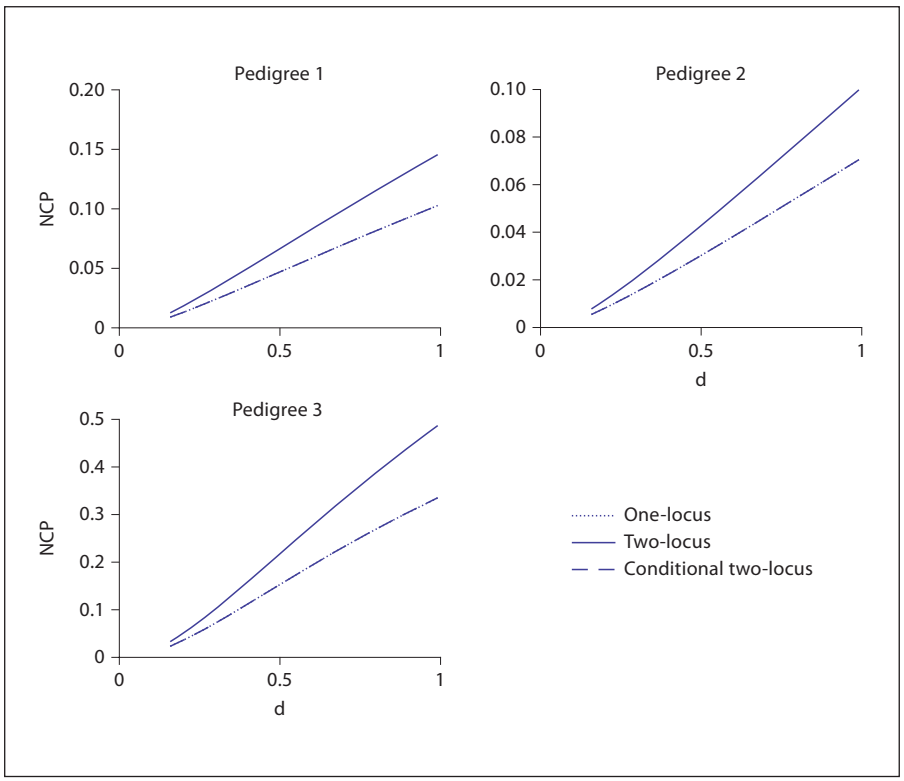

Fig. 3. Maximum one-, two- and conditional two-locus NCP calculations using symmetric *multiplicative* two-locus disease penetrance models (f^2) and displacement $d = d(x)$ with $x = 0.20{:}0.0295{:}0.495$. For details, see figure 2.

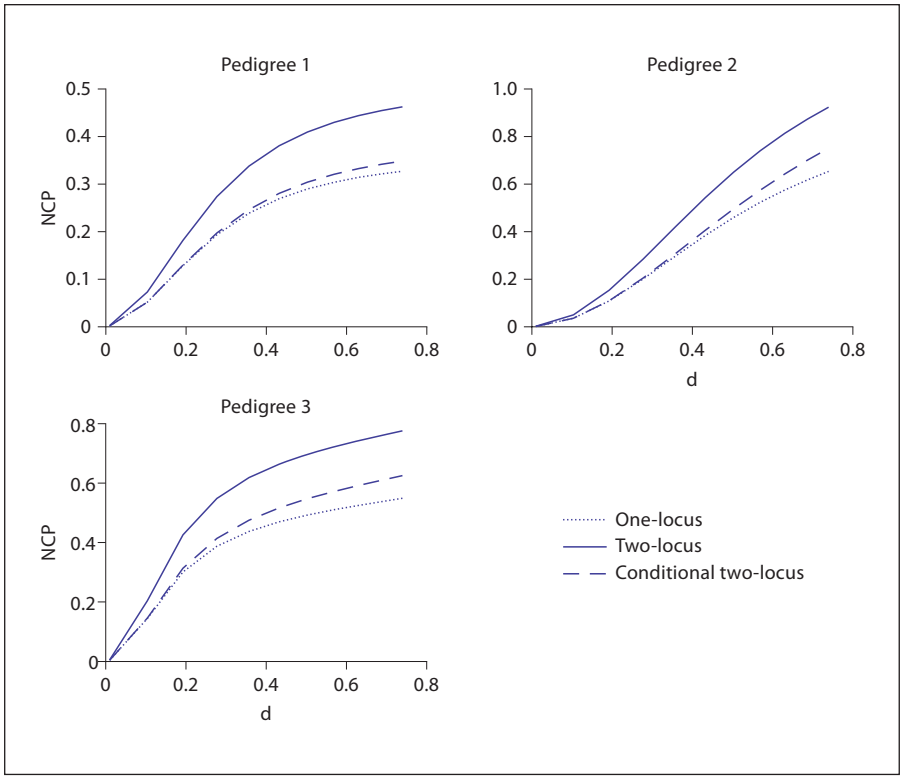

Fig. 4. Maximum one-, two- and conditional two-locus NCP calculations using symmetric *heterogeneity* two-locus disease penetrance models (f^3) and displacement $d = d(x)$ with $x = 0{:}0.024{:}0.24$. For details, see figure 2.

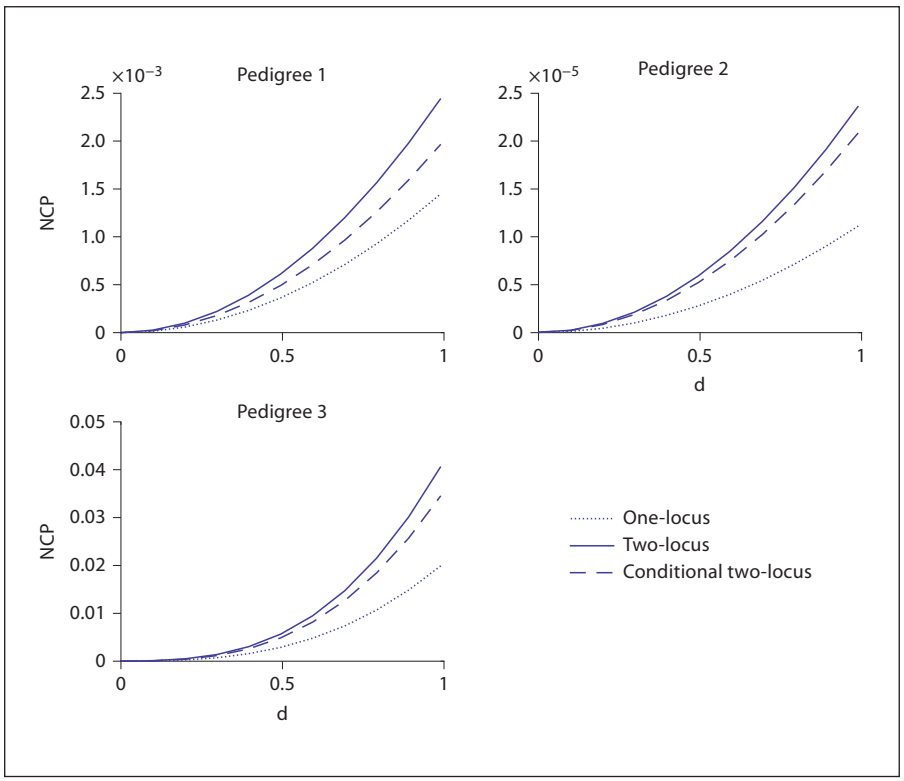

Fig. 5. Maximum one-, two- and conditional two-locus NCP calculations using symmetric *threshold* two-locus disease penetrance models (f^4) displacement $d = d(x)$ with $x = 0{:}0.0495{:}0.495$. For details, see figure 2.

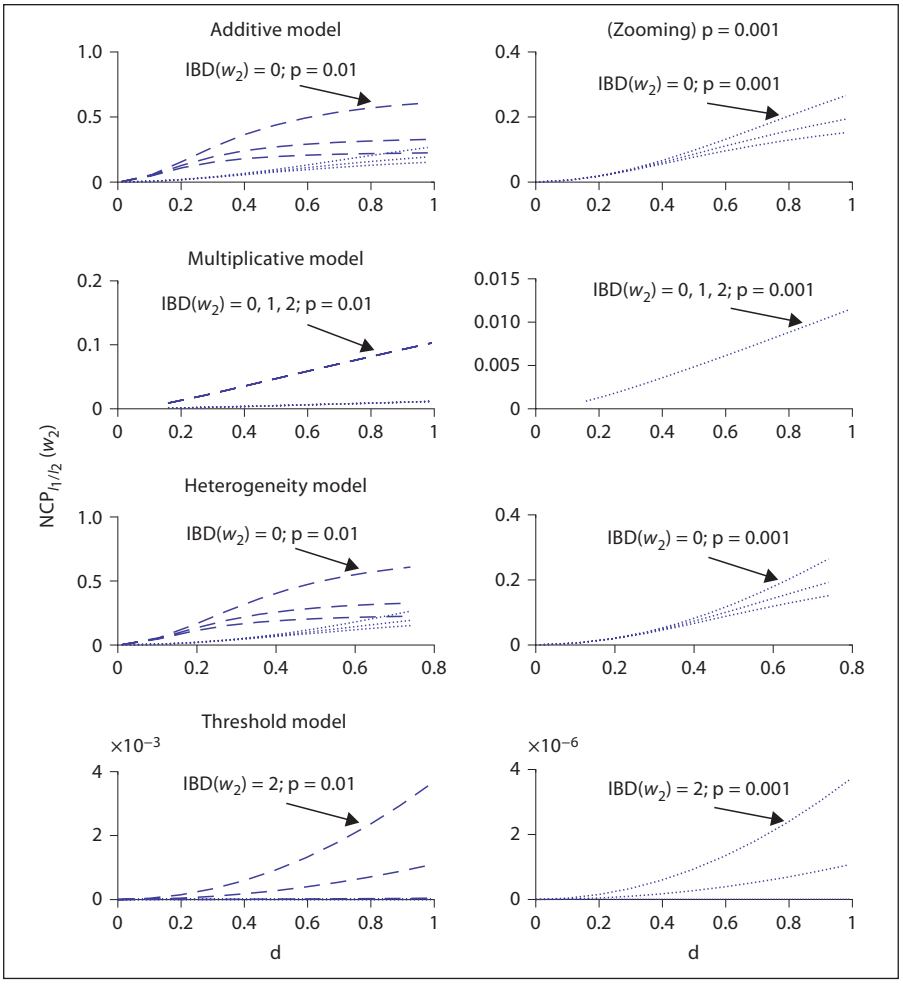

Fig. 6. Optimal conditional noncentrality parameters $\mathrm{NCP}_{l_1|l_2}(w_2)$ for pedigree 1. We assume a fixed conditioning locus l_2, $p = p_1 = p_2 \in \{0.001, 0.01\}$ (dotted and dashed lines), disease prevalence $K = 0.01$, four symmetric two-locus penetrance models (f^1–f^4) and displacements $d = d(x)$. Note that the right-hand side panels are zoomed versions of the left-hand side ones, in order to clearly display the results corresponding to $p = 0.001$.

functions of p for most choices of prevalence K and penetrance parameter x, when the latter two are kept fixed. (Results not graphically displayed; see [48].) Note that, since K and x are held constant and y in Appendix B must be greater than 0, p is in each case restricted to a certain interval.

The average conditional noncentrality parameter for a homogeneous pedigree set with perfect marker data and uniform weights $\gamma_k = \gamma$ can be written as

$$D = \frac{\sum_{w_2} \mathrm{NCP}_{l_1|l_2}(w_2)\bar{S}(w_2)P_2(w_2)}{\sqrt{\sum_{w_2} \bar{S}^2(w_2)P_2(w_2)}}, \qquad (29)$$

where $\bar{S}^2(w_2) = 2^{-m} \Sigma_{w_1} S^2(w_1,w_2)$ is the conditional variance and

$$\mathrm{NCP}_{l_1|l_2}(w_2) = \sum_{w_1} S(w_1,w_2)P(w_1 \mid w_2)/\bar{S}(w_2)$$

the conditional noncentrality parameter for one pedigree when $v(l_2) = w_2$. Notice that any multiplicative constant of S cancels in (29). This does not violate (15), since γ can be varied freely. For the optimal score function (28) (with proportionality constant removed), we get an optimal conditional noncentrality parameter

$$\mathrm{NCP}_{l_1|l_2}(w_2) = 2^{m/2}\sqrt{\sum_{w_1}\left[P(w_1 \mid w_2) - 2^{-m}\right]^2} \qquad (30)$$

and

$$D^2 = \sum_{w_2} \mathrm{NCP}^2_{l_1|l_2}(w_2)P_2(w_2).$$

Hence the conditional noncentrality parameter quantifies how much $P(\cdot|w_2)$ deviates from a uniform distribution and D^2 averages the squared conditional noncentrality parameter with respect to (corresponding probability weights) P_2.

Next, recall pedigree 1 from figure 1. This common pedigree structure refers to an *affected sib-pair (ASP)*

pedigree. Figure 6 displays (30) for an ASP when IBD(w_2), the number of alleles shared identical-by-descent by the affected siblings, equals 0, 1 or 2. For additive and heterogeneity models, $\mathrm{NCP}_{l_1 \mid l_2}(w_2)$ decreases with IBD(w_2), whereas the opposite is true for threshold models. For multiplicative models,

$$P(w_1,w_2) = P_1(w_1)P_2(w_2)$$

when there are no unaffecteds in the pedigree. Hence $P(w_1|w_2) = P(w_1)$ and $\mathrm{NCP}_{l_1 \mid l_2}(w_2)$ is independent of w_2. Note that the conditional NCP, given fixed prevalence K, is increasing with allele frequency p (from $p = 0.001$ to $p = 0.01$), which is consistent with the previous discussion.

5.2 Power and Significance

Methods of Calculation

The significance level and power can be calculated using either: (i) *Analytical approximations* based on Gaussian extreme value theory [51–54]. (ii) *Monte Carlo simulations* [55–58].

The advantage of (i) is fast computations and available explicit expressions. However, it is still only an approximative procedure which, even in the modified versions correcting for nonnormality, for instance [59, 60], may give biased results. Related approaches are, for example, described in [61, 62]. On the other hand, (ii) is more adjustable to complicated situations and do not give biased results in the limit of large Monte Carlo samples. Its drawback is rather the computational burden. Modified simulation algorithms have been suggested in order to deal with this problem, such as importance sampling [63, 64] and the fast but slightly biased replicate-pool method [65, 66].

For conditional two-locus analysis with known conditioning locus, another possibility is to use *permutation testing* when marker data from the pedigrees are fully, or close to, exchangeable. One may note that the procedure outlined in [1], using our general two-locus score function framework, may be generalized to permuting inheritance vectors rather than one-locus NPL scores at the conditioning loci.[7]

With unknown conditioning loci, there is an additional level of uncertainty regarding the actual set of conditioning loci and their inheritance vectors. It seems difficult to adjust analytical approximations, permutation testing and fast simulation procedures to this in proper and convenient ways. Hence we use direct Monte Carlo simulation based on J replicates in all simulations. For instance, estimates $\hat{\alpha}_{\mathbb{Y}}(u)$ and $\hat{\beta}_{\mathbb{Y}}(u)$ of the significance level and power in Section 3.4 are

$$\frac{1}{J}\sum_{j=1}^{J} I\left(p_{\min}^{j} \leq u\right),$$

where $p_{\min}^{j}$ is the minimum p value for the j-th replicate. For all J replicates we simulate marker data along all chromosomes conditional on phenotypes under H_0 and H_1 respectively.

We present calculations using so called *receiver operating characteristic (ROC)* curves [67, 68] by plotting power against significance level for various thresholds.

Monte Carlo Simulations

The power calculations below are performed using score functions S_{pairs} [69] in the one-locus case, and $S_{\mathrm{pairs}}^{2\mathrm{loc}}$(see Appendix C) and $S_{\mathrm{pairs}}^{\mathrm{Cox}}$ with epistatic weights $f(Z) = I(Z \geq 0)$, see (16), in the conditional two-locus case. In addition we also included the NCP-optimal score function S_{opt} of (28), in the one-and conditional two-locus simulations.

We consider homogeneous pedigree sets with equal pedigree weights

$$\gamma_k = 1/\sqrt{N}$$

and four chromosomes of equal length 1.5 Morgans (M) with disease loci l_1 and l_2 located in the middle of the first two chromosomes.

Further, we use symmetric additive (f^1), multiplicative (f^2), heterogeneity (f^3) and threshold (f^4) models, setting the prevalence to $K = 0.01$, the maximum penetrance to $f_{22} = 0.99$ and the disease-allele frequencies $p = p_1 = p_2$ so that f_{00} attains its minimum value 0.

Our simulated one-locus results give that the performance of S_{opt} and S_{pairs} is close to identical which may, to a large extent, follow since these examples involve small pedigrees only, where S_{opt} and S_{pairs} usually should be quite similar. (Results not shown; see [48].) Generally, for small pedigrees S_{pairs} is often close to optimal. For larger pedigrees, S_{opt} often outperforms S_{pairs} to an extent depending on the genetic model.

Next, we compare conditional two-locus power calculations based on a *single known* conditioning locus $y = l_2$ in (17) with using a *single estimated* counterpart $y = \hat{l}_2 = \arg\max_{x \in c(l_2)} Z(x)$. The results are displayed in figures 7–10.

[7] Explicitly, the original set of inheritance vectors $v_k(y)$ at the conditioning loci y is being replaced by the permuted counterpart $v_{\pi_k}(y)$, where $\pi = (\pi_1,\pi_2, ..., \pi_N)$ is a permutation of (1,2, ..., N).

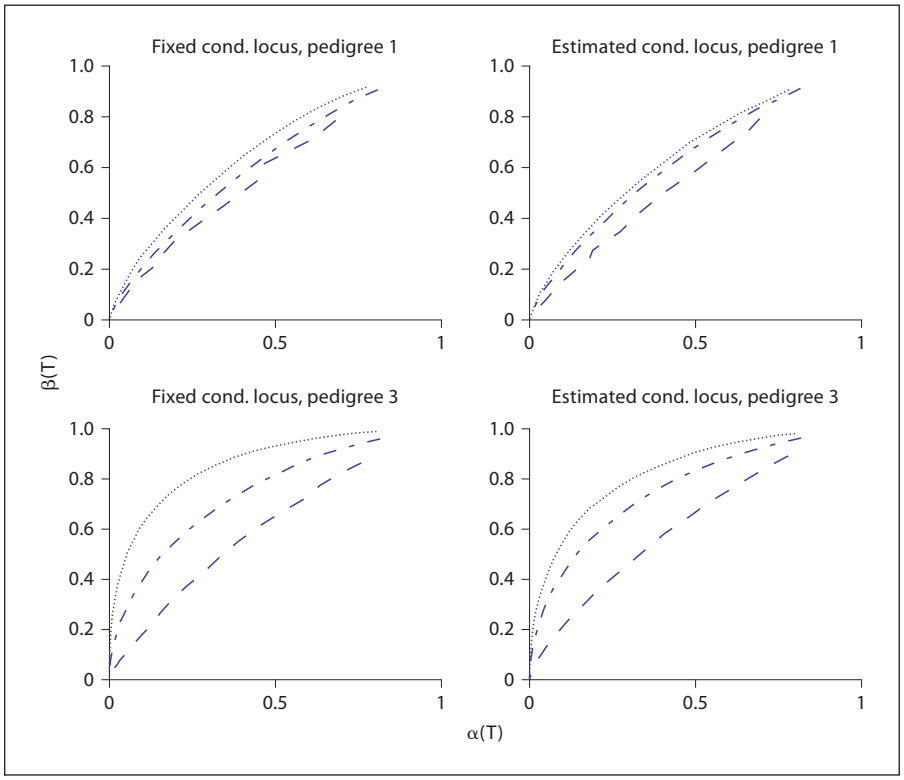

Fig. 7. Conditional two-locus ROC-curves, using pedigrees 1 and 3, under the additive disease model (f^1; $d = 0.99$, $x = 0.2425$) with $K = 0.01$, $p = p_1 = p_2 = 0.0075$ and three score functions ($S^{\text{Cox}}_{\text{pairs}}$, $S^{\text{2loc}}_{\text{pairs}}$ and S_{opt}; dashed, dashed-dotted and dotted lines respectively). The number of pedigrees is $N = 25$, the thresholds $T = 2.0, 2.1, ..., 6.0$ and the number of simulations $J = 10000$.

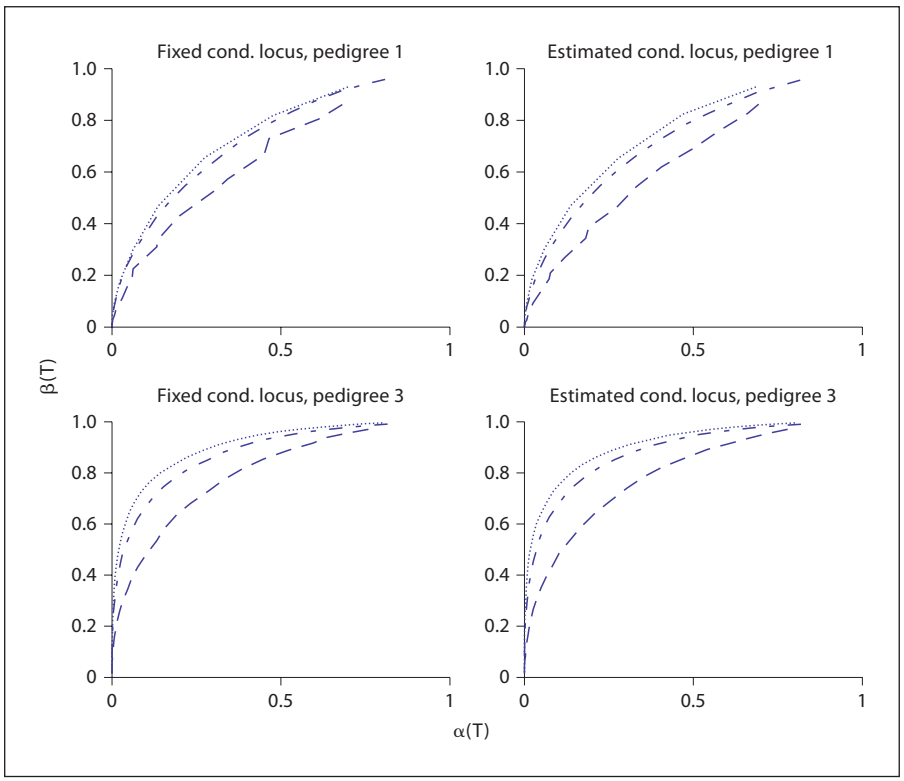

Fig. 8. Conditional two-locus ROC curves, using the multiplicative disease model (f^2; $d = 0.99$, $x = 0.4925$) with $p = p_1 = p_2 = 0.08$. For more details, see figure 7.

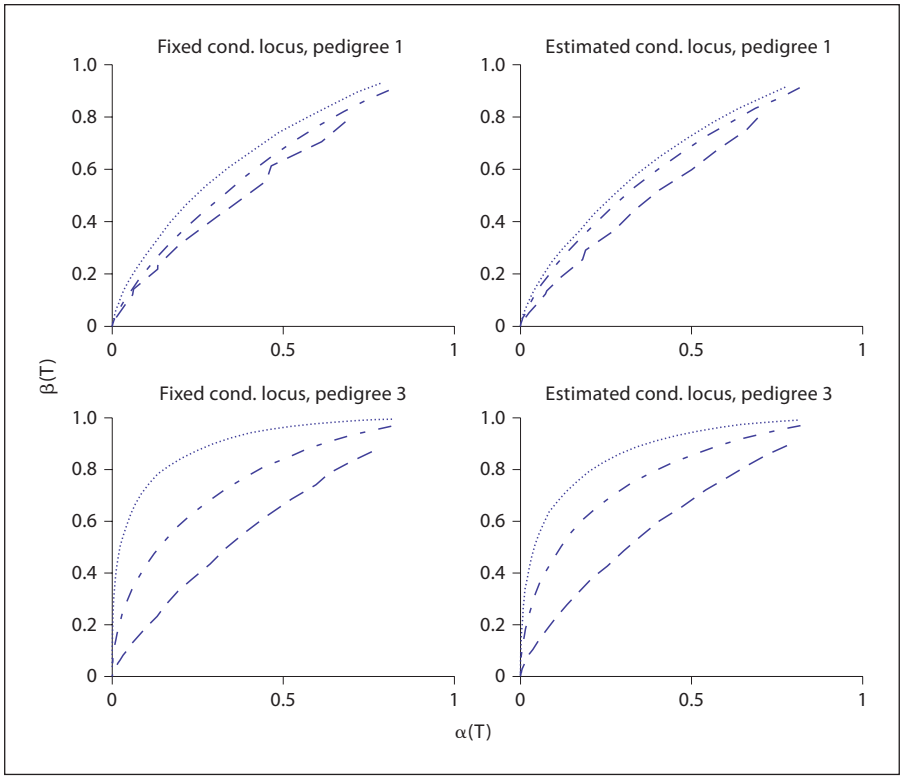

Fig. 9. Conditional two-locus ROC curves, using the heterogeneity disease model (f^3; $d = 0.99$, $x = 0.4450$) with $p = p_1 = p_2 = 0.005$. For more details, see figure 7.

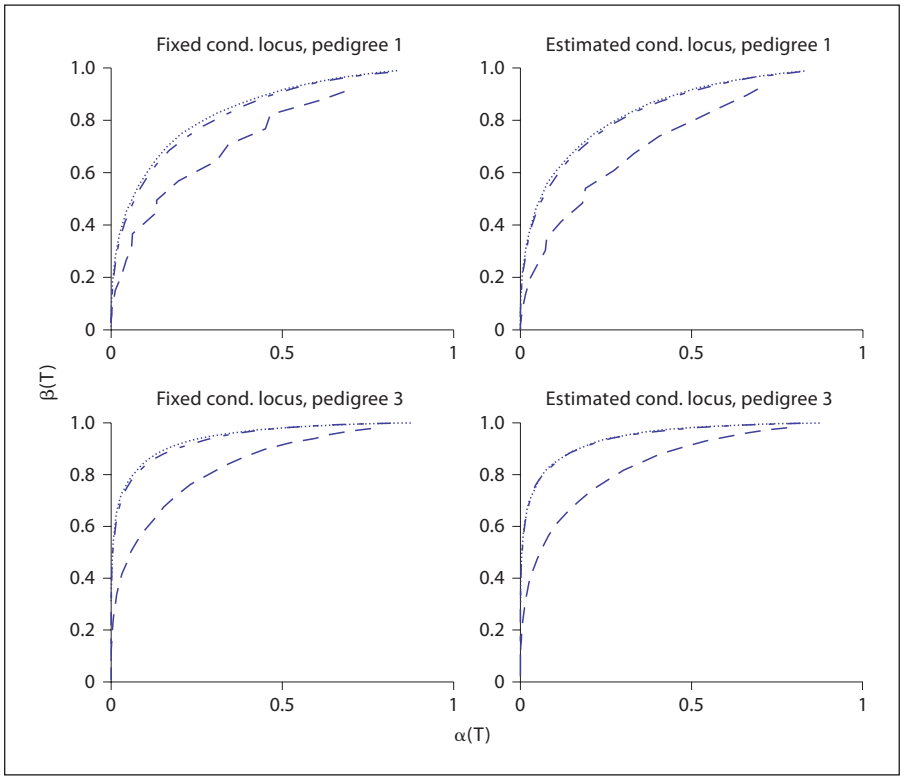

Fig. 10. Conditional two-locus ROC curves, using the threshold disease model (f^4; $d = 0.99$, $x = 0.4900$) with $p = p_1 = p_2 = 0.15$. For more details, see figure 7.

Table 1. Proportions of estimated disease loci that are located on $c(l_1)$, given knowledge that the other disease locus is located on $c(l_2)$ at either a *fixed* or *estimated* position

Gen. mod.	Sc. func.	Pedigree 1		Pedigree 3	
		fixed	estimated	fixed	estimated
f^1 (addi.)	$S_{\text{pairs}}^{\text{Cox}}$	0.5226	0.5789	0.5287	0.5798
	$S_{\text{pairs}}^{\text{2loc}}$	0.5725	0.5879	0.7553	0.7579
	S_{opt}	0.6453	0.6134	0.8702	0.8422
f^2 (mult.)	$S_{\text{pairs}}^{\text{Cox}}$	0.6568	0.6684	0.8073	0.8181
	$S_{\text{pairs}}^{\text{2loc}}$	0.7140	0.7282	0.8967	0.8938
	S_{opt}	0.7325	0.7480	0.9176	0.9185
f^3 (hete.)	$S_{\text{pairs}}^{\text{Cox}}$	0.5258	0.5586	0.5552	0.5918
	$S_{\text{pairs}}^{\text{2loc}}$	0.5689	0.6038	0.7843	0.7886
	S_{opt}	0.6572	0.6243	0.9162	0.8886
f^4 (thre.)	$S_{\text{pairs}}^{\text{Cox}}$	0.7456	0.7609	0.8501	0.8636
	$S_{\text{pairs}}^{\text{2loc}}$	0.8421	0.8611	0.9507	0.9545
	S_{opt}	0.8596	0.8609	0.9573	0.9542

For more details, see figure 7.

Let $\hat{l}_1 = \arg\max_{x \in \Omega \setminus c(l_2)} Z(x,y)$ be the estimated disease locus using the conditional NPL score. In table 1 the proportions

$$r = \hat{P}\,[c(\hat{l}_1) = c(l_1)]$$

of chromosome-wise correctly estimated disease loci are displayed, including only those simulations for which $Z_{\max,y} \geq 3$. Note that this informally corresponds to estimating the first disease locus given the second (known or estimated) one, conditioning on some strength of evidence of linkage.

Further, we consider a *random number* of conditioning loci. We select these loci by setting $z^c = 2.5$ in (19) for all $C = 4$ chromosomes in Ω and all three choices of score functions, see figures 11 and 12, where also the estimated probabilities of being selected as a conditioning locus is shown, i.e.

$$\hat{P}\left(c_i \in \mathbb{C}\right) = \frac{1}{J}\sum_{j=1}^{J} I\left(Z^j_{\max,c_i} \geq 2.5\right); \quad i = 1,2,3,4.$$

Since this case is simulation-wise more complex than the previous cases, we make some comments on the simulation procedure in Appendix D.

The interpretation of power with one mandatory conditioning locus versus a random number of loci is somewhat different. The latter case refers to the power to detect any disease locus (H_1), whereas in the former case power is restricted to detection of disease loci outside the conditioning chromosome c ($\overline{H}_1^c$). Hence, the corresponding ROC-curves are not directly comparable. Moreover, the methods of Sections 3.2 and 3.4 include more multiple testing than those of Sections 3.1 and 3.3. For instance, this implies that a two-locus ROC-curve might be dominated by a one-locus ROC-curve, even though $\beta_{\text{tl}}(z) \geq \beta(z)$ for each threshold z.

6 Discussion

We have outlined and discussed procedures of conditional two-locus NPL analysis. Our primary focus has been approaches that facilitate the calculations of significance levels, power and noncentrality parameters.

Conditional Two-Locus Power Calculations

Given the situation, one may note that in all investigated cases S_{opt} turned out to be the most powerful score function, followed by $S_{\text{pairs}}^{\text{2loc}}$ and $S_{\text{pairs}}^{\text{Cox}}$. A general observation is that, in many cases, the performances of the first two are quite similar, whereas the second one is far more powerful than the third procedure. This behaviour seems to be quite consistent with respect to disease models.

The main implication of this is that $S_{\text{pairs}}^{\text{2loc}}$ clearly outperforms the well-known $S_{\text{pairs}}^{\text{Cox}}$. In addition, the performance of $S_{\text{pairs}}^{\text{2loc}}$ may be improved on, and optimized, by adapting k (see Appendix C) to the genetic model. This is a topic deserving further study.

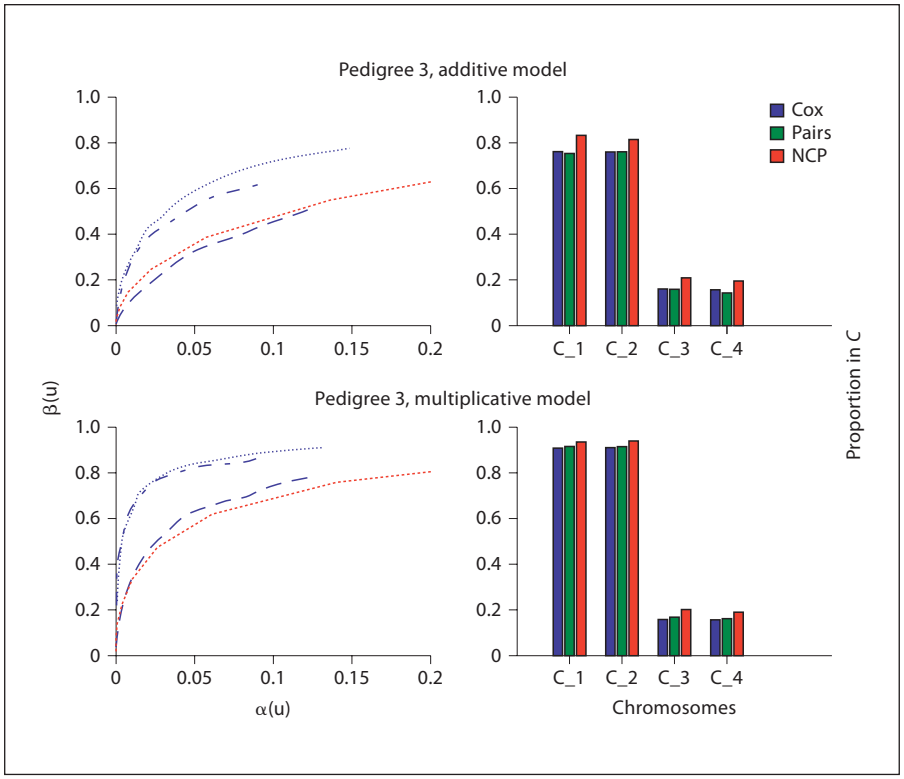

Fig. 11. [Left] Conditional two-locus ROC curves, with a random number of conditioning loci, for pedigree 3 and the additive (f^1) and multiplicative (f^2) disease models, using three two-locus score functions, thresholds $u = 0, 0.0025, ..., 0.25$ and $J = 2500$ simulations. The one-locus ROC curves based on S_{opt} is also displayed (lower dotted line). For further details, see figure 7. [Right] Estimated probabilities for each chromosome of being selected as conditioning locus. Note that conditioning loci are selected through one-locus scores (S_{pairs} for analysis based on S^{Cox}_{pairs} or S^{2loc}_{pairs}; one-locus S_{opt} for analysis based on two-locus S_{opt}).

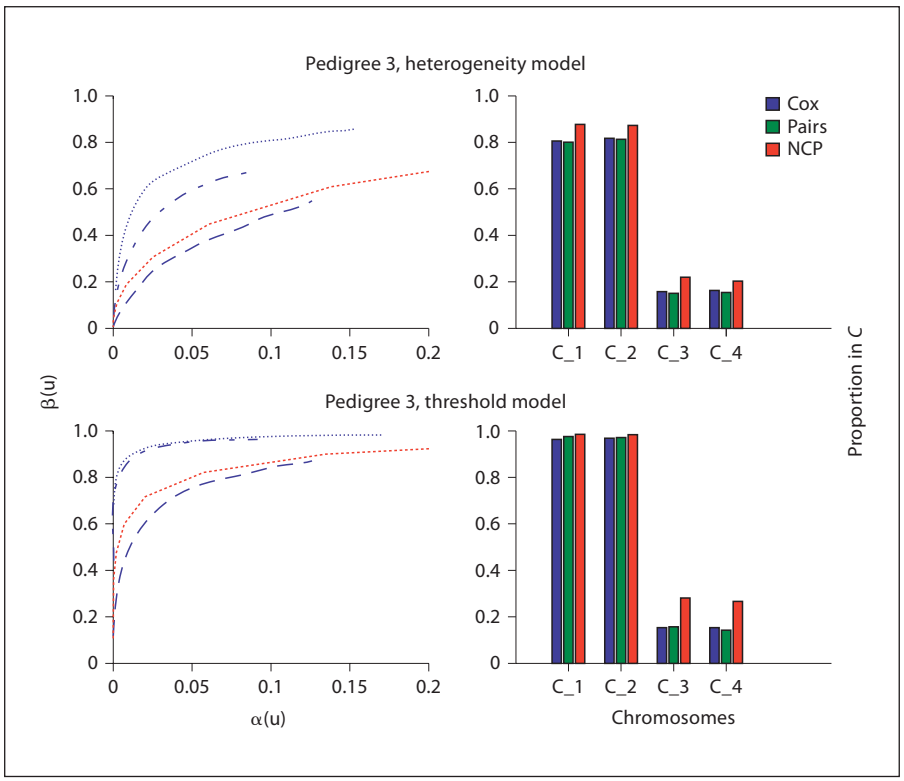

Fig. 12. [Left] Conditional two-locus ROC curves with a random number of conditioning loci, using the heterogeneity (f_3) and threshold (f_4) disease models. [Right] Estimated probabilities for each chromosome of being selected as conditioning locus. For further details, see figures 7 and 11.

Choice of Score Function

The score function S_{opt} requires a known genetic disease model. However, for most real cases, the underlying disease model is, at least to some extent, unknown. Though this is a severe limitation, information from previous *segregation analyses* [70, 71] may be helpful in the sense of suggesting, or at least narrowing down the set of plausible, disease models. Alternatively, genetic model parameters may be estimated simultaneously with the disease locus by means of a mod score analysis [34]. If the model uncertainty is low, S_{opt} might be used directly, whereas a higher degree of model uncertainty calls for adjusted approaches. Either a single robust score function, which performs well under a wide range of genetic models, may be used, or several distinct score functions. For more details on choosing score functions in NPL analysis, see [7, 11, 13, 15, 72].

Comparisons between Methods

Using the four families of disease models derived as in Appendix B, we calculated one-locus and conditional two-locus NCPs and powers. This was done (mainly) on prevalence level $K = 0.01$.

Among our genetic disease models, the threshold model-class is the only one where conditional two-locus NCPs are closer to unconditional two-locus NCPs than to one-locus NCPs. Though, both the additive and heterogeneity classes show considerably higher NCPs for conditional two-locus than for one-locus analyses (fig. 2–5). Further discussion on heterogeneity models and their use in conditional two-locus search is to be found in [29, 42]. Generally, these discussions seem to be consistent with our observations.

For the most informative pedigree structure, pedigree 3, one sees that conditional two-locus power is generally much higher than corresponding one-locus power (fig. 11, 12). The results seem to be consistent between model classes. With respect to our families of disease models, the multiplicative and threshold models S^{2loc}_{pairs} are closest in performance (power) to S_{opt}, and in these cases the relative possibility of real findings is largest. Generally, conditional two-locus S_{opt} and S^{2loc}_{pairs} perform much better than S^{Cox}_{pairs} (fig. 11, 12). For our models, S^{Cox}_{pairs} rarely outperforms the one-locus S_{opt}.

Note that NCP and power performance seems to be highly dependent on K. For instance, for our threshold models, if p is not very large, y in (32, 33) of Appendix B is close to zero. In this case, for large K, most affecteds carry few copies of D, leading to low NCPs and power.

In figures 7–10 we compare conditional two-locus power to the contrasting behaviour of using a conditioning locus of known or unknown location, where in the latter case the conditioning chromosome is known. The proportions r_{fix} and r_{est} of correctly estimated second disease loci are given in table 1. Generally $r_{fix} \approx r_{est}$, which seems surprising at first. The explanation is that, using an estimated conditioning locus, we condition on an inheritance vector corresponding to a high one-locus score, i.e. in many cases consistent with the disease model. Though the differences are small, r_{fix}/r_{est} seems largest and smallest for S_{opt} and S^{Cox}_{pairs} respectively.

Comments on Imperfect Data

In all simulations, we have assumed complete marker data, for which inheritance vectors (up to founder phase uncertainty) is known for all pedigrees at all loci. This is in order to (i) facilitate Monte Carlo simulations and avoid computationally complex algorithms for extracting inheritance vector distributions from data[8], and (ii) compute the maximal possible power that an investigator might expect before genotyping [76]. The complete marker data assumption is fairly realistic when all pedigree members are genotyped with an SNPs density at least, say, 0.1 cM. For larger multigenerational pedigrees, the founder generations are often not genotyped, and hence knowledge of the inheritance vector is less common. However, for marker data to be complete, it actually suffices to know $S_k[v_k(x)]$ in the one-locus case and $S_k[v_k(x),v_k(y)]$ in the (conditional) two-locus cases. Depending on the score function, this may not require all pedigree members to be genotyped. For instance, for pedigree 2, the S_{pairs} and S^{2loc}_{pairs} score functions only depend on IBD-sharing between the two affected first cousins and hence, in principle, it sufficies that these two are genotyped at a sufficiently dense set of markers. This follows since then sets of neighbouring markers might be combined into highly polymorphic ones facilitating the equality of IBS (identical-by-state) and IBD sharing and hence (close to) perfect marker information. More generally, in the same sense, if suffices to genotype the affected pedigree members when the score function only involves IBD-sharing between these.

As a final comment one may also note that all methods based on (7), (12), (17) and (23) are designed to be general enough to allow for imperfect data.

[8] For instance, by implementing the well-known Lander-Green-Kruglyak algorithm [73–75].

Other Approaches

A major current trend in gene localization is to employ genomewide association (GWA) studies. They may be preferable over genomewide linkage studies when the disease allelic variants are common and the effect of the disease gene(s) small [77]. This is related to the *common disease-common variant* hypothesis, which is a strong motivation for the HapMap project. Indeed, the GWA approach has shown recent promise for a number of complex dieseases [78]. Still, other approaches such as genomewide one- or two-locus linkage analysis seem to remain important [79]. Of course, a third possibility is to use combined linkage and association analysis; see e.g. [80] and references therein.

Acknowledgements

All calculations have been performed using MATLAB – The language of technical computing (© The MathWorks, Inc.). For instance, in August 2005 the version 7.1 (release 14) was available. For more information, for example, look up [81].

For editing and typesetting purposes we have used $\mathrm{L^AT_EX}$ – A Document preparation system. For more information, for example, dive into [82].

A Proof of Theorem 1

Let m be the common number of meioses of the homogeneous pedigree set. For a given score function, with weights

$$\gamma_k = 1/\sqrt{N}$$

and standardization (15), the NCP parameters in (26) are defined through

$$A = \sum_{w_1} P_1(w_1) S(w_1),$$
$$B = \sum_{w_1,w_2} P(w_1,w_2) S(w_1,w_2),$$
$$D = \sum_{w_2} \mu(w_2) P_2(w_2),$$

where $\mu(w_2) = \Sigma_{w_1} S(w_1,w_2)P(w_1|w_2)$.

Using the relevant methods in [12], A, B and D are maximized with respect to S under the constraints (4), (9) and (14, 15) respectively. Following an analogous version of Proposition 1 therein, the maximum NCPs turn out to be as in (27) and the corresponding NCP-optimal score functions as in (28). Maximization of D is done in two steps. First,

$$S(w_1,w_2) = C(w_2)[P(w_1|w_2) - 2^{-m}]$$

is derived by maximizing $\mu(w_2)$ for each w_2 subject to a constraint on $\overline{S}^2(w_2)$. Then $\Sigma_{w_2}\ \mu(w_2)P_2(w_2)$ is maximized with respect to $C(w_2)$ subject to (15), which for large samples can be written as

$$\sum_{w_2} \overline{S}^2(w_2) P_2(w_2) = 1.$$

The optimal choice $C(w) = c$ gives the desired solution.

B One-Parameter Families of Genetic Models

For simplicity, we assume that both disease allele frequencies are equal, i.e. $p = p_1 = p_2$, and that the penetrance matrix f in (1) is symmetric. The last two properties imply that the marginal one-locus genetic models are equal. The disease *prevalence* is then defined as

$$\begin{aligned} K = P(\text{affected}) = {} & f_{22}p^4 + 2(f_{21} + f_{12})p^3q + (f_{20} + f_{02})p^2q^2 \\ & + 4f_{11}p^2q^2 + 2(f_{10} + f_{01})pq^3 + f_{00}q^4, \end{aligned} \tag{31}$$

where $q = 1 - p$ is the normal allele frequency at both loci and $f_{ij} = f_{ji}$.

Taking advantage of (31) we calculate a *one-parameter family* of penetrance matrices for the four distinct two-locus disease model types defined in Section 2 as follows: Define penetrance matrices f^1–f^3 through setting $g = h = (y,K + x,K + 2x)$ for the *additive*, *multiplicative* and *heterogeneity* model-classes respectively, i.e. let

$$f^1 = \begin{bmatrix} 2y & K+y+x & K+y+2x \\ K+y+x & 2(K+x) & 2K+3x \\ K+y+2x & 2K+3x & 2(K+2x) \end{bmatrix},$$

$$f^2 = \begin{bmatrix} y^2 & y(K+x) & y(K+2x) \\ y(K+x) & (K+x)^2 & (K+x)(K+2x) \\ y(K+2x) & (K+x)(K+2x) & (K+2x)^2 \end{bmatrix}$$

and

$$f^3 = \begin{bmatrix} y(2-y) & y+(1-y)(K+x) & y+(1-y)(K+2x) \\ y+(1-y)(K+x) & 2(K+x)-(K+x)^2 & 2K+3x-(K+x)(K+2x) \\ y+(1-y)(K+2x) & 2K+3x-(K+x)(K+2x) & 2(K+2x)-(K+2x)^2 \end{bmatrix}.$$

Given K, p and x these classes are then well-defined through

$$y = \begin{cases} -\left[4p(K+x) - K\left(1+2p^2\right)\right]/2(1-p)^2 & \text{(additive)} \\ \left[Kp^2 - 2p(K+x) + \sqrt{K}\right]/(1-p)^2 & \text{(multiplicative)}, \\ \left[Kp^2 - 2p(K+x) + 1 - \sqrt{1-K}\right]/(1-p)^2 & \text{(heterogeneity)}. \end{cases}$$

Moreover, define

$$f^4 = \left[\left\{k_{i+j}\right\}\right] = \begin{bmatrix} K-2y & K-y & K \\ K-y & K & K+x \\ K & K+x & K+2x \end{bmatrix}. \tag{32}$$

directly for the *threshold* model class, which then becomes well-defined through

$$y = -[p^3(p-2)x]/[(p+1)(p-1)^3]. \quad (33)$$

For each of the one-parameter families above, in addition to our initial assumptions, the constraints $0 \leq f_{ij} \leq 1$ define a set of valid models $x_1 \leq x \leq x_2$. The larger x is within this interval, the stronger is the genetic component of the model.

C The Generalized Two-Locus Version of S_{pairs}

The unstandardized one-locus version of S_{pairs} is defined as

$$S_{\text{pairs}}(w) \propto \sum_{i<j} \text{IBD}_{i,j}(w),$$

where summation is over all pairs of affected individuals and $\text{IBD}_{i,j}(w)$ equals the number of alleles shared IBD by the i-th and j-th individual given w.

One may generalize this into a two-locus score function in several ways. We consider

$$S_{\text{pairs}}(w_1,w_2) = \sum_{i<j} \left[\text{IBD}_{i,j}(w_1) + \text{IBD}_{i,j}(w_2)\right]^k,$$

which for $k>1$ may be thought of as capturing epistatic joint pairwise IBD-sharing within a pedigree. The case $k = 1$ corresponds to the additive score function in [27], $S_{\text{pairs}}(w_1,w_2) = S_{\text{pairs}}(w_1) + S_{\text{pairs}}(w_2)$. Throughout all the relevant analyses we use $k = 2$.

D The Random Multiple Conditioning Loci Simulation Procedure

This case is computationally more demanding since p^c in (22) must be computed for all conditioning loci. Each p^c essentially corresponds to a one-locus p value, hence any of the methods described in Section 5.2 can be used.

Defining the original number of genome-wide one-locus simulations as J_1 and the additional number of simulations with respect to p^c in (22) as J_2, one has the following possibilities: (i) Estimate all p^cs *separately* through J_2 simulations using *inner* loops. This occupies less memory, but is computationally demanding. (ii) Save all single-test specific information[9] and estimate p values (p^cs) with respect to all the $|\mathbb{C}^1| + |\mathbb{C}^2| + ... + |\mathbb{C}^{J_1}|$ selected conditioning loci of the J_1 original simulations *simultaneously* using the *same* set of J_2 genome-wide simulated inheritance matrices in the additional (second) run. This occupies significantly more memory, but avoids the need for inner loops.

In our case, we have adopted (ii) and chosen $J_2 = 2500$ when using $S^{\text{Cox}}_{\text{pairs}}$ and $J_2 = 1000$ for $S^{\text{2loc}}_{\text{pairs}}$ and S_{opt}. Further, $J_1 = 2500$ in all three cases.

[9] Basically, this corresponds to the inheritance vectors at all the conditioning loci and their corresponding simulation-wise and chromosome-wise locations.

References

1 Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A: Loci on chromosomes 2 (NIDDM1) and 15 Interact to increase susceptibility to diabetes in Mexican Americans. Nat Genet 1999;21:213–215.
2 Risch N: Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 1990;46:242–253.
3 Holmans P: Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 1993;52:362–374.
4 Risch N: Segregation analysis incorporating genetic markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 1984;36:363–386.
5 Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J: Effects of misspecifying genetic parameters in lod score analysis. Biometrics 1986; 42:393–399.
6 Strauch K, Fürst R, Rüschendorf F, Windemuth C, Dietter J, Flaquer A, Baur MP, Wienker TF: Linkage analysis of alcohol dependence using MOD scores. BMC Genet 2005; 6(suppl 1):S162.
7 McPeek MS: Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 1999;16:225–249.
8 Sengul H, Weeks DE, Feingold E: A survey of affected-sibship statistics for nonparametric linkage analysis. Am J Hum Genet 2001;69: 179–190.
9 Hössjer O: Determining inheritance distributions via stochastic penetrances. J Am Stat Assoc 2003;98:1035–1051.
10 Lange EM, Lange K: Powerful allele-sharing statistics for nonparametric analysis. Hum Hered 2004;57:49–58.
11 Hössjer O: Conditional likelihood score functions for mixed models in linkage analysis. Biostatistics 2005;6;2:313–332.
12 Hössjer O: Information and effective number of meioses in linkage analysis. J Math Biol 2005;50;2:208–232.
13 Ängquist L: Some Notes on the Choice of Score Function in Non-parametric Linkage Analysis. Online 2006 (June). (Free download from 'http://www.maths.lth.se/matstat/staff/larsa/'.)
14 Ängquist L: Pointwise and Genomewide Significance Calculations in Gene Mapping through Nonparametric Linkage Analysis – Theory, Algorithms and Applications. Doctoral thesis 2006;15:2007, Department of Mathematical Statistics, Lund University, Lund.
15 Ängquist L: A Unified Discussion on the Concept of Score Functions Used in Nonparametric Linkage Analysis. Online 2007 (April). (Free down-load from 'http://www.maths.lth.se/matstat/staff/larsa/'.) Under revision for 'Bioinformatics and Biology Insights'.)
16 Lander E, Botstein D: Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 1986;83:19:7353–7357.
17 Schork NJ, Boehnke M, Terwilliger JD, Ott J: Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 1993;53:1127–1136.
18 Ängquist L, Anevski D, Luthman H: Unconditional two-locus nonparametric linkage analysis – on composite null hypotheses with and without gene-gene interaction. Tecknical Report 2005:28, Department of Mathematical Statistics, Lund University, Lund.
19 Li W, Reich J: A complete enumeration and classification of two-locus disease models. Hum Hered 2000;50:334–349.
20 Cordell HJ: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 2002;11: 2463–2468.

21 Strauch K, Fimmers R, Kurz T, Baur MP, Wienker TF: How to model a complex trait. 2. Analysis with two disease loci. Hum Hered 2003;56:200–211.
22 Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003;4:701–709.
23 Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M: Two-locus maximum lod score analysis of a multifactorial trait-joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. Am J Hum Genet 1995;57: 920–934.
24 Farrall M: Affected sibpair linkage tests for multiple linked susceptibility genes. Genet Epidemiol 1997;14:103–115.
25 Cordell HJ, Wedig GC, Jacobs KB, Elston RC: Multilocus linkage tests based on affected relative pairs. Am J Hum Genet 2000;66: 1273–1286.
26 Knapp M, Seuchter SA, Baur M: Two-locus disease models with two marker loci – the power of affected sib-pair tests. Am J Hum Genet 1994;55:1030–1041.
27 Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP: Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models – application to mite sensitization. Am J Hum Genet 2000;66:1945–1957.
28 Barber MJ, Todd JA, Cordell HJ: A multimarker regression-based test of linkage for affected sib-pairs at two linked loci. Genet Epidemiol 2006;30:191–208.
29 Dupuis J, Brown PO, Siegmund D: Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 1995;140:843–856.
30 Almasy L, Blangero J: Multipoint quantitative trait linkage analysis in general pedigrees. Am J Hum Genet 1998;62:1198–1211.
31 Sung YJ, Thompson EA, Wijsman EM: MCMC-based linkage analysis for complex traits on general pedigrees: multipoint analysis with a two-locus model and a polygenic component. Genet Epidemiol 2007;31:103–114.
32 Liang KY, Chiu YF, Beaty TH, Wjst M: Multipoint analysis using affected sib-pairs – incorporating linkage evidence from unlinked regions. Genet Epidemiol 2001;21:105–122.
33 Chiu YF, Liang KY: Conditional multipoint linkage analysis using affected sib pairs – an alternative approach. Genet Epidemiol 2004; 26:108–115.
34 Pinto D, Kasteleijn-Nolst Trenité DGA, Cordell HJ, Mattheisen M, Strauch K, Lindhout D, Koeleman BPC: Explorative two-locus linkage analysis suggests a multiplicative interaction between the 7q32 and 16p13 myoclonic seizures-related photosensitivity loci. Genet Epidemiol 2007;31:42–50.
35 Morahan G, Huang D, Tait BD, Colman PG, Harrison LC: Markers on distal chromosome 2q linked to insulin-dependent diabetes mellitus. Science 1996;272:1811–1813.
36 Davies JL, Kawaguchi Y, Bennett, ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SCL, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farrall M, Barnett AH, Bain SC, Todd JA: A genomewide search for human type 1 diabetes susceptibility genes. Nature 2002;371:130–136.
37 Schulze TG, Buervenich S, Badner JA, Steele CJM, Detera-Wadleigh SD, Dick D, Foroud T, Cox NJ, MacKinnon DF, Potash JB, Berrettini WH, Byerley W, Coryell W, DePaulo JR Jr, Gershon ES, Kelsoe JR, McInnis MG, Murphy DL, Reich T, Scheftner W, Nurnberger JI Jr, McMahon FH: Loci on chromosomes 6q and 6p interact to increase susceptibility to bipolar affective disorder in the national institute of mental health genetics Initiative pedigrees. Biol Psychiatry 2004;56: 18–23.
38 Sham P, Zhao JH, Curtis D: Optimal weighting scheme for affected sib-pair analysis of sibship data. Am J Hum Genet 1997;61:61–69.
39 Nilsson S: Two contributions to Genetic Linkage Analysis. Licentiate Thesis 1999, Chalmers University of Technology, Göteborg.
40 Hössjer O: Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. Ann Stat 2003;31:4: 1075–1109.
41 Tang HK, Siegmund D: Mapping multiple genes for quantitative or complex traits. Genet Epidemiol 2002;22:313–327.
42 Holmans P: Detecting gene-gene interactions using affected sib-pair analysis with covariates. Hum Hered 2002;53:92–102.
43 Donnelly KP: The probability that related individuals share some section of the genome identical by descent. Theor Popul Biol 1983; 23:34–64.
44 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis – a unified multipoint approach. Am J Hum Genet 1996;58:1347–1363.
45 Kong A, Cox N: Allele-sharing models – LOD scores and accurate linkage tests. Am J Hum Genet 1997;61:1179–1188.
46 Biernacka JM, Sun L, Bull SB: Simultaneous localization of two linked disease susceptibility genes. Genet Epidemiol 2005;28:33–47.
47 Ängquist L: Conditional Two-locus NPL-Analyses – Theory and Applications. Master's thesis 2001:E22, Department of Mathematical Statistics, Lund University, Lund.
48 Ängquist L, Hössjer O, Groop L: Strategies for conditional two-locus nonparametric linkage analysis. Technical Report 2007:1; Department of Mathematical Statistics, Lund University, Lund.
49 Ge Y, Dudoit S, Speed TP: Resampling-based multiple testing for microarray data analysis (with discussion.) Sociedad Espãnola de Estadística e Investigación Operativa Test 2003;12; 1:1–77.
50 Risch N: Linkage strategies for genetically complex traits – I. Multilocus models. Am J Hum Genet 1990;46:222–228.
51 Lander ES, Botstein D: Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 1989;121:185–199.
52 Feingold E: Markov processes for modeling and analyzing a new genetic mapping method. J Appl Prob 1993;30:766–779.
53 Feingold E, Brown PO, Siegmund D: Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 1993;53:234–251.
54 Lander ES, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 1995;11:241–247.
55 Boehnke M: Estimating the power of a proposed linkage study: a practical computer simulation approach. Am J Hum Genet 1986; 39:513–527.
56 Ploughman LM, Boehnke M: Estimating the power of a proposed linkage study for a complex genetic trait. Am J Hum Genet 1989;44: 543–551.
57 Ott J: Computer-simulation methods in human linkage analysis. Proc Natl Acad USA 1989;86;11:4175–4178.
58 Terwilliger JD, Speer M, Ott J: Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 1993;10:217–224.
59 Tang HK, Siegmund D: Mapping quantitative trait loci in oligogenic models. Biostatistics 2001;2:147–162.
60 Ängquist L, Hössjer O: Improving the calculation of statistical significance in genome-wide scans (with supplementary material online). Biostatistics 2005;6;4:520–538.
61 Hernández S, Siegmund DI, De Gunst M: On the power for linkage detection using a test based on scan statistics. Biostatistics 2005; 6;2:259–269.
62 Bacanu SA: Robust estimation of critical values for genome scans to detect linkage. Genet Epidemiol 2005;28:24–32.
63 Malley JD, Naiman DQ, Bailey-Wilson JE: A comprehensive method for genome scans. Hum Hered 2002;54:174–185.
64 Ängquist L, Hössjer O: Using importance sampling to improve simulation in linkage analysis. Stat Appl Genet Molec Biol 2004; 3;1;5. (Online journal; 24 pages).
65 Song KK, Weeks DE, Sobel E, Feingold E: Efficient simulation of p values for linkage analysis. Genet Epidemiol 2004;26:88–96.
66 Wigginton JE, Abecasis GR: An evaluation of the replicate-pool method: quick estimation of genome-wide linkage peak p values. Genet Epidemiol 2006;30:320–332.
67 Selin I: Detection Theory. Princeton, Princeton University Press, 1965.
68 Bradley AP: ROC Curves and the χ^2 test. Pattern Rec Lett 1996;17:287–294.

69 Whittemore AS, Halpern J: A class of tests for linkage using affected pedigree members. Biometrics 1994;50:118–127.
70 Khoury MJ, Beaty TH, Cohen BC: Fundamentals of Genetic Epidemiology. New York and Oxford, Oxford University Press, 1993 (Monographs on Epidemiology and Biostatistics, Volume 22).
71 Haines JL, Pericak-Vance MA: Genetic Analysis of Complex Disease. New York, Wiley-Liss, 2006.
72 Hössjer O: Combined association and linkage analysis for general pedigrees and genetic models. Stat Appl Genet Molec Biol 2005; 4;1;11. (Online journal, 42 pages).
73 Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 1987;85:2363–2367.
74 Kruglyak L, Daly MJ, Lander ES: Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 1995;56:519–527.
75 Ziegler A, Koenig IR: A Statistical Approach to Genetic Epidemiology – Concepts and Applications. Weinheim, Wiley-VCH, 2006.
76 Teng J, Siegmund D: Multipoint linkage analysis using affected relative pairs and partially informative markers. Biometrics 1998;54:1247–1265.
77 Risch N, Merikangas K: The future of genetic studies of complex human diseases. Science 1996;273:1516–1517.
78 The Wellcome Trust case control consortium: genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. Nature 2007;447:661–678.
79 Clerget-Darpoux F, Elston RC: Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genomewide association. Hum Hered 2007;64:91–96.
80 Kurbasic A, Hössjer O: A general method for linkage disequilibrium correction for multipoint linkage and association. Under revision for 'Genetic Epidemiology' 2007.
81 Davis TA, Sigmon K: MATLAB Primer, ed 7. Boca Raton, Chapman & Hall/CRC, 2005.
82 Mittelbach F, Goosens M: The L^AT_EX Companion – Tools and Techniques for Computer Typesetting. ed 2. Boston, Addison Wesley, 2004.