



Mathematical Statistics  
Stockholm University

A General Method for Linkage  
Disequilibrium Correction for  
Multipoint Linkage and Association

Azra Kurbasic and Ola Hössjer

Research Report 2006:9

ISSN 1650-0377

**Postal address:**

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

**Internet:**

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



Mathematical Statistics  
Stockholm University  
Research Report **2006:9**,  
<http://www.math.su.se/matstat>

# A General Method for Linkage Disequilibrium Correction for Multipoint Linkage and Association

Azra Kurbasic \* and Ola Hössjer<sup>†</sup>

December 2006

## Abstract

Lately, many different methods of linkage, association, or joint analysis for family data have been invented and refined. Common to most of those is that they require a map of markers that are in linkage equilibrium. However, at the present day, high-density SNPs maps are both more inexpensive to create and they have lower genotyping error. When marker data is incomplete, the crucial and computationally most demanding moment in the analysis is to calculate the inheritance distribution at a certain position on the chromosome. Recently, different ways of adjusting traditional methods of linkage analysis to denser maps of SNPs in linkage disequilibrium have been proposed. We describe a hidden Markov model which generalizes the Lander-Green algorithm. It combines Markov chain for inheritance vectors with a Markov chain modelling founder haplotypes and in this way takes account for linkage disequilibrium between SNPs. It can be applied to association, linkage, or combined association and linkage analysis, general phenotypes and arbitrary score functions. We also define a joint likelihood for linkage and association that extends an idea of Kong and Cox (1997) for pure linkage analysis.

**KEY WORDS:** Hidden Markov model, linkage, association, linkage disequilibrium, SNPs.

---

\*Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Box 118, 221 00 LUND, Sweden. E-mail: [azra@maths.lth.se](mailto:azra@maths.lth.se).

<sup>†</sup>Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: [ola@math.su.se](mailto:ola@math.su.se). Financial support from the Swedish Research Council, contract nr. 626-2002-6286.

## Introduction

Mapping genes due to increased identical by descent (IBD) sharing has become widespread the last two decades and many different methods of linkage, association or joint linkage/association analyses have been invented, all based on IBD sharing within or between families. Common to most of the linkage or joint association/linkage methods is that they require use of markers that are in linkage equilibrium (LE). Lately, it has become customary to use high-density (single nucleotide polymorphisms) SNP maps. The advantage of SNPs is that their greater density will compensate for the smaller amount of information per SNP by creating local haplotypes of SNPs, that function as “super” alleles. Jointly as a haplotype these should have greater linkage information content. Further they are associated with lower genotyping error, cf. Evans and Cardon (2003). On the other hand, SNPs are more likely to be in strong linkage disequilibrium (LD) and misspecification of the LD structure is analogous to misspecification of marker allele frequencies in classical linkage analysis. Many authors have recently noted that misspecification of population haplotype frequencies causes inflation of multipoint (parametric or nonparametric) linkage scores, when some or all founders are not genotyped, see for instance Schaid et al. (2002, 2004), Browning et al. (2004), Broman and Feingold (2004), Huang et al. (2004), Boyles et al. (2005) and Abecasis and Wigginton (2005).

There are different suggestions of what to do for dense marker maps, i.e. clustering SNPs or choosing a subset of “independent” SNPs by retaining a single SNP from each cluster, cf. Abecasis and Wigginton (2005) and Browning et al. (2004). In regions of high LD, haplotype blocks exist and a reduced set of haplotype tag SNPs that identify common haplotypes may be selected in the analysis, as proposed by Boyles et al. (2005) and Stram et al. (2003). Incorporation of flanking markers not in LD with the cluster moderate the LD effect but the false positive rate is still higher than expected (Boyles et al., 2005).

Joint association and linkage analysis face similar problems when marker maps are tight and founders not genotyped, mainly due to presence of the linkage component. Several methods for joint linkage and association analysis have recently been proposed by a number of authors, for instance Xiong and Jin (2000), Göring and Terwilliger (2000), Fulker et al. (1999), Sham et al. (2000), Pérez-Enciso (2003), Li et al. (2005), Cantor et al. (2005), Jung et al.

(2005) and Hössjer (2005a). The methods in these papers are either single point, or, if multipoint, different markers are either grouped into clusters, with no within-cluster recombinations allowed for, or assumed to be in LE. Cantor et al. (2005) mention the possibility of allowing for more general marker haplotype structures by using the Elston-Stewart algorithm (Elston and Stewart, 1971).

In this paper we describe how to incorporate LD into multipoint analysis by generalizing the Lander-Green algorithm. This we do by combining the Markov chain for inheritance vectors used in multipoint linkage analysis (Lander and Green, 1987) with another  $L^{\text{th}}$  order Markov chain that models LD structure of founder marker data. The latter Markov chain has recently been used for singletons by Eronen et al. (2004) as a way to estimate haplotype frequencies. Using hidden Markov models, we quantify incompleteness of marker data by means of the conditional distribution of founder haplotypes and the inheritance vector given marker data.

Our framework incorporates not only joint tests for linkage and association, but also separate tests for linkage and association (see Hössjer, 2005a). It allows for parametric as well as nonparametric choices of score function and various kinds of genetic models involving for instance binary or quantitative phenotypes. In principle, arbitrary pedigree structures are allowed for, although the present implementation of the method is only computationally feasible for small  $L$  and small pedigrees.

To handle incomplete marker data, Kruglyak et al. (1996) used the expected value of the scoring function given marker data. The resulting test is conservative though when information on descent is incomplete since the variance of the test statistic is overestimated. Kong and Cox (1997) proposed a one parameter allele sharing likelihood model. The resulting likelihood ratio (LR) test is very similar to the approach of Kruglyak et al. (1996) for complete marker data but much less conservative otherwise. We define a two-parameter extension of the Kong and Cox model to handle linkage and association jointly. By restricting the parameter space, it is also possible to define separate linkage and association tests.

Finally, we illustrate our method by means of a small simulation study, where marker data is generated with LD structure and then analyzed both under LD and LE assumptions. Our results indicate that i) association tests are much less sensitive to the LD assumption than linkage tests and joint association/linkage tests and ii) that our method is able to reduce inflated linkage scores.

# Multipoint Approach to Linkage and Association

Assume a map of  $K$  markers at loci  $x_1 < \dots < x_K$  along one chromosome, with map distance measured in Morgans. We consider initially one pedigree with  $n$  individuals, of which  $f$  are founders and  $n - f$  nonfounders. Assume pedigree members are numbered so that founders have labels  $1, \dots, f$  and nonfounders labels  $f + 1, \dots, n$ . We let  $b_{i,2k-1}$  be the maternal allele of individual  $k$  at  $x_i$  and  $b_{i,2k}$  the paternal allele ( $1 \leq k \leq n, 1 \leq i \leq K$ ). Further, assume that the marker allele at  $x_i$  is  $d_i$ -allelic, so that  $b_{i,2k-1}$  and  $b_{i,2k}$  can attain  $d_i$  different values. Let

$$M_i = \{(b_{i,2k-1}b_{i,2k}), k \in \mathcal{T}_i\}$$

be observed marker data at  $x_i$ , with  $\mathcal{T}_i \subset \{1, \dots, n\}$  the set of genotyped individuals at  $x_i$ . Here  $(b_{i,2k-1}b_{i,2k})$  is the marker genotype of  $k$  at  $x_i$ , which a priori has unknown phase. (Given data we might know the phase though if  $k$  is a nonfounder.) Notice that  $\mathcal{T}_i = \emptyset$  is possible, in which case we interpret  $x_i$  as a marker locus with completely missing data. Let

$$\mathbf{M} = (M_1, \dots, M_K)$$

be marker data at all loci for all individuals and  $m = 2(n - f)$  the number of meioses in the pedigree. Define  $v_i = (v_{i1}, \dots, v_{im})$  as the inheritance vector at  $x_i$ , with  $v_{il}$  equal to zero or one depending on whether a grandpaternal or grandmaternal allele was transmitted during the  $l^{\text{th}}$  meioses (Donnelly, 1983), and  $b_i = (b_{i1}, \dots, b_{i,2f})$  as the founder allele vector at  $x_i$ . We combine  $v_i$  and  $b_i$  into an allele configuration

$$w_i = (b_i, v_i)$$

at  $x_i$ . In the ideal case of complete marker data and known phase of all founders, we would observe  $w_1, \dots, w_K$ . Hence we interpret  $w_i$  as complete marker data at locus  $x_i$ . Any one-locus association, linkage or combined association and linkage tests statistic at locus  $x_i$  is a function of  $w_i$  if marker data is complete (Hössjer, 2005a). For incomplete marker data, we retrieve information about  $w_i$  from  $\mathbf{M}$  using the multipoint distribution

$$P_i(w) = P(w_i = w | \mathbf{M}). \tag{1}$$

## Markov Model for Multipoint Probability

Usually, (1) is evaluated by means of the Lander-Green algorithm, assuming linkage equilibrium between markers at different loci, see Lander and

Green (1987) and the appendix of Hössjer (2005a) for details. This is in fact the forward-backward algorithm for Hidden Markov Models (HMMs) with  $\{v_i\}_{i=1}^K$  as hidden Markov regime, see e.g. Baum (1972) and Rabiner (1989). For this we need to assume no chiasma interference, so that  $\{v_i\}$  becomes a Markov chain with transition probabilities

$$P(v_{i+1}|v_i) = \theta_i^{|v_{i+1}-v_i|} (1 - \theta_i)^{m-|v_{i+1}-v_i|}, \quad (2)$$

where  $\theta_i$  is the recombination fraction between  $x_i$  and  $x_{i+1}$  and  $|v_{i+1} - v_i| = \sum_{l=1}^m |v_{i+1,l} - v_{i,l}|$  the Hamming distance between  $v_i$  and  $v_{i+1}$ . The size of the state space of  $v_i$  is  $2^m$ .

Our objective is to generalize the Lander-Green algorithm by allowing for LD between markers. We model LD structure by assuming that marker alleles along each founder haplotype form an  $L^{\text{th}}$  order Markov chain, i.e.

$$P(b_{i+1,k}|b_{ik}, \dots, b_{1k}) = P(b_{i+1,k}|b_{ik}, \dots, b_{i-L_i+1,k}) \quad (3)$$

for  $i = 1, \dots, K - 1$ ,  $L_i = \min(L, i)$  and  $k = 1, \dots, 2f$ . The transition probabilities in (3) depend on  $i$  but not on  $k$ . They involve frequencies of haplotypes of length  $L_i + 1$ . Indeed, let  $h_{ik} = (b_{i-L_i+1,k}, \dots, b_{ik})$  denote the  $k^{\text{th}}$  founder haplotype formed by  $L_i$  consecutive loci, the rightmost one being  $x_i$  and  $g_{ik} = (b_{i-L_i,k}, \dots, b_{ik})$  the corresponding haplotype of length  $L_i + 1$ , also having its rightmost allele at  $x_i$ . Then

$$P(b_{i+1,k}|b_{ik}, \dots, b_{i-L_i+1,k}) = \frac{P(g_{i+1,k})}{P(h_{i,k})} = \frac{P(g_{i+1,k})}{\sum P(\tilde{g}_{i+1,k})}, \quad (4)$$

where the sum in the rightmost denominator is taken over all haplotypes  $\tilde{g}_{i+1,k}$  of length  $L_i + 1$  whose  $L_i$  leftmost alleles equal  $h_{ik}$  and whose rightmost allele is at  $x_i$ . For singletons ( $n = 1$ ), the Markov model (3)-(4) has been used by Eronen et al. (2004) for haplotype reconstruction.

The usual LE assumption in linkage analysis corresponds to  $L = 0$ . We will assume  $L > 0$ , and, for this we need to enlarge the state space and consider  $L_i$  adjacent loci simultaneously. Define  $B_i = (b_{i-L_i+1}, \dots, b_i)$  and

$$W_i = (B_i, v_i), \quad (5)$$

for  $i = 1, \dots, K$ . We will assume no segregation distortion ( $\{b_i\}$  and  $\{v_i\}$  independent processes) and Hardy-Weinberg equilibrium ( $\{b_{ik}\}$  independent processes for different  $k$ ). Because of (2) and (3), it follows that  $\{W_i\}$  is a Markov chain, and  $P(W_i|M)$  can be computed by means of the forward-backward algorithm, see the appendix for details. This in turn provides the sought multipoint probabilities (1) through

$$P_i(w) = \sum_{W_i; w_i=w} P(W_i|M).$$

As shown in the appendix, the computational complexity of the forward-backward algorithm without any speedups for SNPs ( $d_1 = \dots = d_K = 2$ ) is

$$O\left(K \cdot 2^{2(L+1)f+2m}\right), \quad (6)$$

which is intractable for all but very small  $L$ ,  $f$  and  $m$ .

## Computational Savings

As a first speedup, we utilize that haplotype phase of all founders is unknown, since marker data is unchanged if any founder's haplotypes are switched at the same time as the mode of allele transmission to all his offspring is switched.

Consider a fixed locus  $x_i$  and, for simplicity, drop index  $i$ , so that  $W_i = W = (B, v)$  and  $h_{ik} = h_k$ ,  $k = 1, \dots, 2f$ . The idea is to subdivide all  $W$  into a number of sets. The allele configurations  $W$  within each set cannot be distinguished from marker data and hence can be treated as a single element. Following Kruglyak et al. (1996), we let  $c_k$  be the inheritance vector of length  $m$  which has ones in all bits corresponding to those meioses where founder  $k$  transmits alleles to his or her children,  $k = 1, \dots, f$ . Let  $\pi_k$  be a function of  $B$  that switches founder haplotypes  $2k - 1$  and  $2k$  and leaves all other founder haplotypes fixed. Since  $B = (h_k)_{k=1}^{2f}$  we get

$$\pi_k(B) = (h_1, \dots, h_{2k-2}, h_{2k}, h_{2k-1}, h_{2k+1}, \dots, h_{2f}).$$

A phase switch of founder  $k$  corresponds to changing  $W$  to  $(\pi_k(B), v + c_k)$ , and this configuration can not be distinguished from  $W$  using marker data. By combining founder phase switches for all  $f$  founders in all  $2^f$  possible ways we get equivalence classes

$$\overline{W} = \left\{ (\pi_1^{\xi_1} \circ \dots \circ \pi_f^{\xi_f}(B), v + \sum_{k=1}^f \xi_k c_k), \xi_k \in \{0, 1\}, k = 1, \dots, f \right\} \quad (7)$$

of size  $2^f$ .

A second speedup of matrix multiplication in the HMM algorithm is achieved by updating individual meioses and founder haplotypes individually. This generalizes an idea of Idury and Elston (1997) for pure lineage analysis.

With both of the abovementioned speedups incorporated, it is shown in the appendix that the total number of operations is

$$O(K \cdot (2f + m)2^{2fL+m-f}). \quad (8)$$

This is still intractable for large  $L$ ,  $f$  and  $m$ , but less so than (6).



## Score Functions

We consider tests of association and/or linkage of a given locus  $x_i$  to a disease locus  $\tau$ . The hypothesis testing problem is formulated as

$$\begin{aligned} H_{0i} &: \tau \text{ unlinked to } x_i, \\ H_{1i} &: \tau = x_i, \epsilon \neq 0, \end{aligned} \tag{9}$$

where  $\epsilon$  consists of genetic model parameters of the disease, with  $\epsilon = 0$  corresponding to no genetic component. In general,  $\epsilon$  involves penetrance parameters and allele frequencies at  $x_i$ . It may also contain association parameters between  $x_i$  and  $\tau$ , although this is not needed when  $x_i$  coincides with disease locus, since the two loci are then in complete association<sup>1</sup>. It is mathematically equivalent to reformulate the null hypothesis as having a 'disease locus'  $\tau$  at  $x_i$  with no genetic effect and rewrite (9) as

$$\begin{aligned} H_{0i} &: \tau = x_i, \epsilon = 0, \\ H_{1i} &: \tau = x_i, \epsilon \neq 0, \end{aligned} \tag{10}$$

The latter formulation (10) was originally introduced by Whittemore (1996) for linkage tests statistics. The advantage of (10) over (9) is that the parameter  $\epsilon$  is used for testing  $H_{0i}$  versus  $H_{1i}$ . In the sequel, we write  $H_{0i} = H_0$ , since the null hypothesis is independent of  $x_i$  as long as we consider markers along one chromosome.

Let  $b = (b_1, \dots, b_{2f})$  and  $v = (v_1, \dots, v_m)$  be arbitrary founder allele and inheritance vectors and  $w = (b, v)$ . Write

$$P_{i,\epsilon}(w) = P(w_i = w | \tau = x_i, \epsilon, Y).$$

for the allele configuration distribution at a disease locus  $x_i$ , conditional on observed phenotypes  $Y$  in the pedigree. In general, it depends on allele frequencies at  $x_i$  and penetrance parameters of the disease. Notice in particular that

$$P_{i,0}(w) = P_0(w_i = w | Y) = P(w_i = w) = 2^{-m} P(b_i = b), \tag{11}$$

where  $P(b_i = b) = \prod_{k=1}^{2f} P(b_{ik} = b_k)$ , and subscript 0 refers to probabilities under  $H_0$ . We will consider retrospective log likelihood

$$l(x_i, \epsilon) = \log P_{i,\epsilon}(\mathbf{M} | Y) = \log \left( \sum_w P(\mathbf{M} | w) P_{i,\epsilon}(w) \right) \tag{12}$$

---

<sup>1</sup>In Hössjer (2005a), such an association component was included, and the additional requirement 'alleles at  $\tau$  and  $x_i$  are not associated' was added to  $H_{0i}$ . This is reasonable for coarser marker maps not including the disease locus itself. For dense marker maps, covering all or most polymorphic loci, we find (9) more natural.

and assume that  $\epsilon = (\epsilon_1, \epsilon_2)$  involves two parameters. The first one,  $\epsilon_1$ , corresponds to association tests and the second one,  $\epsilon_2$ , to linkage tests. For complete marker data, when  $w_i$  is observed, the likelihood score vector

$$l'_i = l'(x_i, (0, 0)) = S(w_i) - \mu_i \quad (13)$$

for some score function  $S(w) = S(w; Y)$  and constant  $\mu_i = E_0(S(w_i)|Y)$ . We regard  $S$  as a function of  $w$ , which also depends on phenotypes  $Y$  and (possibly) genetic model parameters that are known or estimated in advance.  $l'(x_i, \epsilon)$  is the partial derivative of  $l(x_i, \epsilon)$  with respect to  $\epsilon$ . For pure association tests, the derivative is taken with respect to one component,  $\epsilon_1$ , so that  $S = S_1$  is a scalar. For pure linkage tests, the derivative is taken with respect to  $\epsilon_2$ , giving the scalar function  $S = S_2$ . Finally, for combined linkage and association tests, the derivative is taken with respect to both components of  $\epsilon$ , so that  $S = (S_1, S_2)$  is a vector. For more details on likelihoods leading to (13), see Whittemore (1996), McPeck (1999) and Hössjer (2003, 2005b) for linkage tests, Clayton (1999), Whittemore and Tu (2000) and Shih and Whittemore (2002) for association tests and Hössjer (2005a) for combined association and linkage tests.

We consider association and linkage score functions of the form

$$\begin{aligned} S_1(w) &= \sum_{k=1}^n \omega_k (b_{2k-1} + b_{2k}) \\ S_2(w) &= \sum_{1 \leq k < l \leq n} \omega_{kl} \text{IBD}_{kl} \end{aligned} \quad (14)$$

where  $\text{IBD}_{kl}$  is the number of alleles shared IBD by  $k$  and  $l$  and  $\omega_k$  and  $\omega_{kl}$  are weights assigned to individuals and pairs of individuals, both of which depend on phenotypes, and possibly also on fixed genetic model parameters. Both  $S_1$  and  $S_2$  appear in the likelihood score vector (13) for certain low penetrance genetic models.

A nonfounder version of  $S_1$  is defined by conditioning on founder alleles,

$$S_1^{\text{NF}}(w) = S_1(w) - E_0(S_1(w_i)|b, Y). \quad (15)$$

In general, test procedures based on  $S_1^{\text{NF}}$  are less powerful than those using  $S_1$ , since information is lost by conditioning on founder alleles. On the other hand,  $S_1^{\text{NF}}$  is more robust against spurious association due to population admixture, see Clayton (1999) and Shih and Whittemore (2002).

In the sequel, our framework incorporates score functions for pure association ( $S = S_1$  or  $S_1^{\text{NF}}$ ), pure linkage ( $S = S_2$ ) or combined linkage and association

( $S = (S_1, S_2)$  or ( $S_1^{\text{NF}}, S_2$ )). The standardized version<sup>2</sup> of  $S$  at locus  $x_i$  is

$$Z_i(w) = (S(w) - \mu_i)\Sigma_i^{-1/2}, \quad (16)$$

where  $\Sigma_i = \text{Var}_0(S(w_i)|Y)$ . Both  $\mu_i$  and  $\Sigma_i$  are scalars for pure association and linkage score functions, but a  $1 \times 2$  vector and a  $2 \times 2$  diagonal matrix for combined linkage and association, see Hössjer (2005a) for details. If  $S = S_2$ ,  $Z_i$  is independent of  $i$ , but in all other cases it depends on the allele frequencies locus  $x_i$ .

For incomplete marker data, we don't observe  $Z_i(w_i)$ . Instead, we use the multipoint distribution (1) and define<sup>3</sup> the family score

$$\bar{Z}_i = E(Z_i(w_i|\mathbf{M})) = \sum_w Z_i(w)P_i(w) \quad (17)$$

at locus  $x_i$ . For pure linkage tests, this reduces to the family NPL score of Kruglyak et al. (1996). The likelihood score vector (13) then generalizes to

$$l'_i = \sum_w (S(w) - \mu_i)P_i(w) = \bar{Z}_i\gamma_i, \quad (18)$$

where  $\gamma_i = \Sigma_i^{1/2}$  is either a scalar or a diagonal matrix of order 2.

## An Allele Configuration Model

We will now reverse the procedure of the previous section, starting with a score vector  $S(w)$  with both association and linkage components, computing  $Z_i$  according to (16) and then defining a two parameter allele configuration model

$$P_{i,\epsilon}(w) = P_{i,0}(w)(1 + Z_i(w)\gamma_i\epsilon^T), \quad \epsilon = (\epsilon_1, \epsilon_2) \in \Theta, \quad (19)$$

where  $\gamma_i = \text{diag}(\gamma_{i1}, \gamma_{i2})$  is a given diagonal  $2 \times 2$  weight matrix at locus  $x_i$ . A possible choice is  $\gamma_i = \Sigma_i^{1/2}$ , although we will allow for other weighting schemes. In (19),  $\epsilon_1$  and  $\epsilon_2$  quantify strength of association and linkage

---

<sup>2</sup>An alternative standardization  $Z(w) = S_1^{\text{NF}}(w)/\sqrt{\Sigma(b)}$  of the nonfounder score function is possible, with  $\Sigma(b) = \text{Var}_0(S_1(w)|b, Y)$ . However, it is not well defined when  $\Sigma(b) = 0$ . For a nuclear family, this happens

when both parents are homozygote.

<sup>3</sup>In fact, for a reduced state space we let  $\bar{w}$  denote equivalence class (7) with  $L = 1$ . Under the mild requirement that  $S(w)$  and hence also  $Z(w)$  is the same for all  $w \in \bar{w}$  we notice that (17) is equivalent to  $\bar{Z}_i = \sum_{\bar{w}} Z_i(\bar{w})P_i(\bar{w})$ , where  $P_i(\bar{w})P(\bar{w}_i = \bar{w}|\mathbf{M}) = \sum_{\bar{W}_i; \bar{w}_i = \bar{w}} P(\bar{W}_i|\mathbf{M})$ .

respectively, without having direct interpretation in terms of genetic model parameters. The parameter space  $\Theta$  is defined by the requirements  $\epsilon_2 \geq 0$  and  $P_{i,\epsilon}(w) \geq 0$  for all  $w$ . Notice that we only allow for positive values of  $\epsilon_2$ , corresponding to increased allele sharing among affected for binary phenotypes. On the other hand, both positive and negative values of  $\epsilon_1$  are allowed for, since either of the two alleles at  $x_i$  may be associated with the disease. It is possible to restrict  $\Theta$  further as in pure association models ( $\epsilon_2 = 0$ ) and pure linkage models ( $\epsilon_1 = 0$ ). In the latter case, (19) is equivalent to the linear model of Kong and Cox (1997).

It is shown in the appendix that the retrospective log likelihood based on (19) is

$$l(x_i, \epsilon) = \text{constant} + \log \left( 1 + \bar{Z}_i \gamma_i \epsilon^T \right), \quad (20)$$

with a constant depending on marker data  $\mathbf{M}$ , but not on  $\epsilon$ . Hence the score vector  $l'_i$  obtained from (20) is (18). The diagonal entries of  $\gamma_i = \text{diag}(\gamma_{i1}, \gamma_{i2})$  are weights given to the association and linkage components of the family.

## Test Statistics

Consider a collection of  $N$  families (of possibly different structure) with marker and phenotype data  $(\mathbf{M}_1, Y_1), \dots, (\mathbf{M}_N, Y_N)$ . The total retrospective log likelihood for all families is

$$\begin{aligned} l(x_i, \epsilon) &= \sum_{j=1}^N \log P_{i,\epsilon}(\mathbf{M}_j | Y_j) \\ &= \text{constant} + \sum_{j=1}^N \log \left( 1 + \bar{Z}_{ji} \gamma_{ji} \epsilon^T \right), \quad \epsilon \in \Theta, \end{aligned} \quad (21)$$

where  $\bar{Z}_{ji}$  and  $\gamma_{ji} = \text{diag}(\gamma_{ji1}, \gamma_{ji2})$  are the likelihood score vectors and weight matrices of the  $j^{\text{th}}$  family at locus  $x_i$ . The parameter space  $\Theta = \bigcap_{j=1}^N \Theta_j$  is the intersection of the parameter spaces of all individual families.

Let

$$\hat{\epsilon}_i = \arg \max_{\epsilon \in \Theta} l(x_i, \epsilon)$$

be the maximum likelihood estimator of  $\epsilon$  at  $x_i$  and define the log likelihood ratio statistic

$$T_i = 2(l(x_i, \hat{\epsilon}_i) - l(x_i, (0, 0))), \quad (22)$$

as test statistic for (10). The asymptotic distribution of  $T_i$  is  $0.5\chi^2(1) + 0.5\chi^2(2)$  for combined association and linkage tests,  $\chi^2(1)$  for pure association

tests and  $0.5\chi^2(0) + 0.5\chi^2(1)$  for pure linkage tests, see e.g. Self and Liang (1987).

The LR test only depends on the family scores  $\bar{Z}_{ji}$ , but not their variance, which is a computational advantage. On the other hand, a likelihood score test requires calculation of the covariance matrix of

$$l'_i = \sum_{j=1}^N \bar{Z}_{ji} \gamma_{ji},$$

which is intractable and typically requires Monte Carlo simulation of marker data for all families. If the weights are normalized as  $\sum_{j=1}^N \gamma_{ji}^2 = 1$ , we notice that approximately,  $l'_i$  has a (one- or two-dimensional) standard normal distribution for large  $N$  when marker data is complete.

## Estimation of Haplotype Frequencies and Simulation

Looking at the transition probability (3) for founder alleles one realizes that we need to specify  $K - L$  distributions of haplotypes  $g_{i,k}$  of length  $L + 1$ , one distribution for each  $i = L + 1, \dots, K$ . Various methods for estimating haplotype frequencies exists, see e.g. Hodge et al. (1999), Douglas et al. (2001) and Niu et al. (2002). We use the software Haplotyper (Niu et al., 2002).

There are different approaches to pedigree data simulation under LD assumption. One possibility is to first simulate phenotypes and genotypes at the disease locus, starting with an affected proband. Then disease genotypes and affection status of all other individuals is determined, using the disease allele frequencies and specified penetrances, see Bass et al. (2002). This is repeated until the phenotypes meet user-specified ascertainment criteria. Finally, marker genotypes are assigned to all individuals starting with the founders.

We will follow another approach, where marker data is simulated conditionally on phenotypes, see for instance Boehnke (1986), Ploughman and Boehnke (1989) and Ott (1989), Terwilliger et al. (1993). For one single family, simulation from  $P_{i,\epsilon}(\mathbf{M}|Y)$  can be achieved in two steps as follows:

1. Generate  $\mathbf{b} = (b_1, \dots, b_K)$  and  $\mathbf{v} = (v_1, \dots, v_K)$  from  $P_{i,\epsilon}(\mathbf{b}, \mathbf{v}|Y)$ .
2. Compute  $\mathbf{M} = (M_1, \dots, M_K)$  by spreading founder alleles from  $\mathbf{b}$  according to  $\mathbf{v}$  at each locus and hiding genotypes for untyped individuals.

In Step 1 of the above algorithm, we proceed differently under  $H_0$  and  $H_{1i}$ , see the appendix for details.

## Results

For simplicity, we restrict ourselves to binary phenotypes and put  $\omega_k = 1$  if  $k$  is affected and  $\omega_k = 0$  otherwise (unaffected or unknown phenotype). For pairs of individuals we put  $\omega_{kl} = 1$  if both  $k$  and  $l$  are affected and zero otherwise. With these weights,  $S_2 = S_{\text{pairs}}$ , the linkage score function introduced by Whittemore and Halpern (1994).

We only consider sibpair families, all with the same phenotype vector and genotyped family members. Since the families are exchangeable, we use a constant weight matrix

$$\gamma_{ji} = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

in (21). The parents of each family are numbered 1,2, the children 3,4, the phenotype vector is denoted  $Y = (Y_1, \dots, Y_4)$ , where  $Y_k = 1$  if  $k$  is affected and  $Y_k = 0$  if  $k$  is unaffected or has unknown phenotype.

In the simulations, we consider a short part of Chromosome 9. To get a realistic picture of LD, we use CEPH (Utah residents with ancestry from northern and western Europe) population data from the HapMap database (Thorisson et al., 2005) to estimate haplotype frequencies. Information about markers and LD between the markers in the cluster are given in Tables 1, 2, 3 and in Figure 1. The choice of SNPs and map length is influenced by the fact that in most cases significant linkage disequilibrium does not influence markers separated by 0.1 cM in outbreed populations, see Evans and Cardon (2004).

Under the null hypothesis, we simulate *one* data set with 1000 sibpair families and then analyse it under different phenotype and genotyping scenarios. We generate founder haplotypes using the Markov model with  $L = 1$  and haplotype frequencies as in Table 2. The simulated data is then analyzed by means of the HMM algorithm, assuming either LD ( $L = 1$ ) or LE ( $L = 0$ ). In the former case, we use the same haplotype frequencies as in the simulations and in the latter case we use marker allele frequencies from Table 1.

Simulation results with one parent untyped are given in Figures 2-6. Among the association LR tests ( $T_1$  and  $T_1^{\text{NF}}$ ) and linkage LR test ( $T_2$ ),  $T_2$  is by far most sensitive to incorrect LD assumptions in the analysis, followed by  $T_1^{\text{NF}}$  and  $T_1$ . The reason why  $T_1^{\text{NF}}$  is more sensitive to LD than  $T_1$  when parental

marker data is missing, is the conditioning on founder alleles in the definition of  $S_1^{NF}$ , see (15) and Figure 3. For  $T_1$  there is no difference between analyses under LD and LE assumption when parents are unaffected or have unknown phenotypes, see Figures 7 and 8. This holds regardless of which persons are typed for markers. It is not surprising, since  $\omega_1 = \omega_2 = 0$  and hence the parental marker alleles are not included in the definition of  $S_1$ .

As mentioned in the introduction, the sensitivity of linkage scores to LD assumptions depends on which individuals that are genotyped for markers. There is inflation of  $T_2$  when one parent is untyped for markers (Figure 4), but even more so when both parents are untyped, cf. Figures 9 and 10 for the outcome of two different simulations. Since a very short chromosomal region is analyzed, it is not surprising that  $T_2$  is almost constant.

Joint association and linkage analysis ( $T_{combined}$ ) is mainly sensitive to LD due to the linkage component, see Figures 5, 6, 11, and 12.

When simulating under the alternative hypothesis we have 1000 sib pair families and choose one marker as disease locus, see Figures 13-17. The genetic model corresponds to a low penetrant disease with a rather small (7.8%) disease allele frequency. It is detected by  $T_1$ ,  $T_2$  and  $T_{combined}$ , but not by  $T_1^{NF}$ . There is the same kind of sensitivity to LD as when simulating under the null hypothesis.

## Discussion

We have suggested and described a method, based on hidden Markov models, for handling LD between markers in multipoint association and/or linkage analysis. The novel feature, in this context of gene mapping, is to model founder haplotypes by means of a Markov chain. We have also extended the likelihood approach of Kong and Cox (1997) for handling incomplete marker data to association and combined association and linkage tests.

Simulations for sib pair families show that proper handling of LD is most important for incomplete marker data, in particular when both parents are untyped. Further, LD affects linkage analysis much more than association analysis. Combined association and linkage analysis is principally affected by LD through the linkage component. For association tests, conditioning on founder alleles slightly increases sensitivity to LD.

Our work can be extended in several ways. First of all, more extensive simulations are needed for various combinations of pedigrees, genotyping scenarios and phenotypes. The LD structure, in particular  $L$ , could be varied both in

simulation and analysis. Empirical pointwise and genomewide  $p$ -values could be computed, to check the validity of asymptotic results, in particular the  $\chi^2$  mixture approximation of pointwise  $p$ -values. This is important both under correct and incorrect LD-modelling. For instance,  $T_2$  is quite large in Figures 9-10 (see also Table 4) even when LD is modelled correctly, indicating inaccuracy of the nominal  $p$ -value approximation. Alternatively, when families are exchangeable or close to exchangeable, one might use permutation testing to assess both pointwise and genomewide  $p$ -values.

Another possibility is to consider score functions depending on marker data at *several loci*. This amounts to using a pedigreewise log likelihood (20), with

$$\bar{Z}_i = \sum_W Z_i(W) P(W_i = W|M),$$

where  $Z_i(W) = (S(W) - \mu_i)\Sigma_i^{-1/2}$ ,  $\mu_i = E_0(S(W_i))$ ,  $\Sigma_i = \text{Var}_0(S(W_i))$ , and

$$S = S(W) \tag{23}$$

is a score function depending on marker data from all  $L_i$  loci included in  $W = (B, v)$ . We conjecture that (23) is preferable over (14) mainly for coarser

marker maps such that the disease locus is located between two neighbouring markers. In this case, the founder marker haplotype are usually more closely associated to phenotypes than individual founder alleles, indicating that the association component of  $S(W)$  should depend on  $B$ , not just on  $b$ .

In contrast, for dense marker maps including the disease locus, the disease alleles are more associated to phenotypes than any other (combination of) marker alleles, indicating that the association component of  $S$  should be single locus. In this case, information from other markers only serves the purpose of filling in missing data for  $S_1$  due to untyped family members. (In contrast, data might be missing at a given locus for  $S_2$  even when all family members are genotyped, if the locus is not fully polymorphic.)

The present implementation of our algorithm is computationally quite intensive, in spite of the speedups. Additional speedups of the forward-backward algorithm are possible: 1) For certain pedigree structures and marker data configurations, state space can be reduced more effectively than using founder phase symmetry alone, see Gudbjartsson et al. (2000) and Abecasis et al. (2002). 2) In recursive updating of forward and backward probabilities, one could utilize that most elements of the diagonal matrices  $D_i$  and  $D_{i+1}$  at consecutive loci are zero. This is likely to decrease computational complexity in particular when  $L$  is large. 3) Each founder individual has  $2^L$  possible



haplotypes over  $L$  loci. The common haplotypes are those with frequency exceeding a given threshold. When  $L$  is large, each  $h_{ik}$  could be restricted to common haplotypes, whose number is substantially smaller than  $2^L$ . 4) Defining a variable order Markov chain (Eronen et al., 2004), letting  $L_i$  be smaller in regions of low LD. 5) Letting each  $x_i$  correspond to a haplotype block, within which no recombinations are allowed. The number of alleles  $d_i$  can be reduced by considering only common haplotypes, and  $L$  can be chosen much smaller than for single base pair polymorphisms. When  $L = 0$ , this is essentially the approach of Abecasis and Wigginton (2005). With  $L = 1$  we get a first-order Markov chain for haplotype blocks, see e.g. Daly et al. (2001).

A great challenge is to develop faster versions or approximations of our methodology such that genomewide scans based on all polymorphic SNPs become feasible. For instance, if  $K = 6 \cdot 10^6$  SNPs are used along the human genome, at an average distance of 0.5 kb, and if linkage disequilibrium extends over 10 kb (Reich et al., 2001, The international HapMap Consortium, 2005), a value about  $L = 20$  is required to fully capture LD-structure, although a substantially smaller  $L$  is likely to approximate it quite well.

## Appendix

### Forward-Backward Algorithm

It follows from (2), (3) and (4) that  $\{W_i\}$  is a Markov chain with transition probabilities

$$P(W_{i+1}|W_i) = P(v_{i+1}|v_i)P(B_{i+1}|B_i) = P(v_{i+1}|v_i) \prod_{k=1}^{2f} P(b_{i+1,k}|h_{ik}), \quad (24)$$

for  $i = 1, \dots, K - 1$ . The absolute probabilities are

$$P(W_i) = 2^{-m} \prod_{k=1}^{2f} P(h_{ik}) \quad (25)$$

where  $2^{-m}$  is the prior probability of  $v_i$ .

Introduce

$$\begin{aligned} \mathbf{M}_i^- &= (M_1, \dots, M_i), \\ \mathbf{M}_i^+ &= (M_{i+1}, \dots, M_K), \\ \alpha_i(W) &= P(\mathbf{M}_i^-, W_i = W), \\ \beta_i(W) &= P(\mathbf{M}_i^+ | W_i = W), \end{aligned}$$

for  $i = 1, \dots, K$ . The forward probabilities  $\alpha_i(W)$  are computed recursively from left to right and the backward probabilities  $\beta_i(W)$  recursively from right to left. By applying Bayes' rule, it follows that

$$P(W_i = W | \mathbf{M}) \propto \alpha_i(W) \beta_i(W), \quad (26)$$

where the proportionality constant is chosen so that the right-hand side probabilities sum to one.

Let  $\mathcal{W}_i$  denote the set of all possible  $W_i$ . For SNPs we have  $|\mathcal{W}_i| = 2^{2fL_i+m}$ , since  $B_i$  and  $v_i$  can be chosen in  $2^{2fL_i}$  and  $2^m$  different ways. Introduce

$$\begin{aligned} \alpha_i &= (\alpha_i(W)), & 1 \times |\mathcal{W}_i| \text{ forw. prob. vector,} \\ \beta_i &= (\beta_i(W)), & 1 \times |\mathcal{W}_i| \text{ backw. prob. vector,} \\ P_i &= (P_i(W)), & 1 \times |\mathcal{W}_i| \text{ abs. prob. vector,} \\ Q_i &= (Q_i(W, W')), & |\mathcal{W}_i| \times |\mathcal{W}_{i+1}| \text{ transition matrix,} \\ D_i &= \text{diag}(D_i(W)), & |\mathcal{W}_i| \times |\mathcal{W}_i| \text{ diagonal matrix,} \end{aligned} \quad (27)$$

where  $Q_i(W, W') = P(W_{i+1} = W' | W_i = W)$ ,  $P_i(W) = P(W_i = W)$  and  $D_i(W, W) = P(M_i | W_i = W)$ . Recursive algorithms for computing forward and backward probabilities are

$$\begin{aligned} \alpha_{i+1} &= \alpha_i Q_i D_{i+1}, & i = 1, \dots, K-1, \\ \beta_i^T &= Q_i D_{i+1} \beta_{i+1}^T, & i = K-1, \dots, 1, \end{aligned} \quad (28)$$

where  $\beta_i^T$  is the transpose of  $\beta_i$ , with initial conditions

$$\begin{aligned} \beta_K &= (1, \dots, 1), \\ \alpha_1 &= P_1 D_1. \end{aligned}$$

$Q_i$  and  $P_i$  are deduced from (24) and (25) and

$$D_i(W) = 1_{\{(b,v) \rightarrow M_i\}}.$$

Here  $(b, v)$  is the part of  $W = (B, v)$  containing founder alleles at the right-most locus of  $B$  only and  $(b, v) \rightarrow M_i$  means that marker data at  $x_i$  is compatible with spreading of founder alleles  $b$  to nonfounders according to inheritance vector  $v$ . Hence  $D_i$  has only ones and zeros along the diagonal, where the ones correspond to  $W_i$  compatible with  $M_i$ . The set of nonzero diagonal entries of  $D_i$  are found using genetic descent and founder-allele graphs, see Appendix A of Kruglyak et al. (1996) or Sobel and Lange (1996).

## Unreduced Computational Complexity

For SNPs, each row of  $Q_i$  has at most  $2^{2f+m}$  nonzero elements. This is so, since given  $W_i$ , only  $v_{i+1}$  and  $b_{i+1}$  of  $W_{i+1}$  vary freely, and can be chosen in  $2^m$  and  $2^{2f}$  ways. Similarly, each column of  $Q_i$  contains at most  $2^{2f+m}$  nonzero elements. For this reason the total number of operations to update the forward and backward probability vectors in (28) is

$$O(2^{2f+m} \cdot |\mathcal{W}|) = O(2^{2(L+1)f+2m}),$$

where  $\mathcal{W}$  is the set of all  $W = (B, v)$  when  $B$  spans  $L$  adjacent loci. This is the most time-consuming part of the forward-backward algorithm and must be repeated  $K - 1$  times, once for each transition between neighbouring pairs of markers, giving a total complexity (6).

## Reduced Computational Complexity

Let  $\overline{W}_i$  denote the collection of all possible  $\overline{W}_i$ . The forward-backward calculations after founder phase reduction are similar, provided we replace  $P_i$ ,  $D_i$  and  $Q_i$  by their analogues  $\overline{P}_i$ ,  $\overline{D}_i$  and  $\overline{Q}_i$ , which are matrices of dimension  $1 \times |\overline{W}_i|$ ,  $|\overline{W}_i| \times |\overline{W}_i|$  and  $|\overline{W}_i| \times |\overline{W}_{i+1}|$ .

The crucial step is to notice that  $\{\overline{W}_i\}$  is a Markov chain with transition probabilities

$$\overline{Q}_i(\overline{W}, \overline{W}') = P(\overline{W}_{i+1} = \overline{W}' | \overline{W}_i = \overline{W}) = \sum_{W' \in \overline{W}'} Q_i(W, W'), \quad (29)$$

provided the RHS of (29) is independent of  $W \in \overline{W}$ . It is possible to prove that this is the case, see the appendix of Hössjer (2006).

The components of  $\overline{P}_i$  and  $\overline{D}_i$  are

$$\begin{aligned} \overline{P}_i(\overline{W}) &= P(\overline{W}_i = \overline{W}) = 2^f P_i(W), \\ \overline{D}_i(\overline{W}, \overline{W}) &= P(M_i | \overline{W}_i = \overline{W}) = 1_{\{(b,v) \rightarrow M_i\}}, \end{aligned} \quad (30)$$

where, in the last equation,  $W = (B, v)$  is any element of  $\overline{W}$ , and  $(b, v)$  contains only founder alleles from the rightmost locus of  $B$ . Hence,  $\overline{P}_i$  and  $\overline{D}_i$  are computed essentially as for the non-reduced state space.

It turns out that the recursive computation (28) of forward and backward probabilities can be speeded up by updating only one component of  $W_i$  at a time (see also Idury and Elston, 1997). In more detail, write

$$Q_i = Q_{i1} Q_{i2} \dots Q_{i,2f+m}.$$

For  $k = 1, \dots, 2f$ ,  $Q_{ik}$  updates the  $k^{\text{th}}$  founder haplotype by dropping the allele at  $x_{i-L_i+1}$  and adding a new allele at  $x_{i+1}$ . All other founder haplotypes as well as the inheritance vector are kept fixed. For  $k = 2f + j$ ,  $Q_{ik}$  updates the  $j^{\text{th}}$  meiosis of the inheritance vector, while the remaining meioses, as well as the founder haplotypes, are kept fixed. Hence, each row and column of  $Q_{ik}$  has at most two nonzero elements.

Define

$$\overline{Q}_{ik}(\overline{W}, \overline{W}') = \sum_{W' \in \overline{W}'} Q_{ik}(W, W'). \quad (31)$$

After some computations, one finds that

$$\overline{Q}_i = \overline{Q}_{i1} \overline{Q}_{i2} \dots \overline{Q}_{i,2f+m}.$$

Since the total number of nonzero terms on the RHS of (31) along a column or row of  $\overline{Q}_{ik}$  is at most 2, it follows that the recursive updating

$$\begin{aligned} \bar{\alpha}_{i+1} &= \bar{\alpha}_i \overline{Q}_i \overline{D}_{i+1}, \\ \bar{\beta}_i^T &= \overline{Q}_i \overline{D}_{i+1} \bar{\beta}_{i+1}^T, \end{aligned}$$

of forward and backward probabilities has computational complexity (using  $|\overline{\mathcal{W}}_i| \leq |\overline{\mathcal{W}}| = 2^{2fL+m-f}$ )

$$O(2(2f+m)|\overline{\mathcal{W}}|) = O((2f+m)2^{2fL+m-f}).$$

Repeating this at all marker loci yields a total computational complexity (8).

## Derivation of (20)

Let  $L_i(w) = P(\mathbf{M}|w_i = w)$ . Using (1), (11) and Bayes' Theorem, it follows that

$$P_i(w) = C^{-1} L_i(w) P_{i,0}(w), \quad (32)$$

where  $C = P_0(\mathbf{M}) = \sum_w P_{i,0}(w) L_i(w)$ . Let  $L(x_i, \epsilon) = P_{i,\epsilon}(\mathbf{M}|Y)$  denote the retrospective likelihood. Combining (32), (19) and (17) we get

$$\begin{aligned} L(x_i, \epsilon) &= \sum_w L_i(w) P_{i,\epsilon}(w) \\ &= \sum_w L_i(w) P_{i,0}(w) (1 + Z_i(w) \gamma_i \epsilon^T) \\ &= C (1 + C^{-1} \sum_w Z_i(w) L_i(w) P_{i,0}(w) \gamma_i \epsilon^T) \\ &= C (1 + \sum_w Z_i(w) P_i(w) \gamma_i \epsilon^T) \\ &= C (1 + \bar{Z}_i \gamma_i \epsilon^T). \end{aligned}$$

Taking logarithms we arrive at (20), with constant =  $\log C$ .

## Simulating from $P_{i,\epsilon}(\mathbf{b}, \mathbf{v}|Y)$

Here we provide a detailed description on how to simulate  $\mathbf{b}$  and  $\mathbf{v}$  in Step 1 of the algorithm in the simulation section. Under  $H_0$  ( $\epsilon = 0$ ), we simply use

$$P_0(\mathbf{b}, \mathbf{v}|Y) = P(\mathbf{b}, \mathbf{v}) = P(\mathbf{v}) \prod_{k=1}^{2f} P(\mathbf{h}_k),$$

where  $\mathbf{h}_k = (b_{1k}, \dots, b_{Kk})$  is the  $k^{\text{th}}$  founder haplotype of length  $K$ , containing alleles at all marker loci. Hence,  $\mathbf{v}$  and each  $\mathbf{h}_k$  are generated as independent Markov chains from the left to the right (say) with transition probabilities (2) and (4).

Under  $H_1$ , Step 1 is slightly more complicated:

1. (a) Generate the allele configuration  $w_i = (b_i, v_i)$  at the disease locus  $x_i$  from  $P_{i,\epsilon}(w|Y)$ .
- (b) Given  $v_i$ , generate  $\mathbf{v}^- = (v_1, \dots, v_{i-1})$  and  $\mathbf{v}^+ = (v_{i+1}, \dots, v_K)$  from  $P(\mathbf{v}^-|v_i)P(\mathbf{v}^+|v_i)$ . That is, inheritance vectors are generated as two independent Markov chains to the left and right of  $x_i$ . The transition probabilities are given by the right-hand side of (2) in both directions.
- (c) Given  $b_i$ , generate  $B_i$  from  $P(B_i|b_i) = \prod_{k=1}^{2f} P(h_{ik}|b_{ik})$ . That is,  $2f$  founder haplotypes of length  $L_i$  with rightmost locus  $x_i$  are generated independently according to

$$P(h_{ik}|b_{ik}) = \frac{P(h_{ik})}{\sum P(\tilde{h}_{ik})},$$

where the sum in the denominator is taken over all haplotypes  $\tilde{h}_{ik}$  of length  $L_i$  with rightmost allele  $b_{ik}$  at  $x_i$ .

- (d) Given  $B_i$ , generate  $\mathbf{B}^- = (B_1, \dots, B_{i-1})$  and  $\mathbf{B}^+ = (B_{i+1}, \dots, B_K)$  from  $P(\mathbf{B}^-|B_i)P(\mathbf{B}^+|B_i)$ . That is, founder haplotypes are generated independently to the left of

$x_{i-L_i+1}$  and to the right of  $x_i$  as  $2 \cdot 2f$  Markov chains with transition probabilities (4) in direction from left to right and

$$P(b_{i-L_i,k}|b_{i-L_i+1,k}, \dots, b_{i,k}) = \frac{g_{ik}}{\sum \tilde{g}_{ik}} \quad (33)$$

in direction from right to left. In the denominator of (33), the sum is taken over all haplotypes  $\tilde{g}_{ik}$  of length  $L_i + 1$  whose  $L_i$  rightmost alleles equal  $h_{ik}$ .

In Step 1a), we may use a biological model, with  $\epsilon$  containing penetrances and disease allele frequencies. An alternative is to start with a score function  $S$  and disease allele frequencies and then use (19).

## Acknowledgments

The first author was sponsored by the National Research School in Genomics and Bioinformatics and the second by the Swedish Research Council, contract nr 626-2002-6286.

## References

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. (2002). Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30:97-101.
- Abecasis GR, Wigginton JE. (2005). Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. *Am J Hum Genet* 77:754-767.
- Barrett JC, Fry B, Maller J, Daly MJ. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2):263-265.
- Bass MP, Martin ER, Hauser ER. (2002). Pedigree generation for analysis of genetic linkage and association. *Am J Hum Genet*, suppl 71:569.
- Baum LE. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* 3:1-8.
- Boehnke M. (1986). Estimating the power of a proposed linkage study: A practical computer simulation approach. *Am J Hum Genet* 39:513-527.
- Boyles AL, Scott WK, Martin ER, Schmidt S, Li Y-J, Ashley-Koch A, Bass MP, Schmidt M, Pericak-Vance MA, Speer MC and Hauser E. (2005). Linkage disequilibrium inflates type 1 error rates in multipoint linkage analysis when parental genotypes are missing. *Hum Hered* 59:220-227.
- Broman KW and Feingold E. (2004). SNPs made routine. *Nature Methods* 1(2):104-105.
- Browning BL, Brashear DL, Butler AA, Devon DC, Harris EC, Nelsen AJ, Yarnall DP, Ehm MG, Wagner MH. (2004). Linkage analysis using single nucleotide polymorphisms. *Hum Hered* 57:220-227.
- Cantor RM, Chen GK, Pajukanta P, Lange K. (2005). Association testing in a linked region using large pedigrees. *Am J Hum Genet* 76:538-542.

- Clayton D. (1999). A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *Am J Hum Genet* 65:1170-1177.
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander E. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics* 29:229-232.
- Donnelly P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theor Pop Biol* 23:34-64.
- Douglas JA, Boehenke M, Gillanders T, Trent JM, Gruber SB. (2001). Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature Genetics* 28:361-364.
- Elston RC, Stewart J. (1971). A general method for the analysis of pedigree data. *Hum Hered* 21:523-542.
- Eronen L, Geerts F, Toivonen H. (2004). A Markov chain approach to reconstruction of long haplotypes. *Pac Symp Biocomput* 104-115.
- Evans DM, Cardon, LR. (2004). Guidelines for genotyping in genomewide linkage studies: Single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet* 75:687-692.
- Fulker DW, Cherny SS, Sham PC, Hewitt JK. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* 64:259-267.
- Göring HH, Terwilliger JD. (2000). Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* 66:1310-1327.
- Gudbjartsson DF, Jonasson K, Figge ML, Kong A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* 25:12-13.
- Hodge SE, Boehenke M, Spence MA. (1999). Loss of information due to ambiguous haplotyping of SNPs. *Nature Genetics* 21:360-361.
- Hössjer O. (2003). Determining inheritance distributions via stochastic penetrances. *Journal of the American Statistical Association* 98:1035-1051.
- Hössjer O. (2005a). Combined association and linkage analysis for general pedigrees and genetic Models. *Statistical Applications of Genetics and Molecular Biology* 4(1): Article 11.
- Hössjer O. (2005b). Conditional likelihood score functions in linkage analysis. *Biostatistics* 6:313-332. Supplementary material at <http://biostatistics.oupjournals.org/>.
- Hössjer O. (2006). Modelling the effect of inbreeding among founders in linkage analysis. *Theor Pop Biol* 70:146-163.
- Huang Q, Shete S and Amos CI. (2004). Ignoring linkage disequilibrium

- among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet* 75:1106-1112.
- Idury RM, Elston, RC. (1997). A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. *Hum Hered* 47:197-202
- Jung J, Fan R, Jin L. (2005). Combined linkage and association mapping of quantitative trait loci by multiple markers. *Genetics* 170:881-898.
- Kong A, Cox N. (1997). Allele-sharing models: LOD Scores and Accurate Linkage Tests. *Am J Hum Genet* 61:1179-1188.
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am J Hum Genet* 58:1347-1363.
- Lander ES, Green P. (1987). Construction of multilocus genetic maps in humans. *Proc Natl Acad Sci USA* 84:2363-2367.
- Li M, Boehnke M, Abecasis GR. (2005). Joint modelling of linkage and association: Identifying SNPs responsible for a linkage signal. *Am J Hum Genet* 76:934-949.
- McPeck S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol* 16:225-249.
- Niu T, Qin ZS, Xu X, Liu JS. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 70:157-169.
- Ott J. (1989). Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci USA* 86(11): 4175-4178.
- Pérez-Enciso M. (2003). Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: A Bayesian unified framework. *Genetics* 163:1497-1510.
- Ploughman LM, Boehnke M. (1989). Estimating the power for a proposed linkage study for a complex genetic trait. *Am J Hum Genet* 44:543-551.
- Rabiner LR. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77:257-286.
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward, Lander ES. (2001). Linkage disequilibrium in the human genome. *Nature* 411:199-204.
- Self S, Liang K-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J Am Stat Assoc* 82:605-610.
- Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. (2002). Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am J Hum Genet* 71:992-995.



- Schaid DJ, Guenther JC, Christensen GB, Hebring S, Rosenow C, Hilker CA, McDonnell, SK, Cunningham JM, Slager SL, Blute ML, Thibodeau SN. (2004). Comparison of microsatellites versus single-nucleotide polymorphisms in a genomewide linkage screen for prostate cancer-susceptibility loci. *Am J Hum Genet* 75:948-965.
- Sham PC, Cherny SS, Purcell S, Hewitt JK. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 66:1616-1630.
- Shih M-C, Whittemore AS. (2002). Tests for genetic association using family data. *Genet Epidemiol* 22:128-145.
- Sobel E, Lange K. (1996). Descent graphs in pedigree analysis: applications to haplotype mapping, locations scores, and marker-sharing statistics. *Am J Hum Genet* 58:1323-1337.
- Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. (2003). Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum Hered* 55:27-36.
- Terwilliger JD, Speer M, Ott, J. (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet Epidemiol* 10:217-224.
- The International HamMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437:1299-1320.
- Thorisson GA, Smith AV, Krishnan L, Stein LD. (2005). The International HapMap Project Web site. *Genome Research* 15:1592-1593.
- Whittemore A, Halpern J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* 50:118-127.
- Whittemore A. (1996). Genome scanning for linkage: An overview. *Biometrics* 59:704-716.
- Whittemore AS, Tu I-P. (2000). Detection of disease genes by use of family data. I. Likelihood-based theory. *Am J Hum Genet* 66:3128-3140.
- Xiong M, Jin L. (2000). Combined linkage and linkage disequilibrium mapping for genome screens. *Genet Epidemiol* 19:211-234.

## Figures and tables

Marker nr $i$	Marker id	Position (cM)	$P(b_{ik} = 0)$	$P(b_{ik} = 1)$
1	rs79997310	36.5106	0.90	0.10
2	rs1937760	36.5645	0.87	0.13
3	rs12428936	36.6189	0.89	0.11
4	rs9532010	36.6664	0.52	0.48
5	rs95477761	36.7186	0.51	0.49
6	rs7490113	36.7865	0.91	0.09
7	rs1359217	36.8123	0.58	0.42
8	rs9566217	36.8612	0.52	0.48
9	rs7989580	36.9107	0.92	0.08
10	rs7983125	36.9588	0.80	0.20

Table 1: Information about the SNPs used in the study, located along Chromosome 9 and spanning approximatively 0.45 cM.

Haplotype	Markers								
	12	23	34	45	56	67	78	89	910
00	0.87	0.86	0.41	0.11	0.51	0.49	0.20	0.48	0.80
01	0.03	0.01	0.48	0.41	0	0.42	0.38	0.05	0.12
10	0	0.03	0.11	0.40	0.4	0.09	0.32	0.44	0
11	0.10	0.10	0	0.08	0.09	0	0.10	0.03	0.08

Table 2: Estimated population frequencies of haplotypes of length two (Niu et al., 2002).

Marker	2	3	4	5	6	7	8	9	10
1	0.778	0.915	0.104	0.020	0.008	0.002	0.000	0.013	0.007
2		0.708	0.046	0.002	0.100	0.012	0.004	0.007	0.006
3			0.114	0.027	0.006	0.000	0.000	0.011	0.005
4				0.440	0.018	0.008	0.010	0.002	0.007
5					0.108	0.070	0.013	0.045	0.036
6						0.077	0.091	0.008	0.007
7							0.034	0.015	0.048
8								0.045	0.015
9									0.279

Table 3: Table displaying the correlation coefficient  $r^2$  of LD between markers.

$F$	$\lambda_{0.05}$	$\lambda_{0.01}$	$\lambda_{0.001}$	$\lambda_{0.0001}$
$\chi^2(1)$	3.84	6.63	10.83	15.14
$0.5\chi^2(0) + 0.5\chi^2(1)$	2.71	5.41	9.55	13.83
$0.5\chi^2(1) + 0.5\chi^2(2)$	5.14	8.28	12.81	17.50

Table 4: Quantiles  $\lambda_\alpha = F^{-1}(1 - \alpha)$  of different distributions  $F$ .

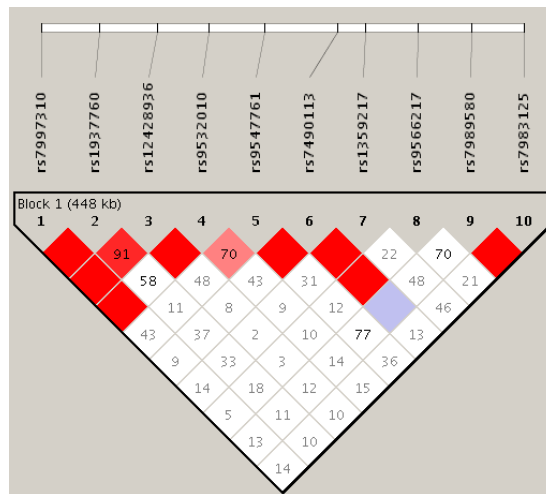


Figure 1: LD structure of the markers with colour scheme displaying Lewontin ( $D'$ ) measure of disequilibrium between markers. The value in the each box represents a hundredth part.  $D'$  prime values of one are not shown (i.e. boxes are empty), Barrett et al. 2004.

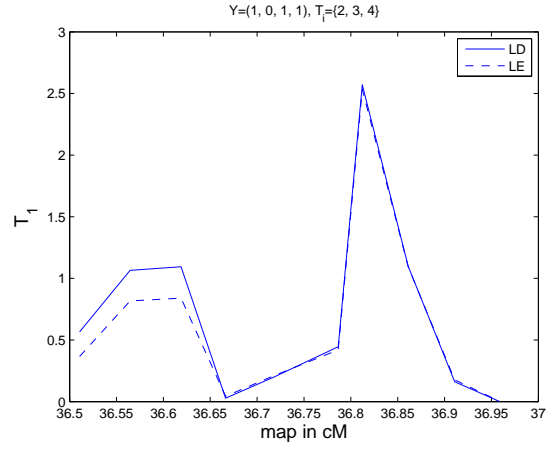


Figure 2: LR statistic  $T$  for association analysis for data set simulated under  $H_0$  when  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{2, 3, 4\}$ .

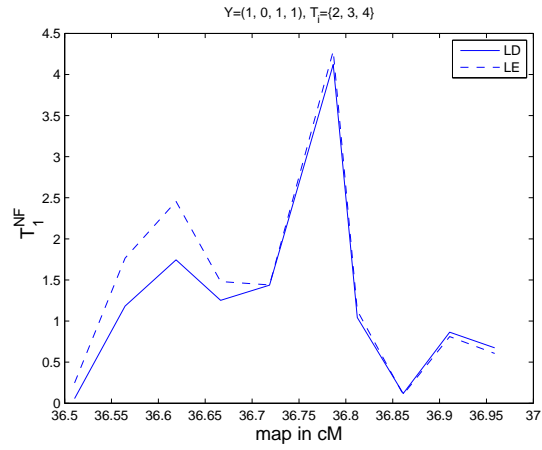


Figure 3: LR statistic  $T$  for association analysis when  $S = S_1^{NF}$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{2, 3, 4\}$ .

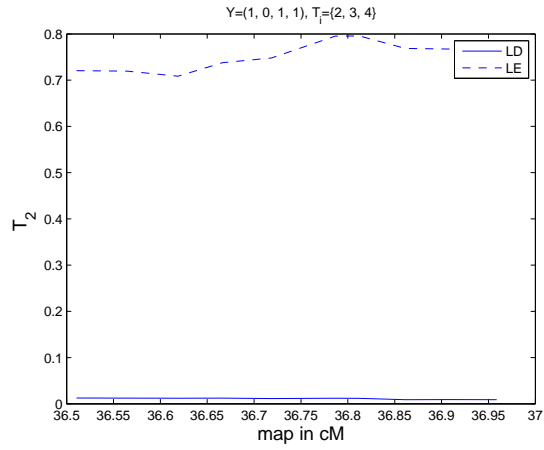


Figure 4: LR statistic  $T$  for linkage analysis when  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{2, 3, 4\}$ .

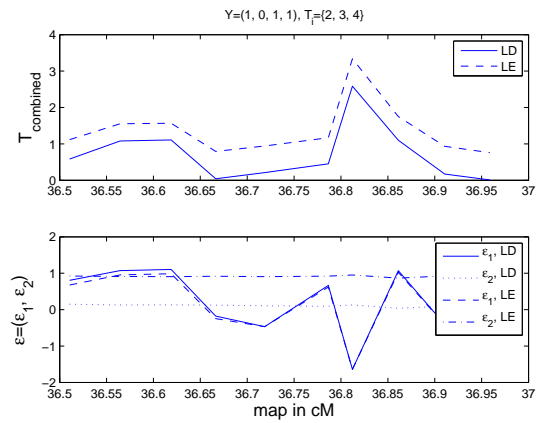


Figure 5: LR statistic  $T$  for combined association and linkage analysis when  $S = (S_1, S_2)$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{2, 3, 4\}$ . Figure below shows  $\hat{\epsilon} = (\hat{\epsilon}_1, \hat{\epsilon}_2)$ .

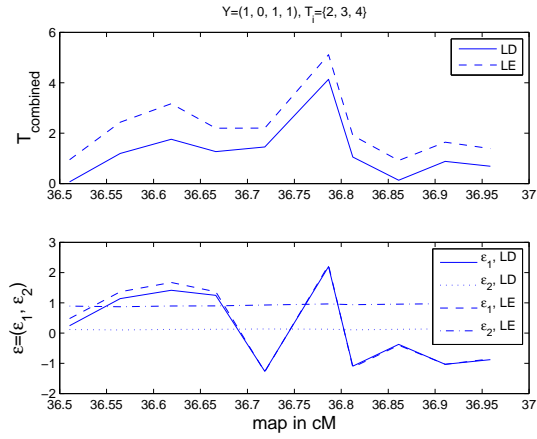


Figure 6: LR statistic  $T$  for combined association and linkage analysis when  $S = (S_1^{NF}, S_2)$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{2, 3, 4\}$ . Figure below shows  $\hat{\epsilon} = (\hat{\epsilon}_1, \hat{\epsilon}_2)$ .

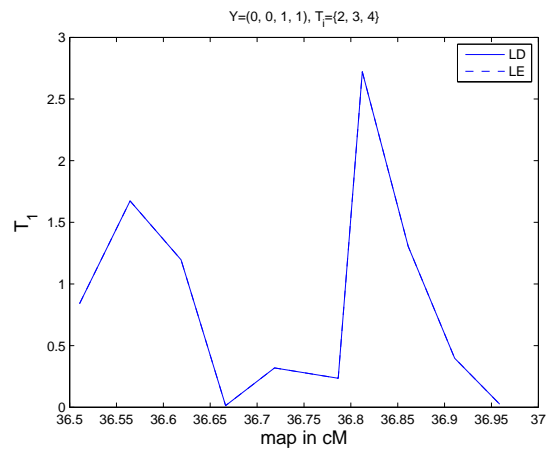


Figure 7: LR statistic  $T$  for association when  $S = S_1$ ,  $Y = (0, 0, 1, 1)$  and  $\mathcal{T}_i = \{2, 3, 4\}$ .

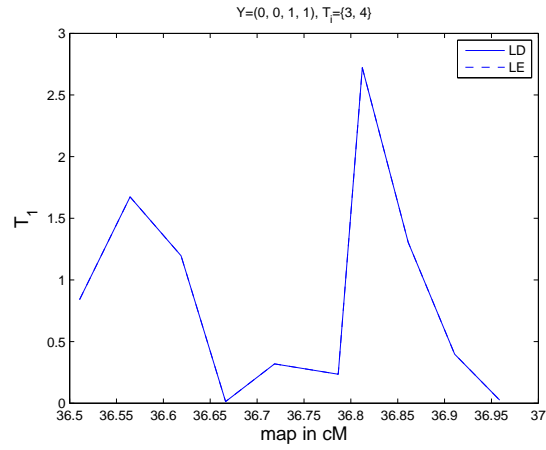


Figure 8: Likelihood ratio statistic  $T$  for association when  $S = S_1$ ,  $Y = (0, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

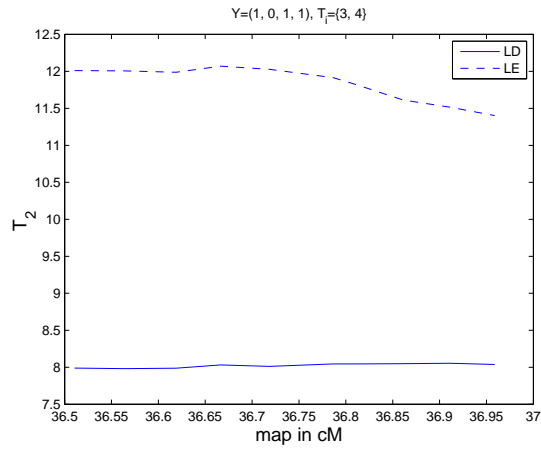


Figure 9: LR statistic  $T$  for linkage when  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .



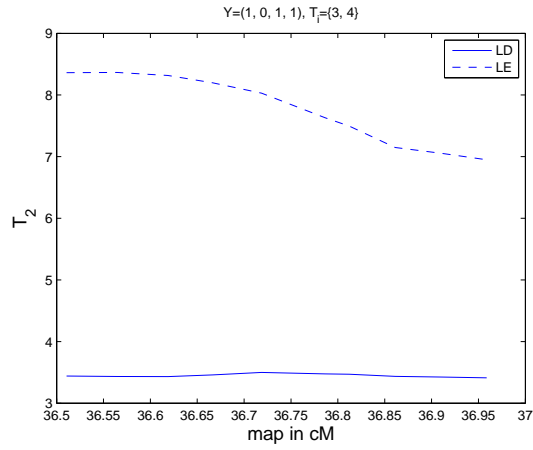


Figure 10: LR statistic  $T$  for linkage when  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

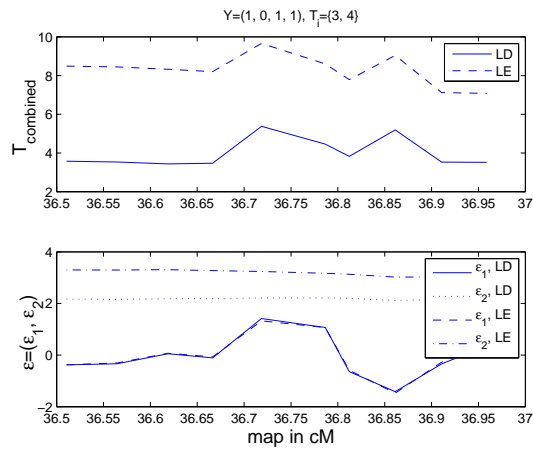


Figure 11: LR statistic  $T$  for combined association and linkage analysis when  $S = (S_1, S_2)$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

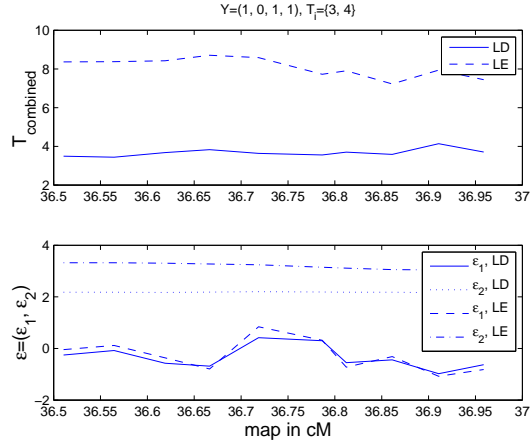


Figure 12: LR statistic  $T$  for combined association and linkage analysis with  $S = (S_1^{NF}, S_2)$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

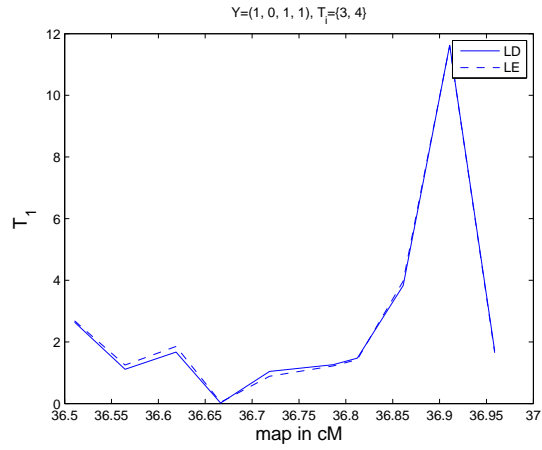


Figure 13: LR statistic  $T$  for association analysis for data simulated under  $H_1$  with marker 9 as a disease locus and penetrance values  $(f_0, f_1, f_2) = (0.10, 0.11, 0.12)$ .  $S = S_1$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

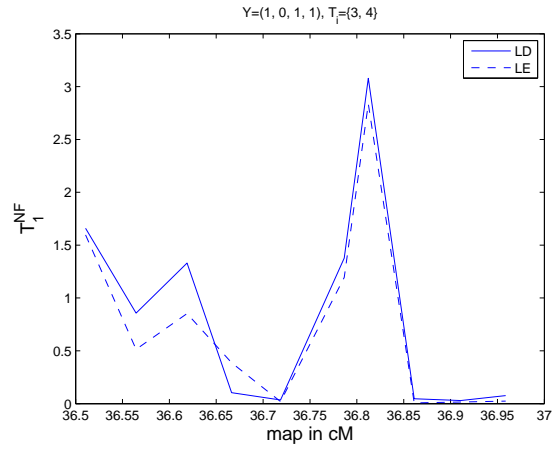


Figure 14: LR statistic  $T$  under  $H_1$  (see Figure 13) for association analysis when  $S = S_1^{NF}$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

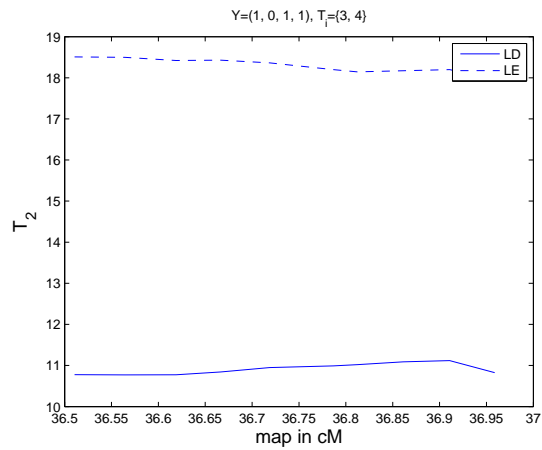


Figure 15: LR statistic  $T$  under  $H_1$  (see Figure 13) for linkage analysis when  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

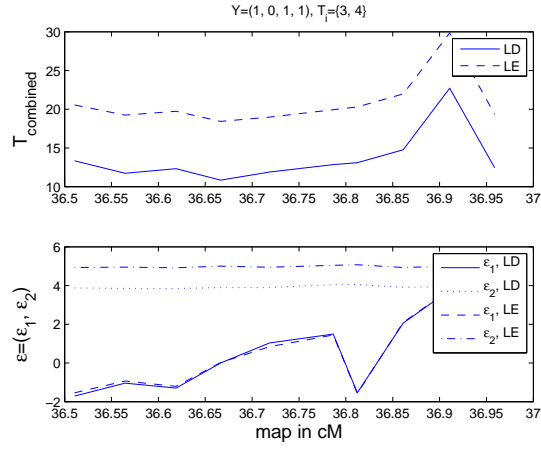


Figure 16: LR statistic  $T$  under  $H_1$  (see Figure 13) for combined association and linkage analysis when  $S = (S_1, S_2)$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .

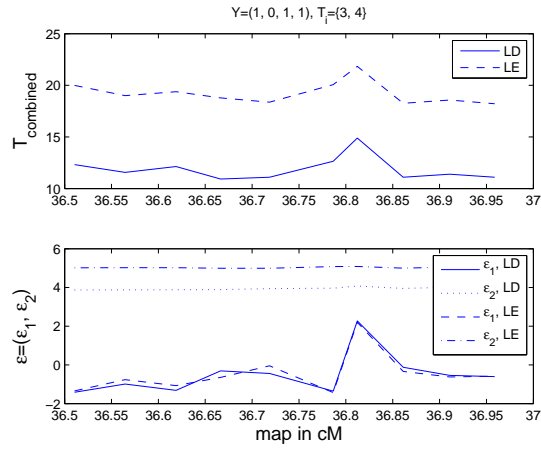


Figure 17: LR statistic  $T$  under  $H_1$  (see Figure 13) for combined association and linkage analysis when  $S = (S_1^{NF}, S_2)$ ,  $Y = (1, 0, 1, 1)$  and  $\mathcal{T}_i = \{3, 4\}$ .