

# Unconditional Two-Locus Nonparametric Linkage Analysis

On Composite Null Hypotheses With and Without Gene-Gene Interaction

Lars Ängquist<sup>1</sup>      Dragi Anevski<sup>2</sup>  
Holger Luthman<sup>3</sup>

21st October 2005

<sup>1</sup>Department of Mathematical Statistics, Lund University, Lund, Sweden.

<sup>2</sup>Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden

<sup>3</sup>Bioinformatics Unit, Department of Endocrinology, Malmö University Hospital, Lund University, Malmö, Sweden

## Abstract

We discuss different aspects of unconditional two-locus nonparametric linkage (NPL) analysis with special emphasis on gene-gene interaction. We interpret this as identical-by-descent (IBD) sharing correlation between two disease loci both having marginal effect. We relate this to the concept of two-locus NPL score functions, the possible importance of using a composite rather than a simple null hypothesis and the corresponding calculation of statistical power. Moreover, we define several classes of score functions and give multiple suggestions on how to incorporate a composite null hypothesis into the analysis. The least favourable two-locus IBD-distribution is discussed, resulting in an upper bound of the two-locus  $p$ -value.

**Key words:** NPL analysis, unconditional two-locus linkage analysis, genetic disease models, IBD-sharing, gene-gene interaction, score functions, composite null hypothesis, least favourable distribution, Monte Carlo simulation, estimation of genetic parameters, power calculations.

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Genetic Disease Models</b>	<b>7</b>
2.1	One-Locus Model . . . . .	7
2.2	Two-Locus Model . . . . .	9
<b>3</b>	<b>Hypotheses for Statistical Tests</b>	<b>12</b>
3.1	One-Locus Hypotheses . . . . .	12
3.2	Two-Locus Hypotheses . . . . .	13
3.2.1	Simple Null Hypotheses . . . . .	13
3.2.2	Composite Null Hypotheses . . . . .	13
<b>4</b>	<b>Score Functions</b>	<b>14</b>
4.1	One-Locus Functions . . . . .	14
4.2	Two-Locus Functions . . . . .	15
4.3	Definitions . . . . .	16
4.4	Examples . . . . .	17
<b>5</b>	<b>Methods for Composite Null Hypotheses</b>	<b>17</b>
5.1	Least Favourable Distribution . . . . .	18
5.2	Significance Calculations Through Simulation . . . . .	20
5.2.1	Least Favourable Distribution Method . . . . .	20
5.2.2	The Estimated One-Locus IBD-Sharing Method . . . . .	21
<b>6</b>	<b>Power Calculations</b>	<b>22</b>
6.1	Simple Null Hypothesis . . . . .	22
6.1.1	Results . . . . .	24
6.1.2	Discussion . . . . .	24
6.2	Composite Null Hypothesis . . . . .	26
6.2.1	Considering Power . . . . .	26
6.2.2	Considering Significance Levels . . . . .	29
<b>7</b>	<b>Discussion</b>	<b>32</b>
<b>8</b>	<b>Acknowledgements</b>	<b>33</b>
	<b>References</b>	<b>33</b>

<i>CONTENTS</i>	3
<b>A One-Locus Allele-Sharing Probabilities</b>	<b>39</b>
<b>B Proof of Theorem 1</b>	<b>40</b>
<b>C Detailed Power Calculation Results</b>	<b>41</b>
<b>D Criteria for Further Inclusion of Cells</b>	<b>44</b>

## 1 Introduction

**Aims and Scope** One-locus nonparametric analysis for assessing genetic influence from a single gene on the development of a disease is well understood. The most common approaches are test methods based on the nonparametric linkage (NPL) score or the maximum lod score (MLS). In this context nonparametric refers to the fact that they presuppose no genetic model for the disease.

In this article we discuss different approaches to unconditional two-locus linkage analysis based on allele-sharing methods through two-locus versions of the NPL score. We use this to look for a joint genetic effect of two different genes on the development of a disease. An important aspect of this kind of analysis is the choice of either including or excluding, in the general hypotheses or models, the possibility of gene-gene interaction with respect to the identical-by-descent (IBD) sharing probabilities. Here we consider, for instance, the choice of a two-locus score function, the importance of allowing for gene-gene interaction, several two-locus null hypotheses and the corresponding calculation of statistical power. We consider both simple and composite null hypotheses and develop several methods for calculating significance levels and power.

Throughout this article we use pedigree sets consisting of affected sib-pairs (ASP) and assume perfect marker information, i.e. that the actual IBD-sharing through the corresponding inheritance vector, at each locus, is known with probability one.

There are two different ways to look for simultaneous effects from two unlinked loci: (i) The most common way is to do a *conditional* search. This means that one calculates the two-locus score conditioning on fixed information from one or several assumed loci of disease genes. The conditioning loci are either fixed or estimated, depending on whether the genes are known or not. In the latter case the conditioning loci may be chosen from the peaks of a one-locus scan. The IBD-information from these loci are used as weights in the second, conditional two-locus scan. See for instance Cox et al. (1999) and Ängquist and Hössjer (2005b) for more details. (ii) The second alternative is to perform an *unconditional* or *simultaneous* search for two disease-causing loci. There are advantages and disadvantages of both approaches. Simple implementations of conditional analysis may not in essence be geared towards capturing general gene-gene interactions and, more impor-

tant, it may not be easy to find the right conditioning loci leading to a loss of information and possible failure to detect susceptibility loci. On the other hand, the conditional approach normally includes significantly less multiple testing and is computationally more feasible.

The main aim of this paper is to compare the strength (power) of different two-locus testing methods, both under simple and composite null hypotheses, and to formulate a test that gives an upper bound on the significance level for the composite alternative.

**Previous Work** Dupuis et al. (1995) treat both conditional and simultaneous search of two unlinked disease-causing loci, in the latter case using a combination of the overall lod score process at the chromosomes. Restricting ourselves to two chromosomes,  $C_1$  and  $C_2$ , they essentially use the sum of two maximum lod score processes  $Z(\cdot)$  as

$$\max_{x_1} Z(x_1) + \max_{x_2} Z(x_2) \text{ with } x_1 \in C_1, x_2 \in C_2.$$

This is a measure using the total lod score on each chromosome, and therefore does not (a priori) capture the individual familial joint genetic effect. A similar approach was used in Strauch et al. (2000) where they suggested two-locus versions of the NPL score functions  $S_{\text{pairs}}$  and  $S_{\text{all}}$  (Whittemore and Halpern, 1994) as sums of the one-locus counterparts and implemented these into GENEHUNTER-TWOLOCUS. Recently, conditional and simultaneous search for quantitative traits and model selection are treated by Tang and Siegmund (2002) and Siegmund (2004) respectively.

Extensions of the MLS method are described by Farrall (1997) and in the two-locus case by Cordell et al. (2000). Zinn-Justin and Abel (1998) consider versions of the two-locus Weighted Pairwise Correlation (WPC) statistic, Lucek et al. (1998) describe a multilocus approach based on neural networks and Doerge and Churchill (1996) outline a permutation test in the multilocus quantitative trait linkage (QTL) setting.

Cox et al. (1999) perform conditional NPL analysis using, in principle, the two-locus pedigree-specific NPL score

$$Z(x_1, x_2) = Z(x_1)f(Z(x_2)) \text{ with } x_1 \in C_1, x_2 \in C_2,$$

where the argument of  $f$  depends on the one-locus NPL score at the fixed conditioning locus  $x_2$  and it produces pedigree weights for computing the total

NPL score when scanning through  $C_1$ . An alternative conditional approach is defined by Chiu and Liang (2004).

Connected, for instance, to the choice of an appropriate score function is the definition of the crucial concepts *interaction*, *epistasis* and *heterogeneity*. We adopt the interpretation of Cox et al. (1999) and Holmans (2002) where epistasis and heterogeneity are viewed as gene-gene interactions with positively and negatively correlated IBD-sharing probabilities. Recently, Vieland and Huang (2003) argued for a contrasting definition through penetrances. This was motivated by insisting on certain biological definitions. Later on, in the same journal volume, two replies of general disagreement with this view containing interesting comments and critique (Farrall, 2003; Cordell, 2003) were published. Holmans (2002) tackled gene-gene interaction through a logistic regression method and related articles directed towards variance components in QTL analysis are e.g. Tiwari and Elston (1998), Purcell and Sham (2004) and Culverhouse et al. (2004).

Disease models for the two- or multilocus cases are discussed by Risch (1990), MacLean et al. (1993), Knapp et al. (1994), Tiwari and Elston (1998) and Kämpe (2001). A complete enumeration of distinct disease models is given in Li and Reich (2000). They reduce the number of models by using permutations based on inherent data symmetry. Constraints for valid two-locus IBD-sharing probabilities are derived by Dudoit and Speed (1998) and Bengtsson (2001).

Several areas within nonparametric linkage analysis in general are discussed by Kruglyak and Lander (1995), Kruglyak et al. (1996), Kong and Cox (1997) and Teng and Siegmund (1998). The performance of different score functions, mainly in the one-locus case, is investigated by Whittemore and Halpern (1994), Davis and Weeks (1997), McPeck (1999), Feingold et al. (2000), Sengul et al. (2001) and Hössjer (2005a).

Finally, interesting recent review papers on two- or multilocus models are Hoh and Ott (2003) and Strauch et al. (2003) and textbooks on linkage analysis in general are Sham (1998) and Ott (1999), whereas a shorter review is given by Ott and Hoh (2000).

**Outline of Article** In *Section 2* we discuss different one- and two-locus genetic models and in *Section 3* the relevant hypotheses (null and alternative) for testing two-locus genetic effects. Connected to this problem is the choice of a proper score function and in *Section 4* we introduce general classes

of regular and restricted two-locus score functions and give some examples. Next, in *Section 5* we discuss the least favourable two-locus distribution when calculating theoretical  $p$ -values for an arbitrary regular score function. We derive both local (pointwise) and global (genome-wide) results. Moreover, some Monte Carlo simulation methods and approaches to significance calculations are discussed. *Section 6* is devoted to power calculations using first the simple null hypothesis and then composite ones. This is made for several score functions and genetic models. We assume constant marginal (one-locus) IBD-sharing probabilities at the two assumed disease loci, but use different strength of the IBD-sharing correlation. Further discussion and conclusions are presented in *Section 7* and the appendices contain some results on allele-sharing probabilities, a proof related to the least favourable distribution, tables with the complete results from the power and significance calculations and some notes on the choice of a restricted score function.

## 2 Genetic Disease Models

**IBD-Sharing** The concept of sharing alleles IBD is of central importance to the remainder of this article. For an underlying pair of individuals, it is defined as the property of inheriting copies of the same founder alleles (see e.g. Sham, 1998; Ott, 1999; Strachan and Read, 2003). Since genotypes consists of two alleles, it is formally possible to, with probabilities depending on the specific pair of relatives, share 0,1 or 2 alleles IBD. At locus  $x$ , we define these events to be the possible outcomes of the corresponding stochastic variable  $IBD(x)$ .

**Assumptions** Throughout, we assume binary phenotypes, biallelic disease loci, no parental imprinting, no interference and random inheritance. Consider Table 1 for some basic notation.

### 2.1 One-Locus Model

Let  $z_i(x) = P(IBD(x) = i|ASP)$  ( $i=0,1$  or  $2$ ) be the probability that an affected sib-pair shares  $i$  alleles IBD at locus  $x$ . If  $l$  is the disease locus, we may express this identity using the following expansion based on conditional

Table 1: Basic notation.

Notation	Meaning
$D$	disease allele
$d$	nondisease allele
$p$	frequency of $D$
$q=1-p$	frequency of $d$
$PG$	parental genotype
$SG$	sib-pair genotype
$ASP$	affected sib-pair
$IBD$	identical-by-descent
$f_i,$ $i=0,1 \text{ or } 2$	penetrances i.e. $P(\text{disease}   i \text{ } D\text{s})$

probabilities,

$$z_i(l) = \frac{\sum_{PG} \sum_{SG} P(IBD(l) = i | PG, SG) P(ASP | SG) P(SG | PG) P(PG)}{\sum_{PG} \sum_{SG} P(ASP | SG) P(SG | PG) P(PG)}. \quad (1)$$

More information on (1) is given in Appendix A. From now on, for ease of notation, we assume that the calculations are performed at a disease locus, i.e.  $z_i = z_i(l)$ . Further, we define the event of no genetic component as random inheritance given affection status, i.e. corresponding to IBD-sharing probabilities  $z = (z_0, z_1, z_2) = (1/4, 1/2, 1/4)$  at the (false) disease locus. One-locus IBD-probability constraints are defined by Holmans (1993).

**Note 1** *The rightmost term in the numerator of (1) depends on the disease allele frequency  $p$ , the second term is a function of the penetrances  $f_i$  and the remaining ones follows through combinatorial arguments given the joint structure of the genotypes of the parents and the sib-pair.*

Next, we will calculate the probabilities in (1) for two distinct fully penetrant one-locus disease models: the recessive and the dominant model. Let  $f = (f_0, f_1, f_2)$  be the penetrance vector.

**Example 1** *(The Recessive Model) We have  $f = (0, 0, 1)$ , i.e. two copies of the disease allele  $D$  (homozygosity) will cause the disease. The IBD-sharing probabilities with respect to different choices of the disease-allele frequency are displayed in Table 2.*

Table 2: IBD-probabilities given an ASP for the recessive model.

$p$	$z_0$	$z_1$	$z_2$
0.01	0.0001	0.0196	0.9803
0.10	0.0083	0.1653	0.8264
0.25	0.0400	0.3200	0.6400
0.50	0.1111	0.4444	0.4444
0.75	0.1837	0.4898	0.3265
1.00	0.2500	0.5000	0.2500

**Example 2** (*The Dominant Model*) We have  $f = (0, 1, 1)$ , i.e. at least one copy of the disease allele  $D$  will cause the disease. The IBD-sharing probabilities with respect to different choices of the disease-allele frequency are displayed in Table 3.

Table 3: IBD-probabilities given an ASP for the dominant model.

$p$	$z_0$	$z_1$	$z_2$
0.01	0.0098	0.4988	0.4914
0.10	0.0813	0.4909	0.4278
0.25	0.1565	0.4856	0.3578
0.50	0.2195	0.4878	0.2927
0.75	0.2443	0.4951	0.2606
1.00	0.2500	0.5000	0.2500

**Note 2** *The maximal probability for sharing two alleles IBD is 1 in the recessive and 0.5 in the dominant case. This means that the ASP is more informative for rare recessive than for rare dominant diseases, see e.g. Hössjer (2005b).*

## 2.2 Two-Locus Model

In the two-locus case, let

$$z_{ij}(x_1, x_2) = P(\text{IBD}(x_1, x_2) = (i, j) | \text{ASP}) \quad i, j \in \{0, 1, 2\}$$

be the probability that an affected sib-pair shares  $i$  and  $j$  alleles IBD at locus  $x_1$  and  $x_2$  respectively. We only consider probabilities at disease loci  $l_1$  and  $l_2$ , i.e.  $z_{ij} = z_{ij}(l_1, l_2)$ .

The two-locus analogue of expansion (1) for calculation of IBD-sharing probabilities is

$$z_{ij} = \frac{\sum_{PG} \sum_{SG} P(\text{IBD} = (i, j) | PG, SG) P(\text{ASP} | SG) P(SG | PG) P(PG)}{\sum_{PG} \sum_{SG} P(\text{ASP} | SG) P(SG | PG) P(PG)}, \quad (2)$$

where the involved loci are assumed to be located on distinct chromosomes,  $PG = (PG_1, PG_2)$  and  $SG = (SG_1, SG_2)$  are the joint parental and sib-pair genotypes with respect to loci  $l_1$  and  $l_2$ . For discussions on two-locus IBD-probability constraints, see Cordell et al. (1995), Dudoit and Speed (1998) and Bengtsson (2001).

**Note 3** *The rightmost term in the numerator of (2) depends on the disease allele frequencies  $p_1$  and  $p_2$ , the second term is related to the two-locus penetrances*

$$f = \begin{pmatrix} f_{00} & f_{01} & f_{02} \\ f_{10} & f_{11} & f_{12} \\ f_{20} & f_{21} & f_{22} \end{pmatrix}, \quad (3)$$

where  $f_{ij}$  refers to the probability of being affected given that the genotypes  $G_1$  and  $G_2$  at loci  $l_1$  and  $l_2$  contain  $i$  and  $j$  copies of the disease alleles  $D_1$  and  $D_2$ .

As for the one-locus case, we will now give two examples of calculation of the IBD-sharing probabilities corresponding to (2). For ease of presentation, we let  $p_1 = p_2$ .

**Example 3** *(The Recessive-Recessive Model) We have*

$$f = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

i.e. joint homozygosity with respect to the disease alleles  $D_1$  and  $D_2$ , at the first and second disease locus respectively, will cause the disease. The IBD-sharing probabilities found when varying the common disease-allele frequency are displayed in Table 4.

Table 4: IBD-probabilities given an ASP for the Recessive-Recessive model.

$z_{ij}$	$p_1=p_2$	$j=0$	$j=1$	$j=2$	$p_1=p_2$	$j=0$	$j=1$	$j=2$
$i=0$	0.01	0.0000	0.0000	0.0004	0.50	0.0123	0.0494	0.0494
$i=1$		0.0000	0.0001	0.0192		0.0494	0.1975	0.1975
$i=2$		0.0004	0.0192	0.9610		0.0494	0.1975	0.1975
$i=0$	0.10	0.0001	0.0014	0.0068	0.75	0.0337	0.0900	0.0600
$i=1$		0.0014	0.0273	0.1366		0.0900	0.2399	0.1599
$i=2$		0.0068	0.1366	0.6830		0.0600	0.1599	0.1066
$i=0$	0.25	0.0016	0.0128	0.0256	1.00	0.0625	0.1250	0.0625
$i=1$		0.0128	0.1024	0.2048		0.1250	0.2500	0.1250
$i=2$		0.0256	0.2048	0.4096		0.0625	0.1250	0.0625

**Example 4** (*The Dominant-Dominant Model*) We have

$$f = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

*i.e.* at least one copy (jointly) of the disease alleles  $D_1$  and  $D_2$ , at the first and second disease locus respectively, will cause the disease. The IBD-sharing probabilities found when varying the common disease-allele frequency are displayed in Table 5

**Note 4** *The maximal probability for simultaneously sharing two alleles IBD at both disease loci is  $1^2 = 1$  in the recessive-recessive case and  $0.5^2 = 0.25$  in the dominant-dominant case.*

Next, we define the marginal (one-locus) IBD-sharing probabilities in this two-locus setting as

$$\begin{aligned} z^1 &= (z_0^1, z_1^1, z_2^1) \text{ with } z_i^1 = \sum_j z_{ij}, \\ z^2 &= (z_0^2, z_1^2, z_2^2) \text{ with } z_j^2 = \sum_i z_{ij}, \end{aligned} \tag{4}$$

corresponding to the row and column totals in the 3x3 two-locus IBD-sharing matrix  $\{z_{ij}\}$  respectively. Now, the general definition of gene-gene interaction of IBD-sharing is given by:

Table 5: IBD-probabilities given an ASP for the Dominant-Dominant model.

$z_{ij}$	$p_1=p_2$	$j=0$	$j=1$	$j=2$	$p_1=p_2$	$j=0$	$j=1$	$j=2$
$i=0$	0.01	0.0001	0.0049	0.0048	0.50	0.0482	0.1071	0.0642
$i=1$		0.0049	0.2488	0.2451		0.1071	0.2380	0.1428
$i=2$		0.0048	0.2451	0.2415		0.0642	0.1428	0.0857
$i=0$	0.10	0.0066	0.0399	0.0348	0.75	0.0597	0.1210	0.0637
$i=1$		0.0399	0.2410	0.2100		0.1210	0.2451	0.1290
$i=2$		0.0348	0.2100	0.1830		0.0637	0.1290	0.0679
$i=0$	0.25	0.0245	0.0760	0.0560	1.00	0.0625	0.1250	0.0625
$i=1$		0.0760	0.2358	0.1738		0.1250	0.2500	0.1250
$i=2$		0.0560	0.1738	0.1280		0.0625	0.1250	0.0625

**Definition 1** *IBD-sharing correlation or gene-gene IBD-sharing interaction* is present when, for at least one pair  $(i,j)$ ,

$$z_{ij} \neq z_i^1 z_j^2, \quad 0 \leq i, j \leq 2, \quad (5)$$

where  $z^1$  and  $z^2$  corresponds to the marginal sharing defined in (4).

**Note 5** For ASPs and multiplicative two-locus penetrance models, i.e. where  $f_{ij} = f_i^1 f_j^2$ , one may prove that  $z_{ij} = z_i^1 z_j^2$ . Here  $f^1$  and  $f^2$  refer to marginal penetrances and  $z_i^1$  and  $z_j^2$  are the corresponding one-locus probabilities (1).

### 3 Hypotheses for Statistical Tests

Crucial for the possibility of performing statistical tests is the definition of proper testing hypotheses, i.e. the null hypothesis ( $H_0$ ) and the alternative hypothesis ( $H_1$ ).

#### 3.1 One-Locus Hypotheses

The standard one-locus hypotheses are commonly defined as

$$\begin{aligned} H_0 : z &= (0.25, 0.5, 0.25), \\ H_1 : z &\neq (0.25, 0.5, 0.25), \end{aligned} \quad (6)$$

where we still implicitly assume that  $z = z(l)$  for disease locus  $l$ .

**Note 6** *An equivalent way of formulating this is in terms of penetrances. A genetic effect of the studied disease is present if  $f_0 = f_1 = f_2$  does not hold.*

### 3.2 Two-Locus Hypotheses

Assume we have two different chromosomes  $C_1$  and  $C_2$ , on which we suspect there are loci  $l_1$  and  $l_2$ , that jointly affect the disease of study. Let  $z = \{z_{ij}\}$  be the two-locus IBD-sharing of interest. We would like to stress that no restriction is put on how the two genes interact. Furthermore we are mainly interested in joint effects, i.e. not in finding out whether one of the loci has an effect on its own. This corresponds to the following general hypotheses

$$\begin{aligned} H_0 &: \text{At most one of the loci } l_1 \text{ and } l_2 \text{ have an effect,} \\ H_1 &: \text{Both loci } l_1 \text{ and } l_2 \text{ have an effect,} \end{aligned}$$

where the null and alternative hypothesis mean no and present joint genetic effect respectively.

Let  $\mathbb{Z}$  denote the set of possible two-locus IBD-probabilities in (2), as derived by Bengtsson (2001). Below we will define the two-locus null hypotheses as a given subset  $\mathbb{Z}_0$  of  $\mathbb{Z}$ . By varying  $\mathbb{Z}_0$  we incorporate both simple and composite null hypotheses. Assuming no a priori information about the location of  $l_1$  and  $l_2$ , we require  $\mathbb{Z}_0$  to be symmetric, i.e.  $z = z_{ij} \in \mathbb{Z}_0$  implies that  $z' = z_{ji} \in \mathbb{Z}_0$ .

#### 3.2.1 Simple Null Hypotheses

The simple null hypothesis assuming absence of both marginal genetic effects and gene-gene interaction is

$$H_0^{(1)} : \mathbb{Z}_0 = \{z = (0.25, 0.5, 0.25)^T * (0.25, 0.5, 0.25)\} \quad (7)$$

and the corresponding alternative hypothesis may be interpreted as

$$H_1^{(1)} : \text{At least one disease loci OR gene-gene interaction.}$$

#### 3.2.2 Composite Null Hypotheses

Using a composite null hypothesis we narrow the region of the alternative hypothesis by allowing for the presence of either or both of the following

properties in the null hypothesis: (i) One disease locus. (ii) Gene-gene interaction. Three distinct options are given:

$$H_0^{(2)} : \mathbb{Z}_0 = \{z ; z^1 = z^2 = (0.25, 0.5, 0.25)\}, \quad (8)$$

where  $z^1$  and  $z^2$  are the marginal sharing vectors defined in (4),

$$H_0^{(3)} : \mathbb{Z}_0 = \{(z^1)^T * (0.25, 0.5, 0.25), z^1 \in A\} \cup \{(0.25, 0.5, 0.25)^T * z^2, z^2 \in A\}, \quad (9)$$

where  $A$  is a given subset of one-locus IBD-probabilities and

$$H_0^{(4)} : \mathbb{Z}_0 = \{z ; z^1 = (0.25, 0.5, 0.25) \text{ or } z^2 = (0.25, 0.5, 0.25)\}. \quad (10)$$

The alternative hypotheses corresponding to  $H_0^{(2)}$ - $H_0^{(4)}$  may be interpreted as

$H_1^{(2)}$  : At least one disease loci,

$H_1^{(3)}$  : Two disease loci OR gene-gene interaction,

$H_1^{(4)}$  : Two disease loci.

**Note 7** We can actually narrow the definition of joint effect even further if restricting  $H_1^{(4)}$  by demanding presence of gene-gene interaction, i.e.

$H_1^{(5)}$  : Two disease loci and gene-gene interaction.

In this work we will mainly consider null hypotheses (7), (9) and (10).

## 4 Score Functions

### 4.1 One-Locus Functions

A score function  $S : \{0, 1, 2\} \rightarrow \mathbb{R}$  is a function that given a certain IBD-sharing gives the corresponding pedigree (ASP) a numeric score (weight). A pedigree-specific NPL score  $Z$  is score function applied to a specific locus or specific loci. The pedigree set analogue combines the  $n$  distinct pedigree scores available and will henceforth simply be referred to as the NPL score.

Let us start with the  $i^{\text{th}}$  sib-pair and define  $IBD_i(x)$  to be the number of alleles they share IBD at locus  $x$ . Now, their pedigree-specific NPL score at

this locus is denoted by  $Z_i(x) = S(IBD_i(x))$ . The standardized (pedigree set) NPL score is then given by

$$Z(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i(x) - \mu(x)}{\sigma(x)}, \quad (11)$$

where  $\mu(x) = E(Z_i(x))$  and  $\sigma^2(x) = V(Z_i(x))$ , the expected value and variance under the null hypothesis.

To test for increased IBD-sharing among  $M$  chromosomes  $C_1, \dots, C_M$  one may define the maximum NPL statistic with respect to the genome scanning region  $\Omega = \cup_{i=1}^M C_i$  as  $Z_{\max} = \max_{x \in \Omega} Z(x)$ .

Let  $\theta = (z, l)$  denote the genetic model parameters, with  $z = (z_0, z_1, z_2)$  as the IBD-sharing probabilities at disease locus  $l$ . Given a threshold  $T$ , define the global power function

$$\beta(T; \theta) = P(Z_{\max} \geq T | \theta), \quad (12)$$

and let  $\alpha(T; \theta) = \alpha(T)$  be the global significance level. This is the constant value of  $\beta(T; \theta)$  when  $\theta \in H_0$ . The global  $p$ -value is  $\alpha(z_{\max})$ , where  $z_{\max}$  is the observed value of  $Z_{\max}$ . The local power function  $\beta_{\text{local}}(T; \theta)$  and significance level  $\alpha_{\text{local}}(T)$  are defined analogously by replacing  $Z_{\max}$  by  $Z(l)$  and then putting  $\theta = z$ , since the position of disease locus does affect neither the pointwise significance level nor the pointwise power.

## 4.2 Two-Locus Functions

Extending the concept of score functions to the two-locus case is straightforward.  $S(i, j)$  simply equals the numeric score connected to jointly sharing  $i$  and  $j$  alleles IBD. The formulation of the pedigree-specific NPL score  $Z_i(x_1, x_2) = S(IBD_i(x_1), IBD_i(x_2))$  for the  $i^{\text{th}}$  ASP at loci  $x_1$  and  $x_2$  now leads to the standardized (pedigree set) NPL score given by

$$Z(x_1, x_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i(x_1, x_2) - \mu(x_1, x_2)}{\sigma(x_1, x_2)}, \quad (13)$$

where  $Z_i(x_1, x_2)$ ,  $\mu(x_1, x_2)$  and  $\sigma^2(x_1, x_2)$  are the two-locus analogues, with respect to the loci  $x_1$  and  $x_2$ , of the pedigree-specific NPL score, expected value and variance under the null hypothesis.

Moreover, the maximum NPL statistic for unlinked loci (i.e. located on distinct chromosomes) over  $\Omega$  is generalized to

$$Z_{\max} = \max_{x_1, x_2 \in \Omega} Z(x_1, x_2), \quad C(x_1) \neq C(x_2), \quad (14)$$

where  $C(x)$  is the chromosome where  $x$  is located. This leads to two-locus equivalents of the global power (12) and significance level, with  $\theta = (z, l)$ ,  $z = (z_{ij})_{i,j=0}^2$  as defined in Section 2.2 and  $l = (l_1, l_2)$  consisting of the two disease loci. The local power and significance level are defined analogously, replacing  $Z_{\max}$  by  $Z(l_1, l_2)$  everywhere, and with  $\theta = z$ .

**Note 8** *The global and local significance levels  $\alpha(T; \theta)$  and  $\alpha_{\text{local}}(T; \theta)$  in general vary with  $\theta \in H_0$  when the null hypothesis is composite.*

**Note 9** *For the two-locus case, we denote the score (weight) matrix by*

$$S = \begin{pmatrix} S(0,0) & S(0,1) & S(0,2) \\ S(1,0) & S(1,1) & S(1,2) \\ S(2,0) & S(2,1) & S(2,2) \end{pmatrix}. \quad (15)$$

**Note 10** *(i) Under a simple null hypothesis and perfect marker information  $\mu(x_1, x_2)$  och  $\sigma^2(x_1, x_2)$  will not depend on the chromosomal positions  $x_1$  and  $x_2$ . (ii) In general, for a composite null hypothesis,  $\mu(x_1, x_2)$  and  $\sigma(x_1, x_2)$  will depend on  $\theta \in H_0$ , see Sections 5.2 and 6 for further details.*

### 4.3 Definitions

A natural restriction on a score function  $S$  for testing  $H_0$  against  $H_1$  is that it is monotone with respect to partial IBD-ordering.

**Definition 2** *A score function such that*

$$\begin{aligned} S(i) &\leq S(j); \quad i \leq j, \\ S(i, j) &\leq S(k, l); \quad i \leq k, \quad j \leq l, \end{aligned} \quad (16)$$

*in the one-locus and two-locus case respectively, is called a **regular** score function.*

Our next definition concerns the two-locus weights given in (15).

**Definition 3** *A score function is called **restricted** if the weights in  $S$  only attain values 0 or 1, otherwise it is **unrestricted**. If  $S(i, j) = 1$  then the  $(i, j)^{\text{th}}$  element is **included** in the **selection** of positive weights.*

#### 4.4 Examples

A commonly used unrestricted one-locus score function is

$$S(i) = i, \quad (17)$$

which is known as the mean sharing score function. Several known score functions, such as  $S_{\text{pairs}}$  and  $S_{\text{all}}$  (Whittemore and Halpern, 1994) reduce to (17) for ASPs. Next, we formulate four different versions of two-locus unrestricted score functions,

$$\begin{aligned} S_1(i, j) &= S(i) + S(j) \quad (\text{additive}), \\ S_2(i, j) &= S(i) S(j) \quad (\text{multiplicative}), \\ S_3(i, j) &= S(i)^2 S(j)^2 \quad (\text{quadratic multiplicative}), \\ S_4(i, j) &= \min(S(i), S(j)) \quad (\text{minimum}), \end{aligned}$$

where  $(i, j)$  refers to joint IBD-sharing. All four examples are defined as functions of the corresponding pair of one-locus scores.

When  $S$  is the mean sharing score function,  $S_1$ - $S_4$  are as in (18). In addition, four restricted score functions  $S_5$ - $S_8$  are introduced in (18).

$$\begin{aligned} S_1 &= \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{pmatrix}, & S_2 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 4 \end{pmatrix}, & S_3 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 4 & 16 \end{pmatrix}, & S_4 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \\ S_5 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, & S_6 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, & S_7 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, & S_8 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}. \end{aligned} \quad (18)$$

## 5 Methods for Composite Null Hypotheses

By the Central Limit Theorem, all score functions (11) and (13) have standard normal distributions asymptotically as  $n \rightarrow \infty$  under a simple null hypothesis. However, different score functions will give rise to different power functions under the alternative hypothesis, implying that they will differ in their power to detect deviances from the null hypothesis.

An extra difficulty occurs when the two-locus null hypothesis is composite. Then, with respect to a given score threshold, the significance level depends on the actual instance of the hypothesis. There are several ways to deal with this: (i) Replace the composite null hypothesis with the simple null hypothesis  $H_0^{(1)}$ . However, by shrinking the null hypothesis we also weaken the meaning of rejecting the test. Even though the power function is increased,

its interpretation changes. (ii) Replace the composite null hypothesis with its least favourable instance. We will then overestimate the  $p$ -values, i.e. perform conservative calculations since the  $p$ -value for the least favourable instance is at least as large as for any other element of the given hypothesis. (iii) A form of projection testing, obtained by first estimating parameters from data and then choosing the instance of the composite null hypothesis corresponding to these parameters.

### 5.1 Least Favourable Distribution

The most conservative test is obtained by maximizing the  $p$ -value over the null hypothesis, or equivalently by replacing  $H_0$  with the least favourable distribution in  $H_0$ . Thus, the significance level for the local test for the least favourable distribution in  $H_0$  is

$$\bar{\alpha}_{\text{local}}(T) = \max_{z \in H_0} \alpha_{\text{local}}(T; z) = \max_{z \in H_0} P(Z(x_1, x_2) \geq T | z), \quad (19)$$

where  $x_1$  and  $x_2$  is any pair of loci on different chromosomes. The global test analogue is

$$\bar{\alpha}(T) = \max_{\theta \in H_0} \alpha(T; \theta) \quad (20)$$

Now, we present a theorem whose proof is given in Appendix B.

**Theorem 1** *Consider the composite two-locus null hypothesis (10). Assume  $\Omega = C_1 \cup C_2$ , perfect marker data and that  $S$  is a symmetric and regular score function. Then there are two least favourable distributions  $z_a = (0, 0, 1)^T * (0.25, 0.5, 0.25)$  and  $z_b = (0.25, 0.5, 0.25)^T * (0, 0, 1)$  for the point-wise significance level, i.e.  $\bar{\alpha}_{\text{local}}(T) = \alpha_{\text{local}}(T; z_a) = \alpha_{\text{local}}(T; z_b)$ . This gives rise to the standardization*

$$\begin{aligned} \mu &= \frac{1}{4}(S(0, 2) + 2S(1, 2) + S(2, 2)), \\ \sigma^2 &= \frac{1}{4}(S(0, 2)^2 + 2S(1, 2)^2 + S(2, 2)^2) - \mu^2 \end{aligned}$$

of  $S$  and (total) NPL score

$$Z(x_1, x_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{S(\text{IBD}_i(x_1), \text{IBD}_i(x_2)) - \mu}{\sigma}.$$

with  $x_i \in C_i$ ,  $i = 1, 2$ . Moreover, the asymptotic local significance level of the least favourable distribution satisfies

$$\lim_{n \rightarrow \infty} \bar{\alpha}_{\text{local}}(T) = 1 - \Phi(T), \quad (21)$$

where  $\Phi$  is the standard normal cumulative distribution function.

Let  $|C_i|$  be the length of  $C_i$ . Then, for the global  $p$ -value,  $z_a$  is least favourable if  $|C_1| < |C_2|$ ,  $z_b$  is least favourable if  $|C_1| > |C_2|$  and both distributions are least favourable if  $|C_1| = |C_2|$ . The global significance level satisfies

$$\bar{\alpha}(T) = \begin{cases} \alpha(T; z_a, l), & \text{if } |C_1| \leq |C_2| \\ \alpha(T; z_b, l), & \text{if } |C_1| \geq |C_2| \end{cases} \quad (22)$$

where  $l = (l_1, l_2)$  and  $l_i \in C_i$ ,  $i = 1, 2$ . Asymptotically

$$\lim_{n \rightarrow \infty} \bar{\alpha}(T) = P\left(\max_{0 \leq x \leq \max(|C_1|, |C_2|)} U(x) \geq T\right) \quad (23)$$

where  $\{U(x)\}$  is a mixture of Ornstein-Uhlenbeck processes, as shown in Appendix B.

**Note 11** The term  $P(\max_x U(x) \geq T)$  can be approximated using extreme value theory of Gaussian processes, see Feingold et al. (1993), Lander and Kruglyak (1995), Tang and Siegmund (2001), Hössjer (2003b) and Ängquist and Hössjer (2005a). These formulas involve the crossover rate  $\rho = -r'_U(0)/2$ , where  $r_U(t) = \text{Cov}(U(x), U(x+t))$  is the covariance function. See Appendix B for details.

The reasoning in the proof shows that a similar result is possible to obtain also for larger sibships, if one limits oneself to studying pairwise IBD-counts.

**Note 12** Theorem 1 can be generalized to  $M > 2$  chromosomes. Then the least favourable distribution  $\theta = (z, l_1, l_2)$  of the global significance level is defined as follows: Let  $C_{\min}$  be the shortest chromosome, assume  $l_1 \in C_{\min}$  and put  $z = (0, 0, 1)^T * (0.25, 0.5, 0.25)$ .

**Note 13** Theorem 1 and its generalization in Note 12 can be formulated for the more restrictive null hypothesis (9), provided  $(0, 0, 1) \in A$  (since then  $z_a$  and  $z_b$  belong to  $H_0$ ). When  $(0, 0, 1) \notin A$  in (9), we can still find least favourable distributions  $z_c = (z^1)^T * (0.25, 0.5, 0.25)$  and  $z_d = (0.25, 0.5, 0.25)^T * z^2$  by maximizing  $\sum_{ij} S(i, j)z_{ij}$  with respect to  $z^1$  or  $z^2$  for

any given symmetric and regular score function  $S$ , provided the maximum, attained at  $z^1 = z^2$ , is unique. However, we no longer have the asymptotic characterization (23) of  $\bar{\alpha}(T)$ , but essentially have to estimate it using simulation. Moreover, the positioning of  $l_1 \in C_{\min}$  is no longer arbitrary, but rather  $l_1$  is placed in the middle of  $C_{\min}$ .

## 5.2 Significance Calculations Through Simulation

Throughout, for a given score threshold  $T$  we denote the unknown  $p$ -value with  $\alpha(T)$  and the estimated counterpart with

$$\hat{\alpha}(T; \theta) = \frac{1}{J} \sum_{j=1}^J I(Z_{\max}^j \geq T), \quad (24)$$

where  $J$  is the number of independent and identically distributed simulations under  $\theta \in H_0$ ,  $Z_{\max}^j$  refers to the maximum of the NPL score  $Z^j(\cdot)$  in the one-locus case and of  $Z^j(\cdot, \cdot)$  in the two-locus case during the  $j^{\text{th}}$  simulation and  $I(A)$  is the indicator function for the event  $A$ . Similarly the local significance level is estimated as  $\hat{\alpha}_{\text{local}}(T; \theta)$ , replacing  $Z_{\max}^j$  by  $Z^j(l)$  or  $Z^j(l_1, l_2)$  in (24).

For further reading on Monte Carlo simulation in linkage analysis see e.g. Boehnke (1986), Ott (1989), Terwilliger et al. (1993), Song et al. (2004) and Ängquist and Hössjer (2004).

### 5.2.1 Least Favourable Distribution Method

Assume a null hypothesis (10) or (9) with the conditions of Note 13 satisfied. The least favourable distribution is defined for a general number  $M$  of chromosomes in Note 12. It has the form  $\theta = (z, l_1, l_2)$ , with  $l_1 \in C_{\min}$ . In order to estimate the global significance level for this  $\theta$ , we can use (24) by requiring  $IBD^j(l_1) = 2$  and that  $\{IBD^j(x)\}$  is distributed as a one-locus process under  $H_0$  when  $x \notin C_{\min}$  for  $j = 1, \dots, J$ . Then  $\hat{\alpha}(T; \theta)$  is a Monte Carlo estimate of  $\bar{\alpha}(T)$ . This is evident from (22) when the number of chromosomes  $M = 2$  and follows similarly for general  $M$ . The reasoning for the local significance level is analogous. We require  $IBD^j(l_1) = 2$  and that  $IBD^j(l_2)$  is distributed as under  $H_0$ .

### 5.2.2 The Estimated One-Locus IBD-Sharing Method

For null hypothesis (9), we estimate the global significance as  $\hat{\alpha}(T) = \hat{\alpha}(T; \hat{\theta})$ , where  $\hat{\alpha}(T; \cdot)$  is the Monte Carlo estimated significance level in (24) and

$$\hat{\theta} = \left( (\hat{z}_0, \hat{z}_1, \hat{z}_2)^T * (0.25, 0.5, 0.25), (\hat{l}_1, l_2) \right) = (\hat{z}, \hat{l}) \quad (25)$$

is the estimated genetic model from data. It consists of the estimated disease locus  $\hat{l}_1$ , defined as the maximizer of the one-locus NPL score,

$$\hat{l}_1 = \arg \max_{x \in \Omega} Z(x), \quad (26)$$

and the estimated IBD-proportions  $\hat{z}_j = |\{i; IBD_i(\hat{l}_1) = j\}|/n$  at  $\hat{l}_1$ . In case  $\hat{z} = (\hat{z}_0, \hat{z}_1, \hat{z}_2) \notin A$ , we project  $\hat{z}$  onto  $A$ , i.e. we choose as our estimated one-locus IBD-probabilities the element of  $A$  closest to  $\hat{z}$  in Euclidean distance. Local significance levels are estimated analogously as  $\hat{\alpha}_{\text{local}}(T) = \hat{\alpha}_{\text{local}}(T, \hat{\theta})$ .

The performance of this estimator depends on the accuracy of the estimate  $\hat{\theta}$  as well as the  $\alpha$ -function itself, with respect to its parameters. One attractive feature is that it is often conservative, rather than anticonservative. This follows since  $\hat{z}$  is based on the inheritance at  $\hat{l}_1$  instead of at the unknown position  $l_1$ . Since these positions might differ, according to random fluctuation of the NPL process, the IBD-sharing is generally overestimated (Hössjer, 2003a). Using a regular score function, keeping the other parameters fixed, this indicates conservative  $p$ -values. The level of conservativeness depends on the difference between the real and estimated IBD-proportions,  $z(l_1)$  and  $\hat{z}(\hat{l}_1)$ . Obviously this depends, for instance, on the length of the genome region  $|\Omega|$ , the position of  $l_1$  with respect to the chromosome  $C(l_1)$  and the strength of the genetic effect.

**Note 14** *The least favourable distribution method corresponds to estimating the genetic effect by  $\hat{z} = (0, 0, 1)$  independent of the actual data and, implicitly, the position of the assumed diseased locus. Although being the most conservative approach for this case, it has the advantage of excluding the need for IBD-sharing estimation.*

We close this section by noting that  $\alpha(T; \hat{\theta})$  is a consistent estimator of  $\alpha(T; \theta)$  when  $n \rightarrow \infty$ .

## 6 Power Calculations

One way to compare the performance of different score functions in the sense of its ability to detect an actual (two-locus) disease is to compute the power (12). We perform power calculations under both simple and composite null hypotheses using the score functions defined in (18). Our aim is to try to get some insight regarding the choice of function and the effect of a possible gene-gene interaction.

### 6.1 Simple Null Hypothesis

We perform local power calculations, i.e.  $l_1$  and  $l_2$  fixed, using the simple null hypothesis (7). Explicitly, this is given by the matrix

$$H_0 : z = \begin{pmatrix} 0.0625 & 0.1250 & 0.0625 \\ 0.1250 & 0.2500 & 0.1250 \\ 0.0625 & 0.1250 & 0.0625 \end{pmatrix}, \quad (27)$$

corresponding to no genetic component and no gene-gene interaction with respect to the disease and loci  $l_1$  and  $l_2$ .

To calculate the power, we assume the presence of two distinct disease loci with equally strong (one-locus) genetic effects,

$$H_1 : z^1 = z^2 = (0.2, 0.4, 0.4), \quad (28)$$

and define three different versions of the alternative hypothesis

$$H_{1A} : z = \begin{pmatrix} 0.04 & 0.08 & 0.08 \\ 0.08 & 0.16 & 0.16 \\ 0.08 & 0.16 & 0.16 \end{pmatrix}, \quad (29)$$

$$H_{1B} : z = \begin{pmatrix} 0.10 & 0.05 & 0.05 \\ 0.05 & 0.25 & 0.10 \\ 0.05 & 0.10 & 0.25 \end{pmatrix}, \quad (30)$$

$$H_{1C} : z = \begin{pmatrix} 0.20 & 0 & 0 \\ 0 & 0.40 & 0 \\ 0 & 0 & 0.40 \end{pmatrix}. \quad (31)$$

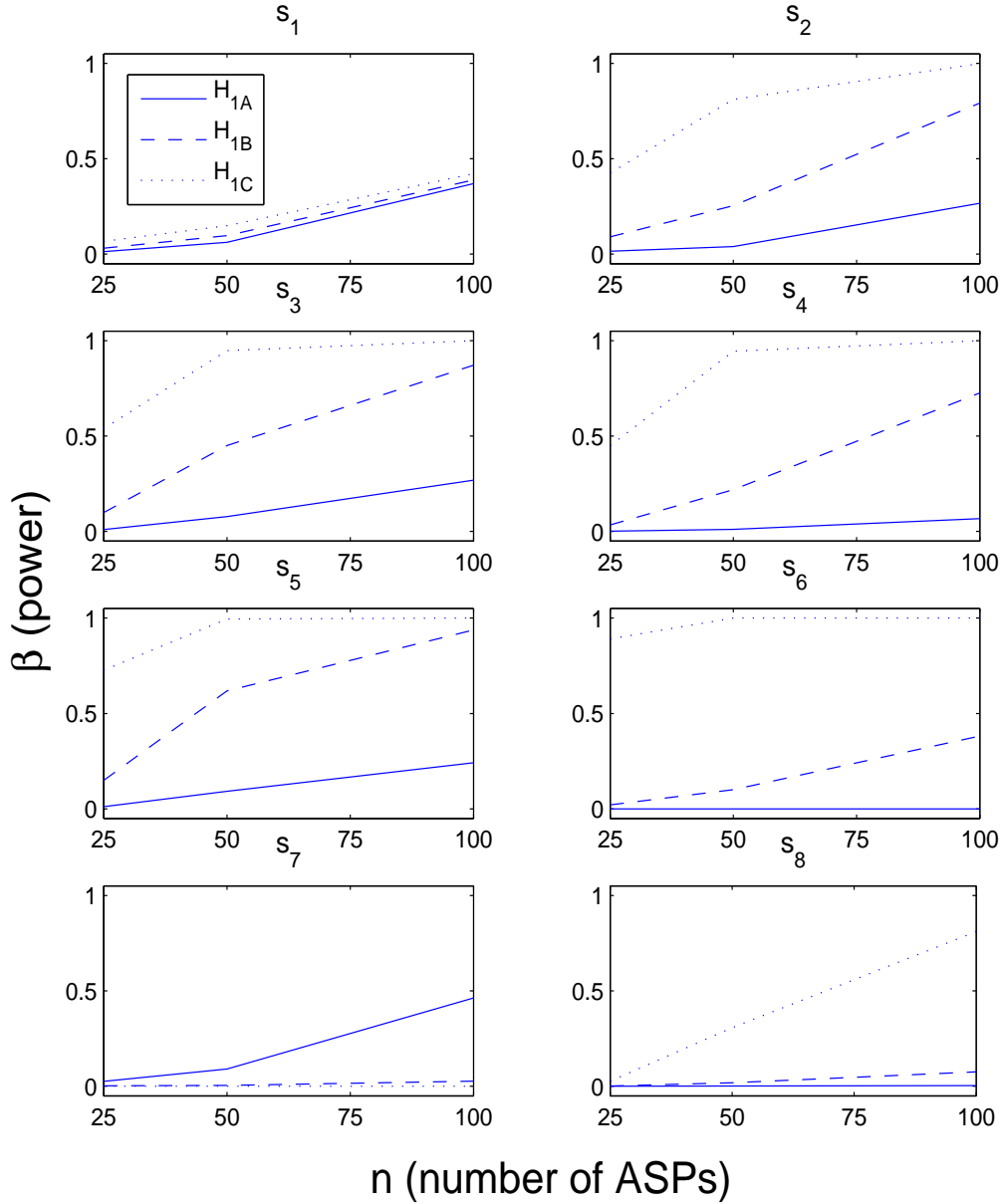


Figure 1: Local power calculations for the eight score functions in (18) using the simple null hypothesis, alternative hypotheses (29)-(31) and number of ASPs  $n=25$ , 50 or 100 respectively. Significance level  $\alpha = 0.00001$ . The number of simulations are  $J=500000$  (thresholds) and  $J=50000$  (power). Note: (i) The significance levels are only approximate due to the discrete NPL score distributions. (ii) The power is approximated by simulation for the unrestricted score functions and exact for the restricted score functions. The latter approach is based on the assumption of binomially distributed data, since then the ASPs are grouped according to their score (0 or 1).

All three hypothesis have marginal (one-locus) probabilities satisfying (28). The difference lies in the amount of IBD-sharing correlation (gene-gene interaction). The hypotheses  $H_{1A}$ ,  $H_{1B}$  and  $H_{1C}$  corresponds to no, some and full interaction respectively. If (31) holds true, it is only possible to jointly share exactly the same number of alleles IBD at both disease loci.

### 6.1.1 Results

We calculate the power with respect to the number of ASPs  $n$ . The results are displayed in Figure 1 and a complete numerical review is given in Table 9-10 of Appendix C.

For all score functions but  $S_7$ , power increases with increasing gene-gene interaction of the alternative.  $S_6$  performs best for strong interaction ( $H_{1C}$ ),  $S_5$  for moderately strong interaction ( $H_{1B}$ ) and  $S_7$  for no interaction ( $H_{1A}$ ). One may note that there is a natural ordering  $S_8 \rightarrow S_4 \rightarrow S_2 \rightarrow S_3 \rightarrow S_5$  with respect to score function structure, where  $S_5$  gives the highest and  $S_8$  the lowest weight to full IBD-sharing at both loci. Of these score functions  $S_8$  has the lowest and  $S_5$  the highest power under all three alternatives. The poor performance of  $S_7$  under  $H_{1B}$  and  $H_{1C}$  is explained by the probability of having a positive score (0.55 for  $H_{1B}$  and 0.40 for  $H_{1C}$ ) compared to the probability under the null hypothesis (0.4375).

### 6.1.2 Discussion

One problem is that the genetic disease model is not known in advance, making the choice of score function a priori difficult. Then trying to find a robust score function that behaves acceptable under a wide range of genetic disease models is attractive.

A possible approach is to use restricted score functions, either one robust or a few contrasting ones to cover the space of disease models. Restricted score functions are easily interpreted and have, given an appropriate choice, good performance. The following definition gives a suggestion, which often is close to optimal.

**Definition 4** *The best possible choice of restricted score function in terms of statistical power is called the **best restricted** score function.*

One situation when this kind of approach might be useful is when the search for a single disease-causing locus has failed and standard two-locus anal-

ysis assuming no gene-gene interaction gives negative answers. Assuming two jointly present disease loci, one source of the preceding failures may be the presence of IBD-sharing correlation. Using appropriate restricted score functions will then, in many cases, be helpful for discovering the disease loci.

**Note 15** For local tests the score function  $S(i, j) = \log(z_{ij1}/z_{ij0})$  is optimal, in the sense that it maximizes the power according to Neyman-Pearson's Lemma. Here we assume IBD-probabilities  $z_{ij0}$  and  $z_{ij1}$  corresponding to a simple null hypothesis and a fixed alternative respectively.

Given a restricted score function  $S$ , we refer to all  $(i, j)$  with  $S(i, j) = 1$  as the set of included cells. Note 15 suggests that the cells  $(i, j)$  should be included in decreasing order of  $z_{ij1}/z_{ij0}$ . If  $n$  is large, the optimal set of included cells can be found by normal approximation, as described in Appendix D.

**Example 5** Consider the standard null hypothesis in (27) and the alternative hypothesis  $H_{1x}$  based on two equally strong disease loci,  $z^1 = z^2 = (0.2, 0.5, 0.3)$ , with varying strength of the gene-gene interaction,

$$H_{1x} : z = \begin{pmatrix} 0.04 + x & 0.10 - x & 0.06 \\ 0.10 - x & 0.25 + x & 0.15 \\ 0.06 & 0.15 & 0.09 \end{pmatrix}, \quad -0.04 \leq x \leq 0.10. \quad (32)$$

We perform local power calculations searching for the best restricted score function, i.e. the optimal selection of positive weights, with respect to several versions of (32) and the results are given in Table 6.

Table 6: Local power calculations using 100 ASPs, significance level  $\alpha = 0.01$ .

$x$	Included cells: Best restricted	Power
0	$\{(2, 2), (1, 2), (2, 1)\}$	0.2355
0.01	$\{(2, 2), (1, 2), (2, 1)\}$	0.2355
0.02	$\{(2, 2), (1, 2), (2, 1), (1, 1)\}$	0.3018
0.05	$\{(2, 2), (1, 2), (2, 1), (1, 1)\}$	0.5454
0.10	$\{(2, 2), (1, 2), (2, 1), (1, 1)\}$	0.8935

For  $x = 0$  a sequential inclusion of cells  $(2, 2)$ ,  $(1, 2)$  and  $(2, 1)$  successively increases the power to 23.55%. This selection is optimal for  $x = 0.01$  as

well, even though  $z_{111} > z_{110}$ , since the increase is too low to positively affect the power ( $0.2331 < 0.2355$ ). Not surprisingly, for  $x \geq 0.02$  the additional inclusion of  $(1, 1)$  leads to a further increase of statistical power. The greatest consistent value  $x = 0.10$  then corresponds to a relatively strong power of 89.35%.

## 6.2 Composite Null Hypothesis

### 6.2.1 Considering Power

**Context** Firstly, we perform global power calculations through simulations, using a genome  $\Omega$  consisting of three chromosomes ( $C_1$ ,  $C_2$  and  $C_3$ ) of equal length ( $|C|=|C_1|=|C_2|=|C_3|=2$  Morgans) and the composite null hypothesis (9) with

$$A = \{(0.2, 0.4, 0.4), (0.25, 0.5, 0.25)\}.$$

Calculations are made with respect to four different score functions  $S_1$ ,  $S_2$ ,  $S_5$  and  $S_7$  in (18), three distinct significance levels and, in analogy with Section 6.1, using the three alternative hypotheses (29)-(31) of varying IBD-sharing correlation.

In order to easily compute the NPL thresholds corresponding to the significance levels, we use three simplifying assumptions regarding the real and estimated genetic models,  $\theta = (z, l_1, l_2)$  in Section 5.2.1 and  $\hat{\theta} = (\hat{z}, \hat{l}_1, l_2)$  in Section 5.2.2 respectively: (i) We consider  $\hat{z} = z = (0.2, 0.4, 0.4)^T * (0.25, 0.5, 0.25)$  as known. (ii) We consider  $l_1$  and  $l_2$  to be positioned *in the middle* of the chromosomes  $C_1$  and  $C_2$  respectively (i.e. at location  $|C|/2$ ). (iii) Given (ii), we consider  $\hat{l}_1 = l_1$  as known.

Moreover, we perform similar simulation analyses using the minor assumptional adjustment: (ii)' We consider  $l_1$  and  $l_2$  to be positioned at the same *random* location on  $C_1$  and  $C_2$  respectively (uniformly distributed over the interval  $[0, |C|]$ ; new generated common location for each simulation). This leads to estimates of the expected value of the power with respect to the common disease loci position.

**Results** The results are displayed in Figure 2-3 and a total numerical review is given in Table 11-12 of Appendix C. One may note that the calculations are performed only for the three stated alternatives but the results in

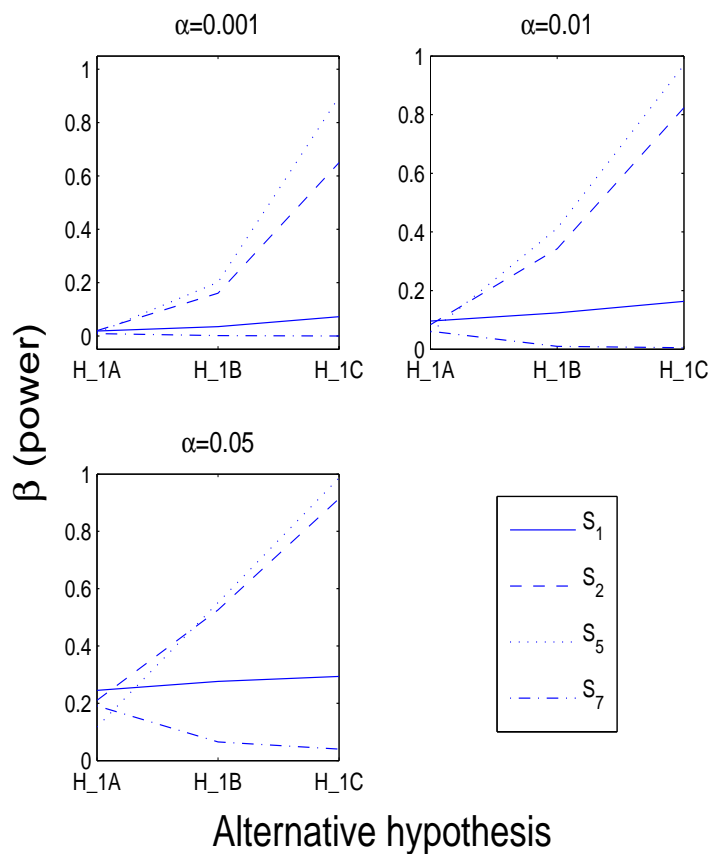


Figure 2: Global power calculations for the score functions  $S_1, S_2, S_5, S_7$  using 3 chromosomes of equal length 2 Morgans, a composite null hypothesis with one disease locus  $z^1=(0.2,0.4,0.4)$ , alternative hypotheses (29)-(31) and number of ASPs  $n=50$ . The significance levels are  $\alpha=0.001, 0.01$  and  $0.05$  respectively. The number of simulations is  $J=5000$  (thresholds) and  $J=2500$  (power).

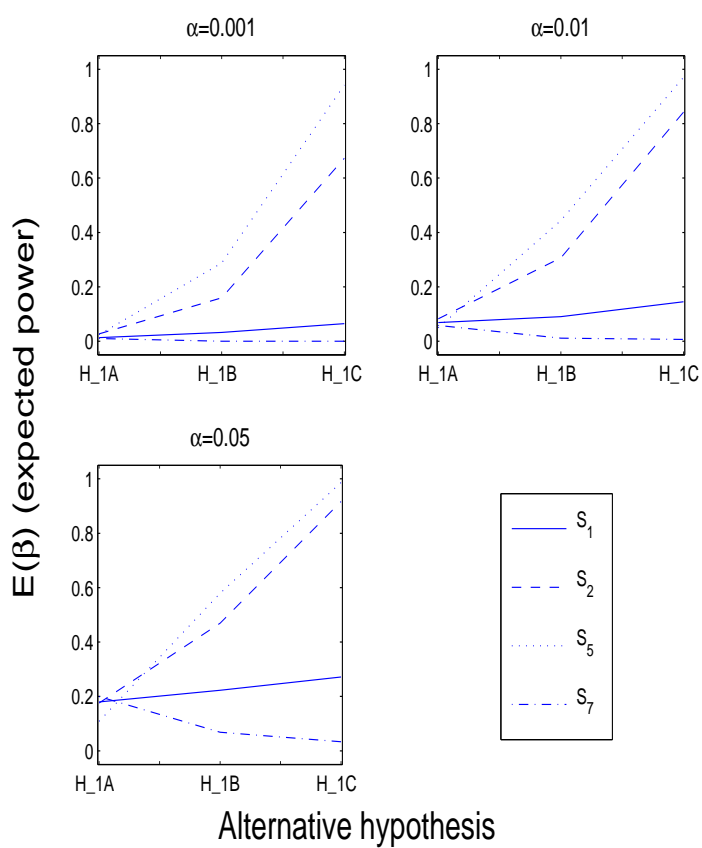


Figure 3: Global calculations of the expected value of the power with respect to the (random) disease loci positions. For further details cf. the caption of Figure 2.

Figure 2-3 may be interpreted as linear interpolations with respect to the set of alternatives

$$H_{1x} : z = \begin{pmatrix} 0.04 + 2x & 0.08 - x & 0.08 - x \\ 0.08 - x & 0.16 + 3x & 0.16 - 2x \\ 0.08 - x & 0.16 - 2x & 0.16 + 3x \end{pmatrix}, 0 \leq x \leq 0.08. \quad (33)$$

Considering  $S_1$ ,  $S_2$  and  $S_5$ , the power (as expected) improves in the direction  $S_1 \rightarrow S_2 \rightarrow S_5$  and  $H_{1A} \rightarrow H_{1B} \rightarrow H_{1C}$ , whereas the results for  $S_7$  is poor and is decreasing with increased IBD-correlation. This is consistent with the discussion in Section 6.1. Generally, the differences between the results when using the assumptions (ii) and (ii)' respectively are small.

### 6.2.2 Considering Significance Levels

**Context** Secondly, for all score functions in (18) we use the estimated significance levels  $\alpha_0$  under the simple null hypothesis (7) and compare it with the corresponding significance levels  $\alpha_1$  (9), using the same thresholds, under three distinct composite null hypotheses, see Figure 4. Here our genome consists of five chromosomes of equal length ( $|C|=2$  Morgans) and the location of the disease locus  $l_1$  is assumed to be in the middle of  $C_1$ . Similar simplifying assumptions as in Section 6.2.1 are used.

Moreover, we perform additional simulations assuming  $l_1$  to be positioned at a random location (uniformly distributed over the interval  $[0, |C|]$ ; new generated location for each simulation). This leads to estimates of the expected value of  $\alpha_1$  with respect to the disease locus position, cf. Figure 5.

**Results** The results show that the discrepancy between  $\alpha_0$  and  $\alpha_1$  might be severe, which shows that (7) and (9) give very different  $p$ -values under a strong one-locus genetic component. Hence, the inclusion of a disease locus in the null hypothesis leads to a considerable loss of power. Generally, the differences between the results using a disease locus of fixed versus random location are small.

**Note 16** *In most cases the significance level discrepancy is enlarged with increasing (disease locus) IBD-sharing. The only exception from this rule is the results for  $S_6$ . The reason for this behaviour is that this score function is not regular since (16) is not satisfied.*

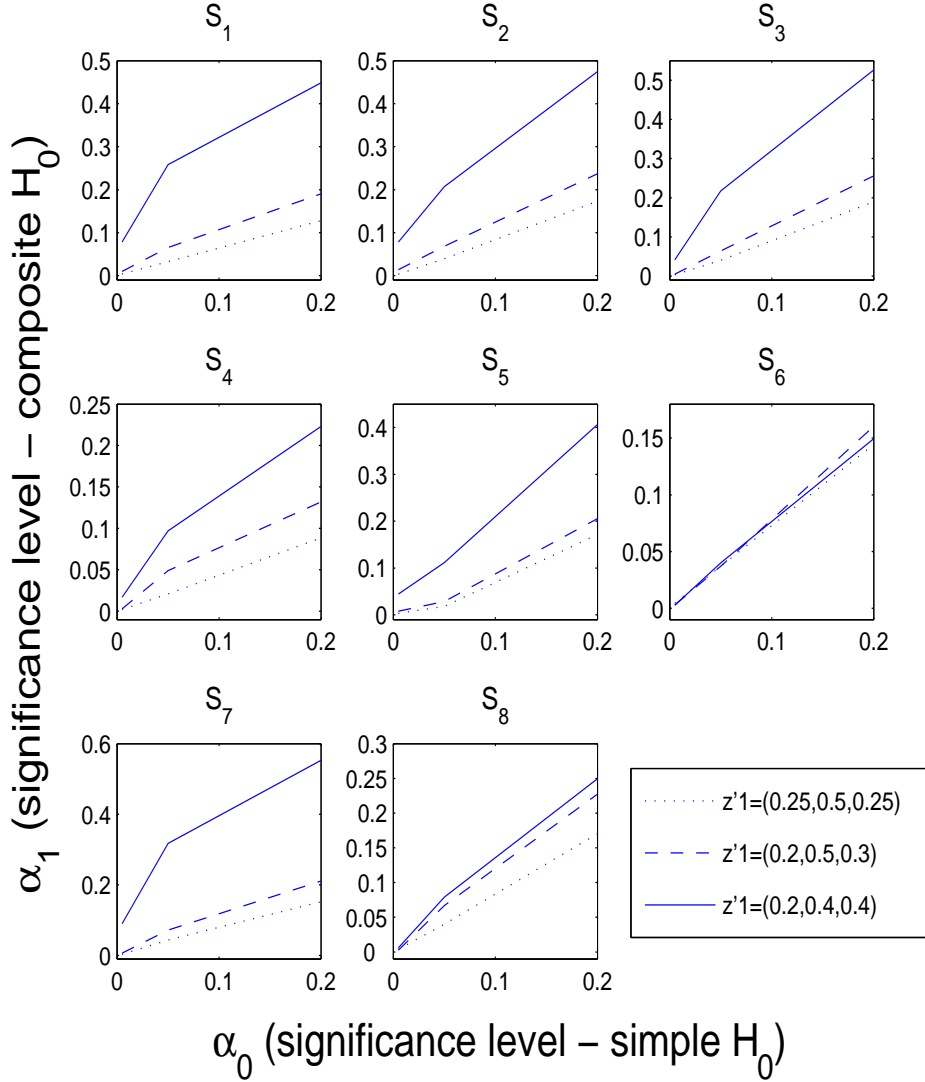


Figure 4: The significance levels using a simple null hypothesis versus the corresponding level for composite alternatives (9) with  $\theta = (z, l)$ ,  $z = (z^1)^T * (0.25, 0.5, 0.25)$ ,  $z^1 \in A = \{(0.2, 0.4, 0.4), (0.2, 0.5, 0.3), (0.25, 0.5, 0.25)\}$ ,  $l = (l_1, l_2)$  and  $l_1$  located in the middle of the first chromosome. Throughout, we use pedigree sets with  $n=100$  ASPs, score functions  $S_1$ - $S_8$  and a genome  $\Omega$  consisting of 5 chromosomes of equal length 2 Morgans. The significance levels of the simple null hypothesis are  $\alpha_0=(0.2,0.05,0.005)$ . The number of simulations is  $J=2000$ . Note: Using a grid of thresholds  $\mathbb{T}$ , we choose the threshold  $T$  as the smallest one giving rise to significance levels *less than or equal to* the appropriate  $\alpha_0$ . This explains why some values seem to be improperly small, e.g. that the dotted line above doesn't exactly follow the diagonal  $\alpha_0 = \alpha_1$ .

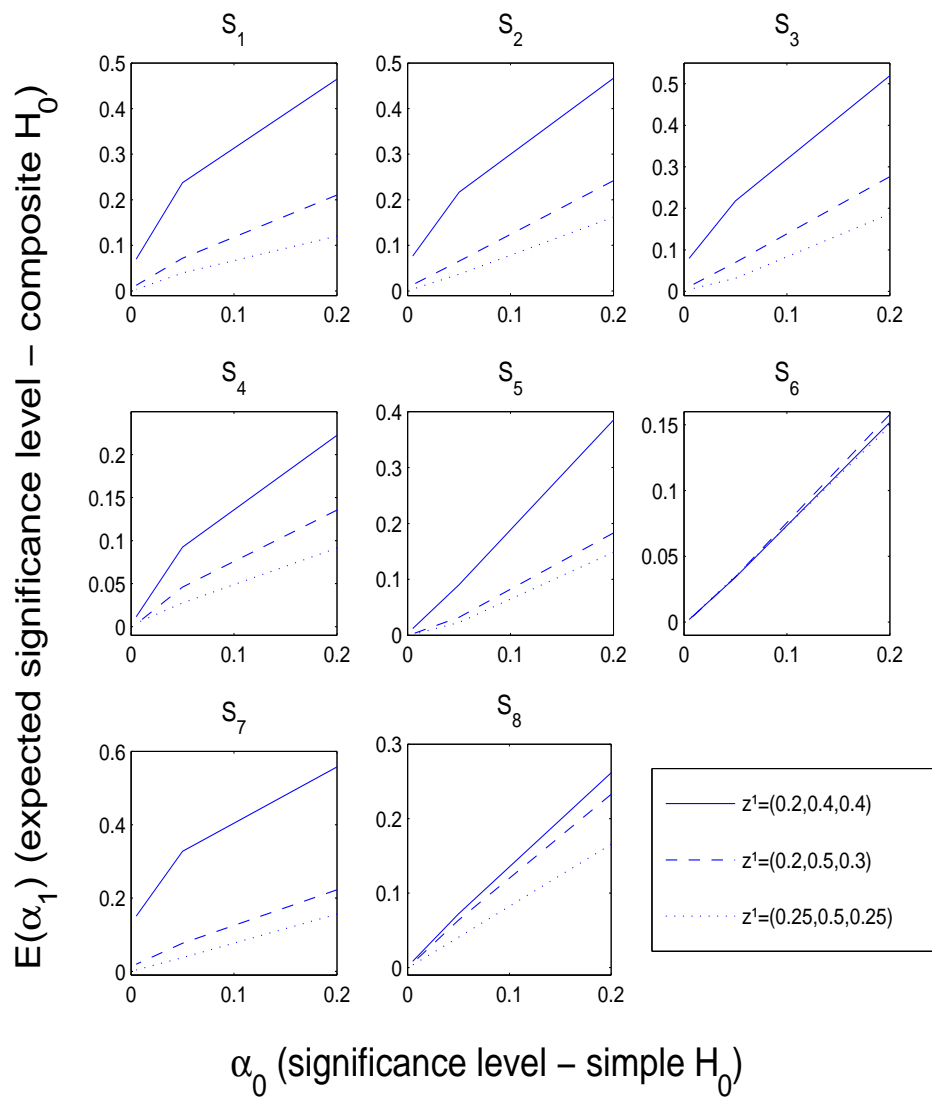


Figure 5: The significance levels using a simple null hypothesis versus the expected value of the corresponding level for composite alternatives (9) with respect to the uniformly random disease locus position. For further details cf. the caption of Figure 4.

## 7 Discussion

We have proposed an approach to simultaneous two-locus genome-wide (global) NPL-based search. It should be noted however that at present looking at a whole genome is very computer-intensive, although we believe that our method is feasible for looking for two-locus linkage and gene-gene interactions over a few chromosomes, where the candidate chromosomes are then suggested by other means.

The main topics of this article are: (i) Within a general framework for two-locus score functions, we are able to give conservative bounds on type-1 errors. In Theorem 1 we give asymptotic  $p$ -values for testing the composite null hypothesis  $H_0$  of no joint genetic effect against the alternative of a joint effect. Asymptotic or simulation-based approximations of these  $p$ -values are easy to implement and, moreover, applicable for a wide class of score functions. (ii) Comparisons between simple and composite null hypotheses. The choice of null hypothesis greatly influences the power of the tests. A simple null hypothesis leads to more powerful tests, but on the other hand the class of alternative hypotheses is so large that a rejected null hypothesis is less informative than for a composite null hypothesis. (iii) Investigations of the effect of gene-gene interaction. For many score functions, gene-gene interaction greatly increases power.

Moreover, we are aware of the fact that all our examples, for instance some cases based on (32)-(33), do not fulfil the two-locus IBD-sharing inequalities given in Bengtsson (2001). Though we believe that to choose somewhat exaggerated examples might be justified according to our aim of clearly and simple presenting the corresponding principles and implications.

Finally, a few words on some alternative methods, of various degree of similarity to our approach, which are plausible to use within the same context:

- I** Replace the estimates  $\hat{z}$  at  $\hat{l}_1$  by the corresponding pedigree-specific NPL scores  $Z_i(\hat{l}_1)$  ( $i = 1, 2, \dots, n$ ). In practice, this reflects a score peak of fixed rather than of random height and might be considered to be an instance of a conditional NPL analysis approach.
- II** Knowledge (or assumption of) a one-locus disease model (i.e. the penetrance vector, disease allele frequency and disease locus position) makes it possible to calculate  $\theta$  under  $H_0$ , avoiding the need to estimate, or to

choose a least favourable, genetic model when calculating significance levels.

- III** A conceptually different approach would be to use the two-locus version of the MLS method (Cordell et al., 1995) to test similar null hypotheses.

## 8 Acknowledgements

This research has been sponsored by the Swedish Research Council and the Wallenberg Laboratory (Department of Endocrinology, Malmö University Hospital, Lund University, Malmö, Sweden).

We would like to thank Professor Ola Hössjer for major comments and suggestions greatly improving the outcome of this project.

## References

- Ängquist, L. and Hössjer, O. (2004). Using importance sampling to improve simulation in linkage analysis. *Statistical Applications in Genetics and Molecular Biology*, 3(1:5). (Electronic journal, 24 pages)
- Ängquist, L. and Hössjer, O. (2005a). Improving the calculation of statistical significance in genome-wide scans. *Biostatistics*, 6(4), 520–538.
- Ängquist, L. and Hössjer, O. (2005b). *Strategies for conditional two-locus nonparametric linkage analysis* (Tech. Rep.). Lund: Department of Mathematical Statistics, Lund University. (Work in progress)
- Bengtsson, O. (2001). *Two-locus affected sib-pair identity by descent probabilities: Constraints, parameterisation and estimation* (Licentiate thesis). Göteborg: Department of Mathematical Statistics, Chalmers University of Technology, Göteborg University.
- Boehnke, M. (1986). Estimating the power of a proposed linkage study: A practical computer simulation approach. *American Journal of Human Genetics*, 39, 513–527.

- Chiu, Y. F. and Liang, K. Y. (2004). Conditional multipoint linkage analysis using affected sib pairs: An alternative approach. *Genetic Epidemiology*, *26*, 108–115.
- Cordell, H. J. (2003). Affected sib-pair data can be used to distinguish two-locus heterogeneity from two-locus epistasis. *American Journal of Human Genetics*, *73*, 1468–1471. (Discussion of article by Vieland and Huang, 2003)
- Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y. and Farrall, M. (1995). Two-locus maximum lod score analysis of a multifactorial trait: Joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. *American Journal of Human Genetics*, *57*, 920–934.
- Cordell, H. J., Wedig, G. C., Jacobs, K. B. and Elston, R. C. (2000). Multi-locus linkage tests based on affected relative pairs. *American Journal of Human Genetics*, *66*, 1273–1286.
- Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I. and Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics*, *21*, 213–215.
- Culverhouse, R., Klein, T. and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiology*, *27*, 141–152.
- Davis, S. and Weeks, D. E. (1997). Comparisons of nonparametric statistics for detection of linkage in nuclear families: Single-marker evaluations. *American Journal of Human Genetics*, *61*, 1431–1444.
- Doerge, R. W. and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics*, *142*, 285–294.
- Dudoit, S. and Speed, T. P. (1998). *Triangle constraints for sib-pair identity by descent probabilities under a general model for disease susceptibility* (Tech. Rep. No. 527). Department of Statistics, University of California, Berkeley.

- Dupuis, J., Brown, P. O. and Siegmund, D. (1995). Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics*, *140*, 843–856.
- Farrall, M. (1997). Affected sibpair linkage tests for multiple linked susceptibility genes. *Genetic Epidemiology*, *14*, 103–115.
- Farrall, M. (2003). Reports of the death of the epistasis model are greatly exaggerated. *American Journal of Human Genetics*, *73*, 1467–1468. (Discussion of article by Vieland and Huang, 2003)
- Feingold, E., Brown, P. O. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *American Journal of Human Genetics*, *53*, 234–251.
- Feingold, E., Song, K. K. and Weeks, D. E. (2000). Comparisons of allelesharing statistics for general pedigrees. *Genetic Epidemiology*, *19*, 92–98. (Supplement 1)
- Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, *4*, 701–709.
- Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics*, *52*, 362–374.
- Holmans, P. (2002). Detecting gene-gene interactions using affected sib pair analysis with covariates. *Human Heredity*, *53*, 92–102.
- Hössjer, O. (2003a). Assessing accuracy in linkage analysis by means of confidence regions. *Genetic Epidemiology*, *25*, 59–72.
- Hössjer, O. (2003b). Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Annals of Statistics*, *31*(4), 1075–1109.
- Hössjer, O. (2005a). Conditional likelihood score functions for mixed models in linkage analysis. *Biostatistics*, *6*(2), 313–332.
- Hössjer, O. (2005b). Information and effective number of meioses in linkage analysis. *Journal of Mathematical Biology*, *50*(2), 208–232.

- Hössjer, O. (2005c). Spectral decomposition of score functions in linkage analysis. *Bernoulli*. (To appear.)
- Kämpe, M. (2001). *Two-locus nonparametric linkage analysis for complex diseases* (Master's thesis No. 2001:E4). Lund: Lund Institute of Technology, Lund University.
- Knapp, M., Seuchter, S. A. and Baur, M. (1994). Two-locus disease models with two marker loci: The power of affected sib-pair tests. *American Journal of Human Genetics*, *55*, 1030–1041.
- Kong, A. and Cox, N. (1997). Allele-sharing models: Lod scores and accurate linkage tests. *American Journal of Human Genetics*, *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. *American Journal of Human Genetics*, *55*, 1347–1363.
- Kruglyak, L. and Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, *57*, 439–454.
- Lander, E. S. and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, *11*, 241–247.
- Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, *50*, 334–349.
- Lucek, P., Hanke, J., Reich, J., Solla, S. A. and Ott, J. (1998). Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. *Human Heredity*, *48*, 275–284.
- MacLean, C. J., Sham, P. C. and Kendler, K. S. (1993). Joint linkage of multiple loci for a complex disorder. *American Journal of Human Genetics*, *53*, 353–366.
- McPeck, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology*, *16*, 225–249.

- Nilsson, S. (2001). *Which genes are involved? - statistical planning and analysis of human genetic samples* (Doctoral thesis No. Ny serie nr 1744). Göteborg: Department of Mathematical Statistics, Chalmers University of Technology and Göteborg University.
- Ott, J. (1989). Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 86(11), 4175–4178.
- Ott, J. (1999). *Analysis of human genetic linkage* (Third ed.). New York: The John Hopkins University Press.
- Ott, J. and Hoh, J. (2000). Statistical approaches to gene mapping. *American Journal of Human Genetics*, 67, 289–294.
- Purcell, S. and Sham, P. C. (2004). Epistasis in quantitative trait locus linkage analysis: Interactions or main effect? *Behavior Genetics*, 34(2), 143–152.
- Risch, N. (1990). Linkage strategies for genetically complex traits: I. multi-locus models. *American Journal of Human Genetics*, 46, 222–228.
- Sengul, H., Weeks, D. E. and Feingold, E. (2001). A survey of affected-sibship statistics for nonparametric linkage analysis. *American Journal of Human Genetics*, 69, 179–190.
- Sham, P. (1998). *Statistics in human genetics*. London: Arnold Applications of Statistics.
- Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika*, 91, 785–800.
- Song, K. K., Weeks, D. E., Sobel, E. and Feingold, E. (2004). Efficient simulation of P values for linkage analysis. *Genetic Epidemiology*, 26, 88–96.
- Strachan, T. and Read, A. P. (2003). *Human molecular genetics* (Third ed.). London and New York: Garland Science.
- Strauch, K., Fimmers, R., Kurz, T., Baur, M. P. and Wienker, T. F. (2003). How to model a complex trait: 2. analysis with two disease loci. *Human Heredity*, 56, 200–211.

- Strauch, K., Fimmers, R., Kurz, T., Deichmann, K. A., Wienker, T. F. and Baur, M. P. (2000). Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: Application to mite sensitization. *American Journal of Human Genetics*, *66*, 1945–1957.
- Suarez, B. K. (1978). The affected sib-pair IBD distribution for HLA-linked disease susceptibility genes. *Tissue Antigens*, *12*, 87–93.
- Tang, H. K. and Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics*, *2*, 147–162.
- Tang, H. K. and Siegmund, D. (2002). Mapping multiple genes for quantitative or complex traits. *Genetic Epidemiology*, *22*, 313–327.
- Teng, J. and Siegmund, D. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics*, *54*, 1247–1265.
- Terwilliger, J. D., Speer, M. and Ott, J. (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genetic Epidemiology*, *10*, 217–224.
- Tiwari, H. K. and Elston, R. C. (1998). Restrictions on components of variance for epistatic models. *Theoretical Population Biology*, *54*, 161–174.
- Vieland, V. J. and Huang, J. (2003). Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected sib-pair data. *American Journal of Human Genetics*, *73*, 223–232.
- Whittemore, A. S. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics*, *50*, 118–127.
- Zinn-Justin, A. and Abel, L. (1998). Two-locus developments of the weighted pairwise correlation method for linkage analysis. *Genetic Epidemiology*, *15*, 491–510.

## A One-Locus Allele-Sharing Probabilities

**One-Locus Case** We will now give some explicit standard results concerning the probabilities appearing in (1). Similar formulas are given in e.g. Suarez (1978) and Nilsson (2001). In Table 7 the six distinct genotype configurations are presented and the (a priori) probabilities of these parental configurations (PG) and the probability of an ASP given such SGs are displayed. Under the assumptions of random mating and Hardy-Weinberg equilibrium the former only depends on the disease allele frequency, whereas the latter is a function of the disease penetrances only.

Table 7: Number, structure, a priori probabilities and the probability of an ASP with respect to the possible distinct genotype configurations.

No.	$PG$ or $SG$	$P(PG)$	$P(ASP SG)$
1	DD-DD	$p^4$	$f_2^2$
2	DD-Dd	$4p^3(1-p)$	$f_1 f_2$
3	Dd-Dd	$4p^2(1-p)^2$	$f_1^2$
4	DD-dd	$2p^2(1-p)^2$	$f_0 f_2$
5	Dd-dd	$4p(1-p)^3$	$f_0 f_1$
6	dd-dd	$(1-p)^4$	$f_0^2$

Next, in Table 8 we present the probabilities for the SGs conditional on the PGs. Obviously, not all the joint SG-PG outcomes are theoretically consistent, i.e. some outcomes are assigned a zero probability, forcing the corresponding conditional probabilities to equal 0. Moreover, the corresponding IBD-probability vector is given for all possible genotype combinations.

**Two-Locus Case** We restrict ourselves to making some remarks on the necessary generalizations with respect to the one-locus case. Comparing with Table 7 this includes: (i) The number of distinct configurations increases to  $6^2 = 36$ . (ii) The probabilities  $P(PG)$  depend on two disease allele frequencies  $p_1$  and  $p_2$ , corresponding to  $l_1$  and  $l_2$  respectively. (iii) The conditional probabilities  $P(ASP|SG)$  depend on the two-dimensional penetrances in (3).

Regarding Table 8 the updating implies: (i) The number of possible (grouped) SGs for each PG configuration ranges from 1 to 36. (ii) The positive probabilities  $P(SG|PG)$  are on the form  $x/256$ ,  $x \in \{1, 2, \dots, 256\}$ ,

Table 8: Probabilities of ( $IBD = 0, IBD = 1, IBD = 2$ ) given the jointly possible combinations of parental genotypes (PG) and sib-pair genotypes (SG).

$PG$	Possible $SGs$	$P(SG PG)$	$P(IBD PG, SG)$
1	DD-DD	1	$[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$
2	DD-DD	$\frac{1}{4}$	$[0, \frac{1}{2}, \frac{1}{2}]$
	Dd-DD	$\frac{1}{2}$	$[\frac{1}{2}, \frac{1}{2}, 0]$
	Dd-Dd	$\frac{1}{4}$	$[0, \frac{1}{2}, \frac{1}{2}]$
3	DD-DD	$\frac{1}{16}$	$[0, 0, 1]$
	Dd-DD	$\frac{4}{16}$	$[0, 1, 0]$
	Dd-Dd	$\frac{4}{16}$	$[\frac{1}{2}, 0, \frac{1}{2}]$
	DD-dd	$\frac{2}{16}$	$[1, 0, 0]$
	Dd-dd	$\frac{4}{16}$	$[0, 1, 0]$
	dd-dd	$\frac{1}{16}$	$[0, 0, 1]$
4	Dd-Dd	1	$[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$
5	Dd-Dd	$\frac{1}{4}$	$[0, \frac{1}{2}, \frac{1}{2}]$
	Dd-dd	$\frac{1}{2}$	$[\frac{1}{2}, \frac{1}{2}, 0]$
	dd-dd	$\frac{1}{4}$	$[0, \frac{1}{2}, \frac{1}{2}]$
6	dd-dd	1	$[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$

since the number of nongrouped configurations is 256. (iii) The conditional IBD-probabilities  $P(IBD|PG, SG)$  are given as 3x3 matrices rather than as vectors.

## B Proof of Theorem 1

Here we prove the statements (22)-(23) involving the least favourable distribution in  $H_0$  and its significance level for the global test (20).

**Proof:** Firstly, note that the validity of the proposition regarding the least favourable distribution ( $z_a$  or  $z_b$ ) follows from the regularity of the score function and that the (one-locus)  $p$ -value is an increasing function of the chromosome length.

Secondly, define  $s_i = (S(2, i) - \mu)/\sigma$  for  $i = 0, 1, 2$ . Then, if  $|C_1| \leq |C_2|$ , under the least favourable distribution  $z_a$  at locus  $l_1 \in C_1$ ,  $x \mapsto Z(l_1, x)$  behaves as a one-locus NPL process on  $C_2$ , with standardized score function  $S$  satisfying  $S(i) = s_i$ ,  $i = 0, 1, 2$ . Since the score function is regular

$$\max_{x_1 \in C_1, x_2 \in C_2} Z(x_1, x_2) = \max_{x_2 \in C_2} Z(l_1, x_2) \quad (34)$$

under the least favourable distribution. Put  $U_n(x) = Z(l_1, x)$ . Then from results in Hössjer (2005c) one finds that the covariance function of  $U_n$  satisfies

$$r_{U_n}(t) = (s_2 - s_0)^2 \frac{1}{8} \exp(-4|t|) + (s_2 - 2s_1 + s_0)^2 \frac{1}{16} \exp(-8|t|) \quad (35)$$

independently of  $n$  under the least favourable distribution. As  $n \rightarrow \infty$ ,  $U_n$  converges weakly to a Gaussian limit process  $U$  with covariance function (35), see Hössjer (2005c) for details. This limit process is a mixture of two Ornstein-Uhlenbeck processes. By (22), (34), (35) and the Continuous Mapping Theorem we obtain (23).

Similarly, if  $|C_1| \geq |C_2|$ , under the least favourable distribution  $z_b$  at locus  $l_2 \in C_2$ ,  $x \mapsto Z(x, l_2)$  behaves like a one-locus NPL process on  $C_1$  with the same score function  $S$ . The rest of the proof is analogous to the case  $|C_1| \leq |C_2|$ . ■

**Note 17** From (35) we also get the crossover rate

$$\rho = -r'_U(0)/2 = \frac{1}{4}(s_2 - s_0)^2 + \frac{1}{4}(s_2 - 2s_1 + s_0)^2 \quad (36)$$

of the limit process  $U$ , which is used to approximate the global significance level (cf. Lander and Kruglyak, 1995).

## C Detailed Power Calculation Results

Here we give a complete presentation of the power simulation results previously displayed in Figure 1 (Table 9-10) and Figure 2-3 (Table 11-12).

Table 9: Local power calculations using 25, 50 or 100 ASPs, a simple null hypothesis under three alternative hypotheses and score functions  $S_1 - S_4$ . Significance level  $\alpha = 0.00001$  (cf. Figure 1).

		$S_1$	$S_2$	$S_3$	$S_4$
$H_{1A}$ :	$N = 100$	0.3695	0.2663	0.2689	0.0677
	$N = 50$	0.0624	0.0397	0.0773	0.0106
	$N = 25$	0.0128	0.0149	0.0102	0.00164
$H_{1B}$ :	$N = 100$	0.3887	0.7916	0.8709	0.7261
	$N = 50$	0.0965	0.2535	0.4496	0.2192
	$N = 25$	0.0307	0.0900	0.0993	0.0348
$H_{1C}$ :	$N = 100$	0.4211	0.9983	0.9998	0.9999
	$N = 50$	0.1491	0.8112	0.9477	0.9449
	$N = 25$	0.0676	0.4253	0.5393	0.4494

Table 10: Local power calculations using 25, 50 or 100 ASPs, a simple null hypothesis under three alternative hypotheses and score functions  $S_5 - S_8$ . Significance level  $\alpha = 0.00001$  (cf. Figure 1).

		$S_5$	$S_6$	$S_7$	$S_8$
$H_{1A}$ :	$N = 100$	0.2424	0.0000270	0.4624	0.00368
	$N = 50$	0.0929	0.0000426	0.0903	0.00159
	$N = 25$	0.0121	0.0000480	0.0255	0.000215
$H_{1B}$ :	$N = 100$	0.9370	0.3822	0.0272	0.0755
	$N = 50$	0.6184	0.1013	0.00449	0.0183
	$N = 25$	0.1494	0.0216	0.00231	0.00157
$H_{1C}$ :	$N = 100$	0.999998	0.9999999992	0.000000392	0.8109
	$N = 50$	0.9943	0.9997	0.00000113	0.3073
	$N = 25$	0.7265	0.8909	0.00000816	0.0274

Table 11: Global power calculations using 50 ASPs, a composite null hypothesis under three alternative hypotheses and score functions  $S_1$ ,  $S_2$ ,  $S_5$  and  $S_7$ . The disease loci  $l_1$  and  $l_2$  are located in the middle of the first chromosome  $C_1$  and second chromosome  $C_2$  respectively.. Significance level  $\alpha=0.001$ ,  $0.01$  or  $0.05$  respectively (cf. Figure 2).

N=50		$S_1$	$S_2$	$S_5$	$S_7$
$H_{1A}$ :	$\alpha=0.001$	0.0184	0.0196	0.0116	0.0088
	$\alpha=0.01$	0.0964	0.0836	0.0580	0.0616
	$\alpha=0.05$	0.2448	0.2100	0.1168	0.1908
$H_{1B}$ :	$\alpha=0.001$	0.0348	0.1604	0.2016	0.0008
	$\alpha=0.01$	0.1240	0.3424	0.4124	0.0100
	$\alpha=0.05$	0.2760	0.5256	0.5516	0.0648
$H_{1C}$ :	$\alpha=0.001$	0.0720	0.6500	0.8968	0.0000
	$\alpha=0.01$	0.1632	0.8240	0.9692	0.0048
	$\alpha=0.05$	0.2936	0.9140	0.9864	0.0400

Table 12: Global calculations of the expected value of the power with respect to the (randomly located) disease loci positions (cf. Figure 3). For further details cf. the caption of Table 11.

N=50		$S_1$	$S_2$	$S_5$	$S_7$
$H_{1A}$ :	$\alpha=0.001$	0.0132	0.0268	0.0240	0.0100
	$\alpha=0.01$	0.0684	0.0824	0.0492	0.0580
	$\alpha=0.05$	0.1796	0.1760	0.1088	0.1996
$H_{1B}$ :	$\alpha=0.001$	0.0320	0.1592	0.2876	0.0000
	$\alpha=0.01$	0.0900	0.3064	0.4436	0.0108
	$\alpha=0.05$	0.2228	0.4692	0.5792	0.0684
$H_{1C}$ :	$\alpha=0.001$	0.0648	0.6732	0.9424	0.0000
	$\alpha=0.01$	0.1448	0.8424	0.9704	0.0064
	$\alpha=0.05$	0.2712	0.9164	0.9864	0.0332

## D Criteria for Further Inclusion of Cells

Let  $A$  be the collection of cells  $(i, j)$  with  $S(i, j) = 1$  and define

$$z_k = z_k(A) = \sum_{i,j \in A} z_{ijk} = P\left(\left( \text{IBD}(l_1), \text{IBD}(l_2) \right) \in A \mid H_k\right), \quad k = 0, 1.$$

According to the discussion in Section 6.1, choose  $A = \{(i, j); z_{ij1}/z_{ij0} \geq t\}$  for some (variable) threshold  $t$ . Let  $Z^n$  denote the nonstandardized NPL score based on  $n$  ASPs. Assume that  $n$  is large enough for a normal approximation of the binomial  $\text{Bin}(n, p)$ -distribution ( $p \in [z_0, z_1]$ ). Then, the null hypothesis is rejected for a pointwise test with significance level  $\alpha$  if

$$Y = \frac{Z^n - nz_1}{\sqrt{nz_1(1-z_1)}} \geq \lambda_\alpha \sqrt{\frac{z_0(1-z_0)}{z_1(1-z_1)}} + \sqrt{n} \frac{z_0 - z_1}{\sqrt{z_1(1-z_1)}}, \quad (37)$$

where  $Y$  has approximately a standard normal distribution under  $H_1$  and  $\lambda_\alpha = \Phi^{-1}(1 - \alpha)$  is the standard normal quantile corresponding to the significance level  $\alpha$ . Hence the power is  $1 - \Phi(y)$ , where  $y$  is the right-hand side of (37). The optimal choice of  $A$ , in terms of pointwise power, is found by minimizing  $y$ .