

A Discussion on Evidential  
Interpretation of Two-Locus  
Nonparametric Linkage Analysis

Lars Ängquist<sup>1</sup>

12th December 2006

<sup>1</sup>PhD Student at Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University, Lund, Sweden.

## **Abstract**

In this article we outline and discuss evidential aspects and interpretations of two-locus, conditional and unconditional, nonparametric linkage (NPL) analysis. Initially we present a quite general multilocus NPL context and then restrict ourselves to two-locus cases leading to and focusing on methods and corresponding significance calculations as given in Ängquist and Hössjer (2006). In the last part of the paper we discuss related evidential properties of positive findings for the discussed kinds of two-locus NPL analysis.

**Key words:** Nonparametric linkage analysis, conditional and unconditional two-locus linkage analysis, conditioning loci, multilocus generalization, significance calculations, evidential interpretation.

## **Contents**

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Nonparametric Linkage Analysis</b>	<b>3</b>
2.1	General Multilocus Case . . . . .	4
2.2	Two-Locus NPL Analysis . . . . .	6
2.2.1	Unconditional Case . . . . .	6
2.2.2	Conditional Case . . . . .	6
2.2.3	Bibliographic Notes . . . . .	7
<b>3</b>	<b>Significance Calculations</b>	<b>8</b>
3.1	General Case . . . . .	8
3.2	Statistical Significance in NPL Analysis . . . . .	9
3.2.1	Unconditional Case . . . . .	9
3.2.2	Conditional Case . . . . .	10
3.2.3	Some Notes on Significance Calculations . . . . .	11
<b>4</b>	<b>Evidential Aspects of Positive Findings</b>	<b>11</b>
4.1	Unconditional Case . . . . .	11
4.2	Conditional Case: A Single Known Conditioning Locus . . . . .	12
4.3	Conditional Case: A Single Estimated Conditioning Locus . . . . .	12
4.4	Conditional Case: A Random Number of Estimated Conditioning Loci . . . . .	12
	<b>References</b>	<b>13</b>

## 1 Introduction

In this article we discuss the sometimes forgotten area of evidential interpretation of significant, or positive, findings originating from the traditional statistical testing paradigm. In some cases these views will contrast each other and we will discuss such issues in the context of genome-wide two-locus, both unconditional and conditional, nonparametric linkage (NPL) analysis. Here genome-wide test statistics are used as a way of dealing with inherent multiple testing.

In *Section 2* we present some basic theory and introduce basic notation with respect to NPL analysis in the multilocus as well as in the two-locus cases. Later, in *Section 3* significance calculations are outlined for both general statistical and NPL analyses. Finally, in *Section 4* we discuss the evidential aspects of positive findings in different kinds of two-locus NPL analysis.

## 2 Nonparametric Linkage Analysis

For a single pedigree, the genetic inheritance pattern at a locus  $x$  is determined by means of the *inheritance vector* (Donnelly, 1983),

$$v(x) = [p_1(x), m_1(x), p_2(x), m_2(x), \dots, p_{(n-f)}(x), m_{(n-f)}(x)], \quad (1)$$

where  $n$  is the number of individuals,  $f$  the number of founders and  $n - f$  the number of nonfounders in the pedigree. The length of  $v$  in (1) equals the number of meioses  $m = 2(n - f)$  and, moreover,  $p_i(x)$  and  $m_i(x)$  equal 0/1 if the  $i^{\text{th}}$  nonfounder's paternal and maternal allele respectively, at locus  $x$ , originate from a grandfather/grandmother.

So called *score functions* are defined with respect to inheritance vectors. They are introduced in order to quantify allele-sharing stratified to be observed within phenotype-groups (affecteds and unaffecteds).<sup>1</sup> For specific discussions related to score functions see, for instance, Whittemore and Halpern (1994), McPeck (1999), Hössjer (2003), Lange and Lange (2004), Hössjer (2005a, 2005b), Ängquist (2006b) and Ängquist and Hössjer (2006).

---

<sup>1</sup>We will assume, as mostly is the case, that allele-sharing corresponds to the concept of alleles being *identical-by-descent (IBD)*.

## 2.1 General Multilocus Case

In the general multilocus case the *unconditional* score function is defined as,

$$S(w_1, w_2, \dots, w_{|l|}); \quad \forall w_i \in \mathbb{V}, \quad (2)$$

where  $|l|$  is the number of underlying loci and  $\mathbb{V}$  is the set including all possible inheritance vectors.<sup>2</sup> Usually  $S$  in (2) is assumed to be in *standardized* form, i.e. that  $E(S) = 0$  and  $V(S) = 1$  under the null hypothesis of no linkage and acceptance of the Mendelian principle of random mating.

To relate the score function  $S$  to marker data MD and to make the underlying loci explicit one may define the *pedigree-specific NPL score* at locus  $x$  as,

$$\begin{aligned} Z(x_1, x_2, \dots, x_{|l|}) &= E [S(v(x_1, x_2, \dots, x_{|l|}))] \\ &= \sum_{w_1, w_2, \dots, w_{|l|}} P_{v(x_1, x_2, \dots, x_{|l|})}(w_1, w_2, \dots, w_{|l|}) S(w_1, w_2, \dots, w_{|l|}), \end{aligned} \quad (3)$$

where

$$\begin{aligned} P_{v(x_1, x_2, \dots, x_{|l|})}(w_1, w_2, \dots, w_{|l|}) \\ = P [v(x_1) = w_1, v(x_2) = w_2, \dots, v(x_{|l|}) = w_{|l|} \mid \text{MD}] \end{aligned} \quad (4)$$

is the conditional probability of joint inheritance vectors given marker data. Letting  $c(x)$  equal the chromosome where  $x$  is located, usually one assumes that

$$\forall i, j : i \neq j \Rightarrow c(x_i) \neq c(x_j).$$

Since inheritance at unlinked loci are independent, where it is assumed that loci located on distinct chromosomes (nonsyntenic loci) are unlinked, it follows that hence (4) simplifies to,

$$P_{v(x_1, x_2, \dots, x_{|l|})}(w_1, w_2, \dots, w_{|l|}) = P_{v(x_1)}(w_1) P_{v(x_2)}(w_2) \cdots P_{v(x_{|l|})}(w_{|l|}). \quad (5)$$

For a pedigree set consisting of  $N$  pedigrees the (total) *NPL score* (Kruglyak et al., 1996) is then defined as a weighted linear combination of the pedigree-specific scores,

$$Z(x_1, x_2, \dots, x_{|l|}) = \sum_{k=1}^N \gamma_k Z_k(x_1, x_2, \dots, x_{|l|}), \quad (6)$$

---

<sup>2</sup>We might call this the  $|l|$ -multilocus analysis approach.

where  $Z_k$  and  $\gamma_k$  denotes the  $k^{\text{th}}$  pedigree-specific score and its weight respectively. The latter is assumed to obey the constraint,

$$\sum_{k=1}^N \gamma_k^2 = 1, \quad (7)$$

in order to preserve the standardized properties of the score.

In a *genome-wide* context one may define the maximum of the *NPL score process* along genome  $\Omega$ ,

$$Z_{\max} = \sup_{x_1, x_2, \dots, x_{|l|} \in \Omega} Z(x_1, x_2, \dots, x_{|l|}), \quad (8)$$

to function as a test statistic.

Fixing information at *conditioning loci*  $x_{i_1}, x_{i_2}, \dots, x_{i_{l'}}$  where,

$$1 \leq i_1 < i_2 < \dots < i_{l'} \leq |l|,$$

on corresponding chromosomes  $c(x_{i_1}), c(x_{i_2}), \dots, c(x_{i_{l'}})$  and using (3) and (6) scanning through, i.e. calculating NPL scores on, the complementary part of the genome,

$$\Omega_{c(x_{i_1}), c(x_{i_2}), \dots, c(x_{i_{l'}})} = \Omega \setminus [c(x_{i_1}) \cup c(x_{i_2}) \cup \dots \cup c(x_{i_{l'}})],$$

one performs a *conditional multilocus NPL analysis*.<sup>3</sup>

Let  $\mathbf{x}$  and  $\mathbf{y}$  denote the vectors of free and conditional loci of lengths  $|l| - l'$  and  $l'$  respectively. In analogy with (8) the maximum conditional multilocus NPL score is given by,

$$Z_{\max, \mathbf{y}} = \sup_{\mathbf{x} \in \Omega_{c(x_{i_1}), c(x_{i_2}), \dots, c(x_{i_{l'}})}} Z(\mathbf{x}, \mathbf{y}), \quad (9)$$

One may note that the most general form of conditional analysis consists of conditioning with respect to actual inheritance vectors and that the standardization procedure is less straightforward, and more involved, in the conditional case (Ängquist and Hössjer, 2006).

**Remark 1** *The multilocus version of the well-known conditional Cox-procedure (Cox et al., 1999; Ängquist, 2001; Ängquist and Hössjer, 2006) may be formulated as,*

$$Z(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^N \gamma_k Z_k(\mathbf{x}) f [Z_k(\mathbf{y})], \quad (10)$$

---

<sup>3</sup>We might call this the  $(|l|, l')$ -multilocus conditional analysis approach.

where  $f(\cdot)$  is a function of the multilocus NPL score defined with respect to the, known or estimated, conditioning loci. Note that this may be seen as a  $[|l| - l']$ -multilocus analysis with weights, which are combinations of pedigree-specific weights  $\gamma_k$  and  $f$ -scores, obeying the constraint

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

Note that the conditioning in this case is made with respect to pedigree-specific NPL scores.

## 2.2 Two-Locus NPL Analysis

In the two-locus case  $|l| = 2$  and the general multilocus score function in (2) is simplified to,

$$S(w_1, w_2); \quad (w_1, w_2) \in \mathbb{V} \times \mathbb{V}.$$

### 2.2.1 Unconditional Case

In the unconditional case

$$Z(x_1, x_2) = \sum_{k=1}^N \gamma_k Z_k(x_1, x_2); \quad c(x_1) \neq c(x_2), \quad (11)$$

where  $\gamma_k$  still are, of course, pedigree weights satisfying (7). Moreover, the two-locus NPL maximum,

$$Z_{\max} = \sup_{x_1, x_2 \in \Omega} Z(x_1, x_2), \quad (12)$$

follows immediately from (8). For a more thorough discussion on this case for affected sib-pairs see Ängquist et al. (2005).

### 2.2.2 Conditional Case

A two-locus conditional analysis correspond to a single conditioning loci  $y$ , i.e.  $l' = 1$ , and a genome scan through the free chromosomes  $\Omega_{c(y)} = \Omega \setminus c(y)$ . Further, the maximum conditional two-locus NPL score is given by,

$$Z_{\max, y} = \sup_{x \in \Omega_{c(y)}} Z(x, y). \quad (13)$$

**Remark 2** *The Cox-procedure in (10) may be formulated as,*

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

*in the two-locus case. This is a multiplicative two-locus score which is analogous to a one-locus NPL score process along  $\Omega_{c(y)}$ , where the combined pedigree-weights depends, for instance, on a function of one-locus NPL scores at locus  $y$ .*

If prior evidence for disease loci to use as conditioning loci is poor, or at least vague, one might consider to use a *random* number of conditioning loci to use in the sense of (13) and to correct the results for multiple testing. One such procedure is outlined in Ängquist and Hössjer (2006) and will be shortly reviewed below.

Select conditioning chromosomes with respect to one-locus NPL scores as,

$$\mathbb{C} = \{c ; Z_{\max}(c) \geq z_c\}, \quad (14)$$

where  $Z_{\max}(c)$  is the NPL maximum over Chromosome  $c$  and  $z_c$  is a given chromosome-dependent score-threshold. Now, if one considers the locations corresponding to these chromosome-wise NPL score maximums,

$$y_c = \arg \max_{x \in c} Z(x) = \arg Z_{\max}(c),$$

one might select conditioning loci through,

$$\mathbb{Y} = \{y_c ; c \in \mathbb{C}\}. \quad (15)$$

Note that this guarantees that all (possibly zero) selected loci are located on distinct chromosomes.

### 2.2.3 Bibliographic Notes

In Strauch et al. (2000) the well-known score functions  $S_{\text{pairs}}$  and  $S_{\text{all}}$  (Whitemore and Halpern, 1994) was extended to the two-locus case, through summation of one-locus scores, and implemented into the program GENEHUNTER-TWOLOCUS. Li and Reich (2000) presented a complete list of distinct two-locus disease models for the binary phenotype-fully penetrant-biallelic disease

loci setting. Discussions on two-locus disease models and gene-gene interaction may be found in e.g. Bengtsson (2001), Holmans (2002) and Ängquist et al. (2005).

The Cox-procedure is applied in, for example, Schulze et al. (2004). A conditional approach based on generalized estimating equations is developed in Liang et al. (2001) and Chiu and Liang (2004). See also Dupuis et al. (1995).

For two-locus or multilocus analysis based on the maximum lod score (MLS) approach see Farrall (1997) and Cordell et al. (2000), and references therein. See also the reviews in Strauch et al. (2003) and Hoh and Ott (2003).

Other similar, but slightly different approaches to NPL analysis are given in e.g. Kong and Cox (1997), Zinn-Justin and Abel (1998) and Zinn-Justin et al. (2001).

### 3 Significance Calculations

The traditional statistical machinery of interpreting results and drawing conclusions is performed within the field constituted of the three closely related paths of *confidence interval construction*, *hypothesis testing* and *p-value calculation*. In some cases one might define distinct *evidential-based* approaches, more or less, contrasting the traditional procedures.

#### 3.1 General Case

Consider a *test statistic*  $T(X)$  which is a function of the underlying (possibly multivariate) random variable  $X$ . Find a *confidence interval*<sup>4</sup>  $C_\alpha$  with respect to a given significance level  $\alpha$  such that,

$$P_{H_0} [T(X) \in C_\alpha] = 1 - \alpha, \quad (16)$$

rejecting  $H_0$  if your result  $T(x) = t \in \bar{C}_\alpha$  for outcome  $X = x$ . Here  $\bar{C}_\alpha = \Omega \setminus C_\alpha$  is the complementary set called the *critical region*. This shows the one-to-one correspondence between hypothesis testing and confidence interval construction. Under a specific instance  $\lambda \in H_1$  of the alternative hypothesis

---

<sup>4</sup>Note that this is a somewhat *arbitrary* concept, i.e. the confidence interval is not uniquely defined through  $\alpha$ .

the power function corresponding to (16) is defined as,

$$\beta(T, C_\alpha) = P_\lambda [T(X) \in \bar{C}_\alpha], \quad (17)$$

where  $\beta(T, C_\alpha) \geq \alpha$  for reasonable, or *unbiased*, tests and instances of (17) close to 1 correspond to very strong, or *powerful*, statistical tests.

Finally, the procedure of *p*-value calculation calls for defining a critical region  $\bar{C}(t)$ , for finding  $T(X) = t$ , as

$$p(t) = P_{H_0} [T(X) \in \bar{C}(t)]. \quad (18)$$

In (18)  $p(t)$  is called the *p*-value corresponding to result  $t$  and  $\bar{C}(t)$  correspond to, so to speak, defining a confidence interval  $C(t)$  where  $t$  is located on the boundary between the regions. Usually this may be rephrased as  $p(t)$  being the probability of finding a value at least as *extreme* as  $t$  under the null hypothesis  $H_0$ .

Confidence intervals under strict acceptance of the *likelihood principle* is often called *support intervals*.<sup>5</sup> See e.g. Edwards (1992) and Clayton and Hills (1993).

## 3.2 Statistical Significance in NPL Analysis

### 3.2.1 Unconditional Case

For unconditional two-locus genome-wide NPL analysis based on (12), as with standard one-locus analysis, the natural test hypotheses are,

$$\begin{cases} H_0 : \text{No disease locus on } \Omega, \\ H_1 : \text{At least one disease locus on } \Omega, \end{cases} \quad (19)$$

which leads to the genome-wide significance level-function,

$$\alpha(z) = P_{H_0}(Z_{\max} \geq z), \quad (20)$$

and corresponding power-function,

$$\beta(z) = P_{H_1}(Z_{\max} \geq z). \quad (21)$$

In the context of (16)-(17) this constitutes a test rejecting  $H_0$  when  $Z_{\max} \geq z$ , i.e. for extreme positive scores.

---

<sup>5</sup>This means that inclusion of values into the confidence interval is based solely on which ones having the highest likelihoods. If the likelihood curve includes several local maximums this might lead to confidence intervals consisting of several nonconnected parts. More frequently this leads to nonsymmetric intervals surrounding a *maximum likelihood (ML)*-estimate.

### 3.2.2 Conditional Case

For conditional two-locus analysis based on the test statistic (13) one may restrict (19) by removing the conditioning chromosome  $c(y)$  from  $\Omega$ ,

$$\begin{cases} \bar{H}_0^{c(y)} : & \text{No disease locus outside chromosome } c(y). \\ \bar{H}_1^{c(y)} : & \text{At least one disease locus outside chromosome } c(y), \end{cases} \quad (22)$$

which leads to genome-wide conditional significance,

$$\alpha_y(z) = P_{\bar{H}_0^{c(y)}}(Z_{\max,y} \geq z | \text{MD}_{c(y)}), \quad (23)$$

and conditional power,

$$\beta_y(z) = P_{\bar{H}_1^{c(y)}}(Z_{\max,y} \geq z | \text{MD}_{c(y)}), \quad (24)$$

Note that (23)-(24) depends on the marker data  $\text{MD}_{c(y)}$  from Chromosome  $c(y)$ .

A conditional two-locus analysis based on a random number of conditioning loci selected using (14)-(15) leads to a refined approach based on conditional two-locus  $p$ -values of form (23),

$$p_c = \alpha_{y_c}(Z_{\max,y_c}); \quad y_c \in \mathbb{Y}. \quad (25)$$

The minimum  $p$ -value found using (25),

$$p_{\min} = \begin{cases} \min_{c \in \mathbb{C}} p_c & \text{if } \mathbb{C} \neq \emptyset, \\ 1 & \text{if } \mathbb{C} = \emptyset, \end{cases} \quad (26)$$

is then used as a test statistic, rejecting  $H_0$  whenever  $p_{\min}$  is *smaller than or equal* a given threshold  $u$ . Note that here the null hypothesis equals the standard one in (19),<sup>6</sup> hence ending up with a genome-wide (global) significance level,

$$\alpha_{\mathbb{Y}}(u) = P_{H_0}(p_{\min} \leq u), \quad (27)$$

and power,

$$\beta_{\mathbb{Y}}(u) = P_{H_1}(p_{\min} \leq u). \quad (28)$$

---

<sup>6</sup>This may be seen as taking the *intersection*, over all  $y_c \in \mathbb{Y}$ , of null hypotheses (22).

### 3.2.3 Some Notes on Significance Calculations

Significance level and power as in (23) and (24) may be calculated using either: (i) *Analytical approximations* based on Gaussian extreme value theory (Lander and Botstein, 1989; Lander and Kruglyak, 1995; Tang and Siegmund, 2001; Ängquist and Hössjer, 2005). (ii) *Monte Carlo Simulations* (Ploughman and Boehnke, 1989; Terwilliger et al., 1993; Malley et al., 2002; Ängquist and Hössjer, 2004). The advantage of (i) is fast computations and available explicit expressions. 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. Here the drawback is rather the computational burden.

An alternative or complementary approach to linkage analysis significance calculations is the *fuzzy significance concept* which is described by, for example, Thompson and Geyer (2005), Thompson (2006) and Ängquist (2006a). For a more general fuzzy view, consider Geyer and Meeden (2005).

## 4 Evidential Aspects of Positive Findings

Let us assume  $|l|$  true disease loci constituting a genetic disease model. In practise, these disease loci are fully, or partially, unknown or hidden. Hence, the test statistic  $T(x_1, x_2, \dots, x_{|l|})$ , if correctly rejecting  $H_0$ , does not necessarily suggest disease loci  $\arg \max_{x_1, x_2, \dots, x_{|l|} \in \Omega} T(x_1, x_2, \dots, x_{|l|})$  located on true disease regions or chromosomes.<sup>7</sup>

The remainder of this section will be dedicated to such evidential interpretations and problems of, conditional as well as unconditional, two-locus NPL analysis.

### 4.1 Unconditional Case

In this case  $T(x_1, x_2) = Z_{\max}$  in (12). The corresponding test hypotheses are defined in (19) and the natural interpretation is that a significant result represent statistical evidence, at the predefined significance level, for at *least one* disease locus among the pair of loci  $(\hat{l}_1, \hat{l}_2) = \arg Z_{\max}$ .

---

<sup>7</sup>In the multilocus case,  $T$  may correspond to (8)-(9), and in the two-locus case it may be identified with (12)-(13) and (26).

## 4.2 Conditional Case: A Single Known Conditioning Locus

Here one disease locus,  $l_2$ , is assumed to be known. Based on the test statistic  $T(x_1, x_2) = Z_{\max, y}$  in (13), the test hypotheses (22) imply that a significant result lead to a finding of *an additional* disease locus  $\hat{l}_1 = \arg Z_{\max, y}$ .

## 4.3 Conditional Case: A Single Estimated Conditioning Locus

This case is similar to the preceding one, though the difference is that  $l_2$  is not anymore assumed to be known but rather estimated from an initial one-locus analysis. Two alternatives are that: (i) Estimate the disease locus through  $\hat{l}_2 = \arg \max_{\Omega} Z(x)$ . (ii) Estimate the disease locus using  $\hat{l}_2 = \arg \max_{c(l_2)} Z(x)$ . In the latter case we simply assume that the actual disease chromosome  $c(l_2)$  is known.

**Remark 3** *It may not be unusual that the estimated conditioning locus procedure shows to be more powerful in finding  $l_1$  than the known conditioning locus procedure. This might seem odd but may be explained as that in cases where  $\hat{l}_2 \notin c(l_1) \cup c(l_2)$ , on one hand, one increases power with respect to the conditional analysis when scanning through both of the true disease chromosomes in the conditional sweep. On the other hand, if present one is not able to use any allele-sharing correlation with respect to  $l_1$  and  $l_2$ . According to this behaviour the interpretation in these cases are somewhat different.*

**Remark 4** *Vaguely speaking, one may redescribe this as follows. With increasing uncertainty regarding  $l_2$ , the evidence corresponding to estimate  $\hat{l}_2$  decreases but, at the same time, the evidence for estimate  $\hat{l}_1$  increases. In other words, with decreasing information on  $l_2$  the interpretation tends to that of a one-locus finding.*

## 4.4 Conditional Case: A Random Number of Estimated Conditioning Loci

In this complex case  $T(x_1, x_2) = p_{\min}$  in (26) and through (19) a significant result gives: (i) Some evidence for  $\hat{l}_2$  as a disease locus. (ii) Some evidence

for  $\hat{l}_1$  as an additional disease locus. The strength of evidence according to (i)-(ii) above are contrasting but both clearly score threshold-dependent.

The tuning parameters  $z_c$  from (14) are interesting. With increasing values they will lead to large general differences in the sizes of  $\mathbb{C}$  under  $H_0$  and  $H_1$ . Actually, this may be seen as that the evidence for  $\hat{l}_2$  as a disease locus will increase while the interpretation of  $\hat{l}_1$  is more involved. In some sense the evidence corresponding to  $\hat{l}_1$  will be masked for large thresholds  $z_c$ , since the null hypothesis of (19) will be rejected only for (at least) one strong genetic component, probably identified with the region of the conditioning locus, in which case an additional disease locus will be hard to discover.

Implications may be stated as that the evidential strength of finding disease loci is somewhat decreased since the set of interesting chromosomal positions underlying rejections of  $H_0$  through  $p_{\min}$  is generally enlarged.<sup>8</sup> For example, the complete set of loci  $(\hat{l}_1, y)$ ,  $y \in \mathbb{Y}$  from (25) underlying (26) might be interesting for exploratory, descriptive or data mining purposes.

## References

- Ängquist, L. (2001). *Conditional two-locus NPL-analyses: Theory and applications* (Master's thesis No. 2001:E22). Lund: Department of Mathematical Statistics, Lund University.
- Ängquist, L. (2006a, October). *Interpreting significance in nonparametric linkage analysis: Fuzzy p-values and information levels*. (Free download from homepage: '<http://www.maths.lth.se/matstat/staff/larsa/>'.)
- Ängquist, L. (2006b, June). *Some notes on the choice of score function in nonparametric linkage analysis*. (Free download from homepage: '<http://www.maths.lth.se/matstat/staff/larsa/>'.)
- Ängquist, L., Anevski, D. and Luthman, H. (2005). *Unconditional two-locus nonparametric linkage analysis: On composite null hypotheses with and without gene-gene interaction* (Tech. Rep. No. 2005:28). Lund: Department of Mathematical Statistics, Lund University.

---

<sup>8</sup>We generally extend our search for disease loci, therefore increasing the possibility of finding *something* significant, but reducing the certainty about *what* we have found. To explore - or to be sure - that's the question!

- 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. (2005). Improving the calculation of statistical significance in genome-wide scans. *Biostatistics*, 6(4), 520–538.
- Ängquist, L. and Hössjer, O. (2006). *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.
- Chiu, Y. F. and Liang, K. Y. (2004). Conditional multipoint linkage analysis using affected sib pairs: An alternative approach. *Genetic Epidemiology*, 26, 108–115.
- Clayton, D. and Hills, M. (1993). *Statistical models in epidemiology*. Oxford: Oxford University Press.
- 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.
- Donnelly, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology*, 23, 34–64.
- 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.

- Edwards, A. W. F. (1992). *Likelihood: Expanded edition* (Second Edition ed.). New York: John Hopkins University Press.
- Farrall, M. (1997). Affected sibpair linkage tests for multiple linked susceptibility genes. *Genetic Epidemiology*, *14*, 103–115.
- Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and p-values. *Statistical Science*, *20*(4), 358–387. (With discussion.)
- Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, *4*, 701–709.
- Holmans, P. (2002). Detecting gene-gene interactions using affected sib pair analysis with covariates. *Human Heredity*, *53*, 92–102.
- Hössjer, O. (2003). Determining inheritance distributions via stochastic penetrances. *Journal of the American Statistical Association*, *98*, 1035–1051.
- 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.
- 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*, *58*, 1347–1363.
- Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, *121*, 185–199.
- Lander, E. S. and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, *11*, 241–247.

- Lange, E. M. and Lange, K. (2004). Powerful allele-sharing statistics for nonparametric analysis. *Human Heredity*, *57*, 49–58.
- Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, *50*, 334–349.
- Liang, K. Y., Chiu, Y. F., Beaty, T. H. and Wjst, M. (2001). Multipoint analysis using affected sib-pairs: Incorporating linkage evidence from unlinked regions. *Genetic Epidemiology*, *21*, 105–122.
- Malley, J. D., Naiman, D. and Bailey-Wilson, J. (2002). A comprehensive method for genome scans. *Human Heredity*, *54*, 174–185.
- McPeck, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology*, *16*, 225–249.
- Ploughman, L. M. and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics*, *44*, 543–551.
- Schulze, T. G., Buervenich, S., Badner, J. A., Steele, C. J. M., Detera-Wadleigh, S. D., Dick, D., Foroud, T., Cox, N. J., MacKinnon, D. F., Potash, J. B., Berrettini, W. H., Byerley, W., Coryell, W., Jr, J. R. D., Gershon, E. S., Kelsoe, J. R., McInnis, M. G., Murphy, D. L., Reich, T., Scheftner, W., Jr, J. I. N. and McMahon, F. J. (2004). Loci on chromosomes 6q and 6p interact to increase susceptibility to bipolar affective disorder in the National Institute of Mental Health Genetics Initiative pedigrees. *Biological Psychiatry*, *56*, 18–23.
- 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.
- Tang, H. K. and Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics*, *2*, 147–162.

- Terwilliger, J. D., Speer, M. and Ott, J. (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genetic Epidemiology*, *10*, 217–224.
- Thompson, E. A. (2006). *Uncertainty in inheritance: Assessing evidence for linkage* (Tech. Rep. No. 498). Department of Statistics, University of Washington, Seattle, Washington.
- Thompson, E. A. and Geyer, C. J. (2005). *Fuzzy p-values in latent variable problems* (Tech. Rep. No. 481). Department of Statistics, University of Washington, Seattle, Washington.
- 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.
- Zinn-Justin, A., Ziegler, A. and Abel, L. (2001). Multipoint development of the weighted pairwise correlation (wpc) linkage method for pedigrees of arbitrary size and application to the analysis of breast cancer and alcoholism familial data. *Genetic Epidemiology*, *21*, 40–52.